Understanding Complex Constructions: A Quantitative Corpus-Linguistic Approach to the Processing of English Relative Clauses

Dissertation zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Philosophischen Fakultät

der Friedrich-Schiller-Universität Jena

von Daniel Wiechmann, MA

geboren am 27.11.1974 in Itzehoe

Gutachter

1. Prof. Dr. Holger Diessel (FSU Jena)
2. Prof. Dr. Volker Gast (FSU Jena)
3. Prof. Dr. Anatol Stefanowitsch (Universität Bremen)

Tag der mündlichen Prüfing: 9.4.2010

TABLE OF CONTENTS

1 Introduction

1.1 Aims of this study

The present study will scrutinize English relative clause constructions to further explore the relationship between the shapes of grammars (and by implication, any given particular grammar) and the cognitive processes that motivate the forms licensed by a grammar. In line with the central tenets of cognitive linguistics, it is assumed that the human linguistic system develops under strong constraints from general cognition, most notably constraints from mechanisms of categorization, (symbolic) representation and on-line processing. In the attempt to contribute to the further development of the cognitive approach to language, the present thesis draws on ideas from cognitive psychology and computational linguistics/AI research and tries to connect these more firmly to linguistic theorizing as envisaged in cognitive construction grammar (Langacker 1987, 2008; Goldberg 1995, 2006). Following research into exemplar-based language processing (Bod 1998, Daelemans 2002, Daelemans & van den Bosch 2005, Skousen 1989, Skousen et al. 2002), special emphasis is put on the role of memory and analogical processes. It is argued that in order to understand the nature of linguistic knowledge, it is advantageous to entertain a unified conception of linguistic representation and processing. In the exemplar-based view, language learning is simply the storage of linguistic experience in memory, and language processing is the use of established memory structures. Following these avenues of research, the present account assumes that language experience can be represented by a corpus of parsed utterances, so that a distributional analysis of such a corpus can yield meaningful insights into the way humans behave linguistically, e.g. which structures are likely to cause more processing difficulties than others. In order to connect these ideas to ideas developed in cognitive construction grammar, special attention will be paid to the notion of a schema or a schematic construction, which is viewed as the theoretical construct in linguistics corresponding to coherent classes of exemplars in the psychological models of linguistic knowledge embraced.

Throughout this work, relative clause constructions (henceforth RCCs) will be

portrayed as multi-clause constructions in the sense of construction grammar and so are conceived of as complex signs, i.e. pairings of form and meaning/function (Langacker 1987, Lakoff 1987, Goldberg 1995, 2006). The formal pole of signs in a linguistic system can be as concrete as a phonological gestalt, as in the case of constructions traditionally referred to as morphemes, e.g. the [*-er*$_{\text{AGENTIVE/INSTRUMENTAL}}$] suffix, but it can also be highly abstract as in the case of argument structure constructions, e.g. the ditransitive construction [$S_{\text{TRANSFER-SCHEMA}}$ [NP_{AGENT} $V_{\text{MEANS OF TRANSFER}}$ $NP_{\text{BENEFICIARY}}$ NP_{THEME}]]. The form of such high level constructions can be described only in terms of abstract linguistic categories as their formal poles are highly variable at the lexical level. Much of the early work in construction grammatical description has focused on showing that linguistic signs can also assume various types of intermediate levels that have both a fixed and a variable part (e.g. the "*What's* X *doing* Y" construction, cf. Kay & Fillmore 1999).

A complete characterization of a high level construction, such as an English RCC, will have to take into account its constitutive constructions and their properties. This is to say that an actual utterance of an RCC always simultaneously instantiates a number of lower level constructions. Hence, a given utterance of a particular type of complex sentence can be described as a particular fixation of all variable slots of its component constructions. In other words, any actual instance of an RCC can be viewed as a particular state (or configuration) of a highly variable system. The notion of a configuration (of a state space) will receive special attention throughout this work and so I will portend to it at various occasions before I define it more rigorously in later sections. A first indication of how this notion will be put to use is presented in the discussion of the structures in (1) and (2), which exemplify a scenario in which both clausal constituents of the RCC, the main clause (MC) and the relative clause proper (RC), are monotransitive:

(1) John likes the team that won the season.
(2) The guy John despises fancies another team.

Abstracting away from the lexical material in (1) and (2), we can disclose some important

dimensions along which English RCCs can vary. Figures 1 and 2 schematically represent three of these dimensions to give the reader a taste of the degree of variability of the construction under scrutiny and the range of possible configurational states:

			Ø	**Ø**		Ø
$SUBJ_{MC}$	[V	[OBJ_{MC}	[**R**	$SUBJ_{RC}$	**V**	**OBJ_{RC}**]]]
John	*likes*	*the team*	*that*		*won*	*the season*

Figure 1: Object RCC [monotrans—monotrans]

	Ø	Ø		**Ø**	
[$SUBJ_{MC}$	[**R**	**$SUBJ_{RC}$**	**V**	OBJ_{RC}]]	VP_{MC}
The guy	*that*	*John*	*despises*		*fancies another team*

Figure 2: Subject RCC [monotrans—monotrans]

The first dimension of contrast illustrated in Figures 1 and 2 concerns what is often termed the external syntax of the relative clause and pertains to the attachment site of the RC. In English, modification by means of a RC is rather unconstrained, i.e. a RC can modify any given nominal in a clause. In Figure 1 it modifies the head noun of the direct object NP of the main clause, whereas in Figure 2 it modifies the head of the subject NP. Ever since the cognitive revolution in the sixties (Miller 2003), linguists have tried to relate properties of linguistic structures to properties of the human processing system by investigating the effects of such structural differences on the processing of the corresponding structures. Modifying a VP-internal nominal is often presumed to introduce less processing difficulty than a modification of the subject nominal (center embedding >> right embedding).[1] In this view,

[1] The expression 'x >> y' should be read as 'x is harder to process than y' or 'the processing demand of x is

processing difficulty is a function of the structural properties of the linguistic expression and so invites talking of a structure's processing demand. A very pertinacious belief in psycholinguistics holds that—all other things being equal—center embedded structures are harder to process than right embedded ones. It is a central goal of this study to show that claims of that type not only are too general and hence need to be qualified, but they are also likely to lead to an inadequate conception of the underlying processing mechanism. It will be argued that a crucial conceptual error results from the seemingly innocent ceteris paribus condition, i.e. the 'all other things being equal' condition: as soon as we have a closer look at actual, contextualized language data, we begin to see that things are hardly ever equal.

But let us return to the examples presented as (1) and (2): A second dimension of contrast illustrated in these examples concerns the "internal syntax" of the RC. English makes use of the gapping strategy of relativization (cf. § 1.1.2), in which one of the syntactic roles inside the relative clause is omitted. In order to felicitously interpret the sentence, the element that gets modified by the RC (often called the *head* or *pivot* of the RC) has to be "(re-)inserted" into this "gap". The type of role that gets relativized—i.e. the role played by the head within the relative clause proper—provides the label of the type of RC, so that (1) instantiates a subject relative clause whereas (2) exemplifies an object relative clause. Again, all other things being equal, object relatives are often considered to introduce greater processing demands than subject relatives, which again is consistent with the idea that processing difficulty is a function of linguistic (structural) complexity. It will be argued that, very much like the claim concerning the processing demand and type of embedding, a statement of such type is too strong and ultimately misleading.

Thirdly, relative clauses are usually introduced by an element R (for *relativizer*), which can be omitted under certain circumstances. It is this phenomenon, the omission of non-obligatory relativizers in English, which will later serve as the testing ground of the central hypothesis pursued in this work, namely that processing difficulty is a direct function

---

greater than that of y'.

of the degree of entrenchment of a constructional configuration, i.e. a particular type of RCC. The hypothesis holds that optional relativizers are likely to be dropped, when the RCC under production instantiates a highly entrenched configuration such that the greater the level of entrenchment, the greater the likelihood of relativizer omission.

So far we have seen three variables that have been suggested to potentially influence the processing of a RCC. Crossing these 3 factors, 'type of embedding', 'gap role', and 'presence of a relativizer', already leaves us with over 20 logically possible configurations (of which some, e.g. gap role = SUBJ & R = Ø, are not permitted in Standard English).[2] However, many other factors have been suggested to act on a RCC's processing demand (e.g. the value for animacy, concreteness and definiteness of the head, the presence of uniqueness adjective, and thematic role ordering to name but a few). Of course, the combinatorial properties of the investigated system, i.e. the sheer number of possible configurations, will become more complex with the number of variables that are introduced for its description and very quickly the patterns become complex enough to escape detection by the naked eye.

The present study will thus provide the first comprehensive survey and analysis of the distributional properties of a large number of potentially relevant factors and will make use of a collection of multivariate statistical procedures to do justice to the combinatorial complexity of the phenomenon.

The central hypothesis—that processing difficulty of an linguistic expression E is best viewed as a function of its ease of activation and hence the degree of entrenchment of E—is closely tied to the idea that frequency information is a driving force behind many mechanisms that shape language as both, a cognitive entity realized in the minds of individual language users and also as a cultural artifact that changes over historical time. However, even though it is assumed that usage frequencies eventually develop their own

[2] The exact number of possible pattern does of course depend on the number of factor levels we distinguish, which is subject to some debate when it comes to the question of how many types of relatives we wish to distinguish.

causal powers, frequency based explanations beg the question when they are not accompanied by an account of why certain constructions (or constructional configurations) are as frequent as they are or why some constructions are preferred over others. This is to say that—in the view taken here—a fully explanatory account of why languages develop the forms they do (and similarly, why certain constructions are harder to process than others) has to go beyond arguments that make reference only to frequencies of use. While accounts that pertain to differences in structural complexity (e.g., Gibson 1998, Hawkins 1994, 2004) surely can contribute to an explanation of observable distributional differences and frequency effects, the present work will argue that explanations based exclusively on structural properties are incomplete and that a pattern's frequency crucially depends on the semantic/functional pole of the construction. Simply put, in order to understand processing asymmetries among a set of patterns, it is important to keep an eye on what is done with those patterns in communicative contexts. Hence, it is argued that frequency distributions are in turn best explained from a functional linguistic perspective that adds semantic and functional properties of linguistic expressions to purely structural ones. In that view it is natural to conceive of the development of grammar as a process subject to pressures to optimize the efficiency of linguistic communication.

A large-scale distributional analysis of English relative clause constructions, as envisaged here, requires data sets that are both a) sufficiently large so that the analysis can reveal interesting effects and relationships and also b) ecologically valid in so far that the constructions under investigation reflect actual speaker's solutions to functional, viz. communicative, pressures. To meet both of these requirements, a quantitative corpus-based approach to the issue was opted for and the study is based on a balanced corpus of contemporary British English (ICE-GB R2). The distributions of configurational patterns obtained from these corpus data will be submitted to multivariate statistical analysis. Conceptually speaking, the goal of the analysis employed is twofold: first, to describe the correlational structures among a set of descriptor variables in order to determine degrees of typicality and entrenchment and second, to show that these degrees of typicality and entrenchment have in fact direct consequences for language users' grammatical choices in situations that allow for grammatical variation. The application of a collection of advanced

statistical techniques (hierarchical agglomerative cluster analysis, (hierarchical) configural frequency analysis, *k*-optimal pattern identification) also serves a secondary goal of the study: to push forward a trend that is already well under way, namely to introduce more rigid methodologies into (corpus) linguistic inquiry and show that such corpus linguistic techniques are indeed a valuable methodological complement to experimental ones in the attempt to understand human language processing.

In summation, the study will provide a quantitative corpus linguistic perspective on the representation and processing of English relative clause constructions. Using data from a balanced corpus of present day British English and a collection of multivariate statistical techniques, the study aims at providing insights about the (cognitive) organization of high level linguistic structures, viz. their degree of typicality and entrenchment, and attempts to show how this organization acts on the use of these structures and ultimately the forms of grammars.

Let me close this introductory note with a brief sketch of the structure of this study. The remainder of this chapter is dedicated to describing the phenomenon, motivate why it was chosen and elaborate on the vantage points from which the study is conducted. Chapter 2 will provide a more detailed view of the theoretical background of this study and introduce the concepts and notions that figure in the proposed analysis. Chapter 3 introduces the corpus data and variables investigated in this study. It will also host a series of bivariate analyses that aim at disclosing systematic differences across registers. The emphasis on the (potential) distributional differences across modalities, i.e. spoken versus written language, is considered to be important, because language processing via the auditory channel imposes different demands on the processing system than language processing via the visual channel. Chapter 4 will motivate, introduce and apply the corpus-based methodologies for the assessment of entrenchment values of higher level constructions so as to identify those patterns that serve as models for conventional utterance types. It will also propose and apply a new methodology geared to adequately relate the detected patterns to each other on the basis of their similarity and will explicate how specific predictions about an expression type's processing demand can be derived from its distance to a deeply entrenched pattern in the relevant state space. The

last empirical step will test the model (locally) by way of trying to predict optional relativizer omission. Finally, we will close the study with a general discussion of the obtained results in the light of available experimental data and compare the present account against their treatment in competing views on language processing.

1.1.1 Why relative clause constructions?

Having stated the most general goal of the study, viz. a better understanding of the relationship between processing demand and the shape of grammars, we may now turn to the phenomenon under investigation, English relative clause constructions. But before we have a closer look at that phenomenon, let me explain why it was chosen in the first place.

There are numerous (interconnected) reasons why relative clause constructions constitute a particularly interesting object of investigation in the present context: RCCs are structurally complex, yet very frequent and highly productive cross-linguistically. This combination of properties has attracted a huge amount of research in many domains of linguistic inquiry: from grammatical theory and linguistic typology to first and second language acquisition and language processing. Given its relevance for these and other areas of research, RCCs can be considered a bridging phenomenon and in this section I shall elaborate on some the reasons why RCCs enjoy this privileged status.

First of all, the choice of an object of inquiry is motivated by its inherent properties and so a first reason to be presented here is methodological in nature. Since this study is interested in the relationship between forms of grammar(s) and the processing of linguistic constructions, the set of candidate phenomena is delimited to those structures that exhibit some property that is suspected to influence processing difficulty.[3] Ideally, the phenomenon

[3] Unless stated otherwise, I will use the terms "processing demand" and "processing difficulty" interchangeably, even though the former arguably is better conceived of as a objective property of a linguistic form, while the latter pertains to a more subjective property that may be ascribed to the objects doing the actual processing, i.e. individual language users. However, there are reasons to believe that the value of a construction's

would of course incorporate *all* those factors that have been proposed to have an impact on the processing difficulty. While it may not be possible to ensure the inclusion of all potentially relevant factors, it is of vital importance to look at a large set of factors together—as opposed to looking at them in isolation—simply because it is only then that we can inquire about their relative impacts on the overall processing difficulty of an expression type. This, in turn, necessitates the use of multivariate statistical models. Given that we need to turn to rather sophisticated models, we should choose a phenomenon where such models have the greatest potential for non-trivial insights (cf. Turchin 2003 for a detailed discussion of this point) To put it in a nutshell, our phenomenon should (1) not be too simple, (2) ideally incorporate all the factors that are presumed to influence the processing difficulty of an utterance type and (3) utilize the capacity of the mathematical/statistical models that we believe need to be employed to understand the domain of interest, i.e. human language.

Secondly, in addition to being a phenomenon of the right kind in the methodological sense described above, it is interesting to look closely at RCCs because of their central role in the (idealized) system that we call the English language. I take it to be desirable to focus on phenomena that are central to a characterization of grammar and that cannot possibly be conceived of as peripheral (by any definition of what may constitute the property of BEING PERIPHERAL). This, however, is far from being uncontroversial. In fact, the attitude towards marginal phenomena is orthogonally different in mainstream generative linguistics and most of functional/cognitive linguistics: while many linguists—especially in generative in the Chomskyan tradition—stress the importance of accounting for the very "core" of grammar, i.e. the relatively stable states of the language faculty (Chomsky 1986), there are certainly others who have argued for the importance of the periphery. In fact, it seems fair to say that research into construction grammars—and other monostratal grammars—has started not from the alleged core but from the extreme edge of the periphery (Fillmore et al 1988 and Kay and Fillmore 1999 represent influential case studies, Bender and Flickinger 1999 presents an

processing difficulty is likely to converge on the value of its processing demand, if we investigate (representative) groups of language users.

interesting discussion of the core-periphery distinction).[4] In construction grammars, relatively general patterns, which are usually considered core phenomena, and more idiomatic patterns, which are usually considered peripheral, are on equal footing. Since more idiomatic constructions involve are larger set of conditions that license a narrower class of actual sentences and more general constructions impose fewer restrictions of possible instantiations, it appears sensible to start developing the theoretical framework from the specific to the general. Fillmore and colleagues write:

> "*It appears to us that the machinery needed for describing the so-called 'minor' or 'peripheral' of the sort which has occupied us here* [namely the let alone-construction; DW] *will have to be powerful to be generalized to more familiar structures, in particular those represented by individual phrase structure rules"*
>
> (Fillmore et al. 1988:28]

While the role of marginal phenomena in linguistic theory building remains a matter of controversy and is hence dependent on one's perspective, there certainly is a consensus in the linguistic community that any adequate theory of language should be able to account for the core phenomena. In other words, explaining the core phenomena is a necessary requirement that any theory of language has to meet, whereas the role of marginal phenomena remains controversial.

A third—and arguably most important—reason to investigate relative clauses stems from the central role these constructions play within and across linguistic disciplines. Given their high frequency in language use and their high degree of productivity cross-linguistically, relative clauses constitute a central issue of research into natural language. The

[4] By *monostratal grammars* I mean grammars that pose only a single layer of syntactic representation and no transformations.

structural complexity of relative clauses and the heterogeneous class of patterns unified under that label has inspired an enormous amount of theoretical linguistic research a lot of which was based on English (or typologically related languages). English relative clause construction exhibit at least one filler-gap dependency, a phenomenon that has intrigued linguists for decades (at least since Ross (1967/1986); Alexiadou et al. 2000 presents a collection of papers dealing with numerous syntactic issues pertaining to relative clause from a generative perspective). From such analyses, scholars in the Chomskyan tradition have often tried to deduce universal linguistic principles, e.g. universal constraints on *wh*-movement. The strong implications of the grammatical analyses of such extraction phenomena have had a strong impact on neighboring fields most notable linguistic typology. Linguistic typologists have long challenged the generative approach of research by showing that inferences from a single language to natural language as a class are problematic. Some typologists are convinced that there is a universal set of (formal) features capable of defining relative clauses cross-linguistically and demarcating relative clauses proper from other noun-modifying clausal constructions or complex noun phrase constructions (cf. Matsumoto 1997, 2007 for a discussion). Relative clauses have also attracted a lot of attention in psychological domains of language study, in both first and second language acquisition (e.g. Flynn and Foley 2004; Diessel and Tomasello 2005; Kidd et al. 2007, Brandt et al. 2008 and references therein) and also in sentence processing (Hsiao and Gibson 2003). Recent years have brought a considerable increase in interdisciplinary work in the attempt to bring together the insight gained in these fields as well as reconcile apparent incongruities.

(English) relative clauses thus meet all the desiderata: they are grammatically complex, yet highly frequent and thus central to any account of processing and grammar. The next section will present a closer look at their linguistic properties.

1.1.2 Characterizing English relative clause constructions

English relative clauses have been described on many occasions in numerous theoretical frameworks and the present study does not aim to put forth yet another specific

analysis of the construction. Nothing that interests us here depends on a particular analysis. Instead this section will focus on those properties of relative clauses that are largely uncontroversial. The characterization here is thus provided foremost to introduce and clarify some central notions that figure in the descriptions of relative clause constructions and on the distributions of the corresponding variables. However, a few comments on the width of theoretical perspectives and their developments as well as some justification on certain theoretical decisions in the characterization provided here (e.g. the syntactic status of *that* in *that*-relatives) are in order.

Early work in generative grammar gave rise to a number of treatments of the syntax of relative clauses. The interested reader may turn to Smith (1964) or Chomsky (1965) for a *Determiner S* analysis, to Ross (1967/1986) for an *NP-S* analysis, or to Stockwell et al. (1973) for a *NOM-S* analysis.[5] Within the generative tradition the discussion soon focused on the contrasts between appositive (non-restrictive) and restrictive RC and ways to express correlations between syntactic attachment and semantic composition (cf. Jackendoff 1977, Bach & Cooper 1978, McCawley 1982). The advent of the DP-hypothesis (Abney 1987) reintroduced Jackendoff's hierarchical distinction within two-level *X-bar theory,* which was the dominant interpretation within the 'Government and Binding' (GB) phase lasting until Kayne's introduction of the *antisymmetry hypothesis* (Kayne 1994). In this view, restrictive modifiers of type PP or AdjP as well as appositive RCs are analyzed by an underlying raising structure. A comprehensive state-of-the art overview of headed relative clauses and their treatment in generative grammar is presented in Bianchi (2002a, 2002b). For purposes of illustration, we will have a look at the dominant analysis during the GB phase (Figure 3)

[5] The names of the mentioned analyses indicate the postulated sisterhood of the constituents in question, i.e. the label *NP-S analysis* indicates that in that analysis, the extracted NP constituent—what we will call the *head* of the RC—is the sister of the relative clause proper.

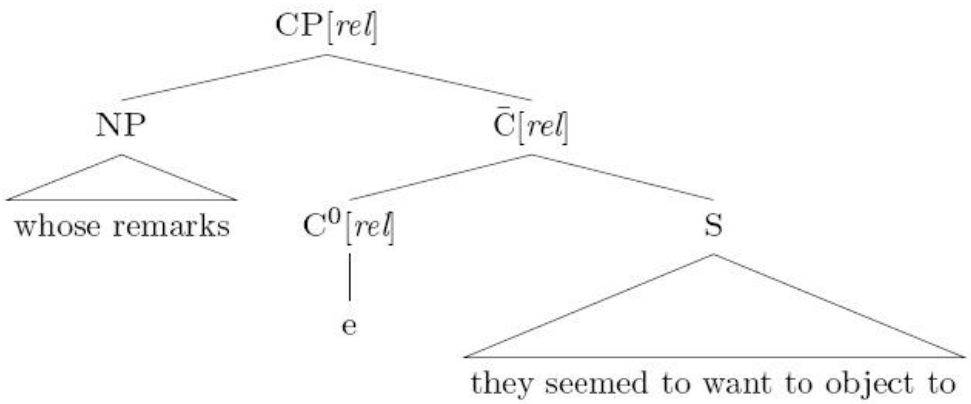

Figure 3: Dominant RC syntax in Government in Binding (from Sag 1997)

As shown in Figure 3, relative clauses (CP[*rel*]) are viewed as projections from a phonologically empty element (C^0[*rel*]). This is not the right place to comment on the descriptive adequacy of such an analysis because the treatment is motivated by the general architecture of the linguistic theory in which the analysis is framed. The postulation of the invisible element is in accordance with more general assumptions of the theory (e.g. the *projection principle*, cf. Chomsky 1981) and I take it that arguing against this particular analysis would mean to argue against a particular theoretical framework, which is not what this section is attempting to provide. However, as discussed in Sag (1997), certain invisible elements—and certainly the C^0 element here—lack sufficient independent, i.e. theory external, motivation. To the extent that we wish to assume only such invisible elements for which such motivation can be provided, we should look out for theoretical treatments that do not make use of invisible C^0. Sag (1997) provides such an analysis for restrictive RCs framed in (a constructionist version of) HPSG. Arnold (2004) presents a related analysis for non-restrictive RCs. We will have a look at a treatment of RCs in Cognitive Grammar (following Langacker 2008) later in this section, but for now let us turn to the properties of RCCs that help us understand the phenomenon in a more theory-neutral manner.

The first thing to say about relative clauses is that—even though most linguists probably believe to have a rather clear idea of what a relative clause is—it is very hard to determine a fixed set of properties that is generally accepted and that may serve as a definition of the notion. If pressed for a definition, a typical answer would certainly include

the idea that a relative clause is a type of dependent clause that usually modifies a nominal element. In English, RCs are usually introduced by a relativizer (R), which is anaphorically related to the modified nominal. Our hypothetical linguist might continue his explication by adding that this element R derives its semantic interpretation from its antecedent, the *head* (or *pivot*) of the clause. These innocent looking claims about RCs already invite a number of questions some of which are rather specific, e.g. whether or not this R element always is realized as a relative pronoun or whether we should distinguish different syntactic types of relativizers. Others questions, however, are very fundamental and cast into doubt whether a true theory-neutral description is possible at all. A good example of such a heavy question would involve the decision of whether we should treat relative clauses as (a) a syntactic phenomenon—so that the class of relative clauses is unified by a set of structural properties—or (b) a semantic one—so that the class is unified by commonalities in meaning or function—or maybe (c) a combination of the two. Linguists in the Chomskyan tradition would surely opt for the first view and would hence try to answer the question of what exactly is the unifying structural class that RCs belong to. An intuitively appealing answer is to treat RCs as clausal constituents and associate them with the category S'/CP (cf. Bianchi 1999, Hoekstra 1992, Rizzi 1997). However, linguists have also argued for the idea that they are projections of the verb and as such belong either to the category VP (Sag 1997), IP (Doherty 1993), or TP (Afarli 1994). Other linguists may observe that RCs are structurally very diverse and conclude that their essential unifying feature is semantic in nature. One of these linguists is Ronald Langacker, who defines RCs as follows:

> "[*A*] *relative clause is one invoked to characterize a nominal referent identified as a participant in the clausal process*." (Langacker 2008:426)

This definition excludes (alleged) RCs like the one in (3), which clearly comments on a proposition, namely that the person referred to by the pronominal expression *she* does not like to watch soccer.

(3) She doesn't like to watch soccer, which is ok.

The example in (3) and the question of whether it should be taken to include a RC brings up a more general problem: not only is it hard to develop an intensional definition of what it means to be a RC, it also appears far from trivial to determine the class of RC extensionally. Consider the triple of examples given as (4)-(6).

(4) The exhaust was developed by some engineers [who were working in this company].
(5) The exhaust was developed by some engineers [working in this company].
(6) The exhaust was developed by some engineers [in this company].

It appears we are facing a demarcation problem here. The example is (4) is uncontroversial: we have an overt pronominal relativizer introducing a finite subordinate clause which restricts the set of possible referents by way of specifying an additional predicate. That is to say the referent of the NP in the *by*-phrase, *some engineers*, must not only meet the condition of being an engineer, it must also meet the additional requirement of having worked in a particular company for some time. The more reduced expressions in (5) and (6) can be employed to express (roughly) the same thought, which illustrates the problem. Whereas (5) is often considered an example of a non-finite participial RC (but cf. Quirk et al. 1985 for counter-arguments), (6) may either be treated as a verbless relative clause—a construct that might raise a few eyebrows—or simply as a prepositional phrase. We may tentatively agree at this point that (4) is likely to score higher on some typicality judgment task than (6) as far as membership of the RC class is concerned. But this difference in typicality does not really help us in our attempt to rigidly define RCs.

Having addressed a potential problem of a formal definition of RCs, let us return to the functions RCs may serve in the linguistic system of English. Relative clauses are often said to be employed to either restrict the set of potential referents that the modified nominal points to (these RCs are called *defining* or *restrictive* RCs) or provide some additional,

characterizing information about that referent (*non-defining*, *non-restrictive* RCs). Example (7) presents an example of an appositive RC, while (8) presents a restrictive one.

(7) This is John, who stole my car (appositive)
(8) This is the guy who stole my car. (restrictive)

Notice that the difference between (7) and (8) reduces to a difference in head properties: while the head of the RC in (7) is a proper name, the head of RC in (8) is realized as a quantified lexical noun phrase. A result of this difference is that a comprehender of (7) is likely to not use the information presented in the RC to identify the object that the speaker of (7) intends to refer to. In most communicative contexts, the use of a proper name such as *John* presupposes that the addressee has sufficient information about the intended referent, hence making additional property ascriptions in the form of a RC rather superfluous for the task of fixing reference. Semanticists have long pointed out the special status of proper names and have even suggested that this class of expressions can be taken to refer directly within a fixed context of utterance. Ideas of direct reference go back to John Stuart Mill (Mill 1862) and continue to stimulate discussions in philosophical logic (for a discussion, cf. Recanati 1993). In contrast, (8) is usually treated as a definite description, which semantically describes a complex *sense* (in the sense of Frege 1892), which is then used to determine reference. The task for the hearer in this scenario is to identify the individual that uniquely matches the content of the description and for this task the information within the RC proper is very helpful, in fact necessary. In the attempt to reflect this distinction in terms of underlying structure, linguists have argued that sentences like (7) and (8) differ in terms of their syntax. Figure 4 presents an influential example of a syntactic contrast that has been suggested in Jackendoff (1977).

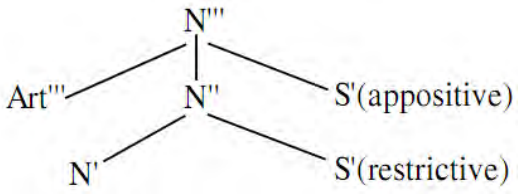

Figure 4: Structural difference between restrictive and appositive RC (Jackendoff 1997)

Even though the motivation for postulating different underlying structures appears to be traceable and comprehensible, the present study will not assign different structures to sentences like (7) and (8) for the rather pragmatic reason that deciding whether or not a RC is "in the business of fixing reference" seems impossible on the basis of observable characteristics and is hence subject to the linguist's interpretation. While I do not think that anything is truly gained by the syntactic differences illustrated in Figure 4 (at least for the purposes of this study), the introduction of subjective judgments jeopardizes the consistency of the general approach taken here, namely to rely as much as possible on observable (or at least linguistically uncontroversial) quantities and derive more complex claims in a data-driven fashion.

Howsoever one's position on these issues may be, there can be no doubt about the conception that structurally speaking English relative clauses form a rather heterogeneous class: as a dependent clause, RCs usually do not occur in isolation but are embedded in another clausal structure, which in the simplest of conceivable scenarios is the corresponding main clause.[6] Thus, in order to fully characterize relative clause constructions, we will consider certain properties of the embedding clause as well. In other words, the description of RCCs involves not only structural properties of the actual RC, e.g. its *internal syntax*, but also features of the dominating clause, i.e. the *external syntax* of the RC. With respect to the former, we can distinguish different sub-types of relative clauses with regard to the grammatical function played by the modified nominal, i.e. the head, within the relative clause

6 I will exclude from the discussion so called *free relatives (aka fused relative, independent relatives, nominal relative)*, i.e. relative clauses that do occur in isolation as in "Whatever you say, my lady".

proper.

At this point it is helpful to look at some observations made in the area of linguistic typology. Keenan and Comrie (1977) have observed that the languages of the world differ in what grammatical roles are accessible for relativization. The accessibility of the roles can be described on the basis of in a cross-linguistic hierarchy known as the *Accessibility Hierarchy* (AH), which exhibits the following order:

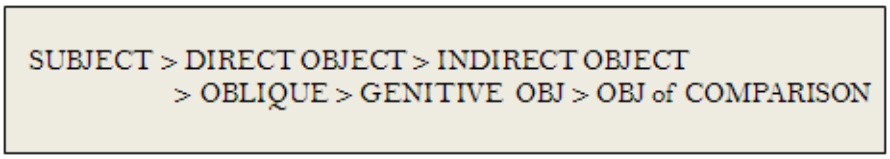

Figure 5: Accessibility Hierarchy following Keenan and Comrie (1977)

Using this hierarchy, Keenan and Comrie proposed an interesting generalization about natural languages: if a language allows relativization on an arbitrary role, it will also allow relativization on any role higher on the AH.[7] In English, all roles are accessible for relativization. With respect to the grammatical means employed for relativization, Keenan and Comrie distinguish four *relativization strategies*, i.e. strategies regarding how the grammatical role of the relativized NP is represented within the relative clause. The strategies postulated for English are known as the *gapping* (or *obliteration*) *strategy* and the *relative pronoun strategy*.[8] In the case of gapping, the relativized NP is not formally represented in the relative clause at all; instead there is a "gap" in the structure where the respective item would normally be expected. Consider (9) for exemplification:

(9) The dentist [that I saw __ last week] is a moron.

7 This is actually a little oversimplified: Keenan and Comrie qualified their statement a little more; the actual hypothesis states that if a language can relativize any position on the AH with a *primary strategy*, then it can relativize all higher position with that strategy.

8 The remaining strategies are *non-reduction* and *pronoun retention*.

In English, complements follow their subcategorizors. Hence, the object of *saw* would be expected in the position denoted by the gap ('__') in the example above. Many syntactic theories thus describe such English RCs as externally headed structures and assume that the head has been moved out of its canonical position into the position immediately to the left of the clause. The primary strategy in English, however, is the relative pronoun strategy, which in fact can be considered a European strategy in so far as it predominant in European languages but rather atypical outside Europe (cf. Lehmann 1984 for a discussion). An example of this strategy, in which the internal role of the head is signaled by way of case marking on a clause initial pronominal relativizer, is presented in (10):

(10) The dentist [whom I hate] is indeed a moron.

In contrast to the form *that* used in (9), the form *whom* does provide the hearer of (10) with sufficient information regarding the grammatical role played by the head noun by means of (object) case marking. This case marking potential of *wh*-elements introducing the relative clause is often taken as an indication of a difference in syntactic status of the relativizing element in question (e.g. Huddleston and Pullum 2002:1056-57) and this in turn has direct implications for the description of relativization strategies used in English. It is thus worthwhile to have a closer look at the syntactic category status of *that* in *that*-relatives. If we assume *that* and *wh*- relativizers to be associated with different parts of speech, we must be prepared to assign different syntactic representations to RCs introduced by the respective elements. Figure 6 may serve as an illustration.

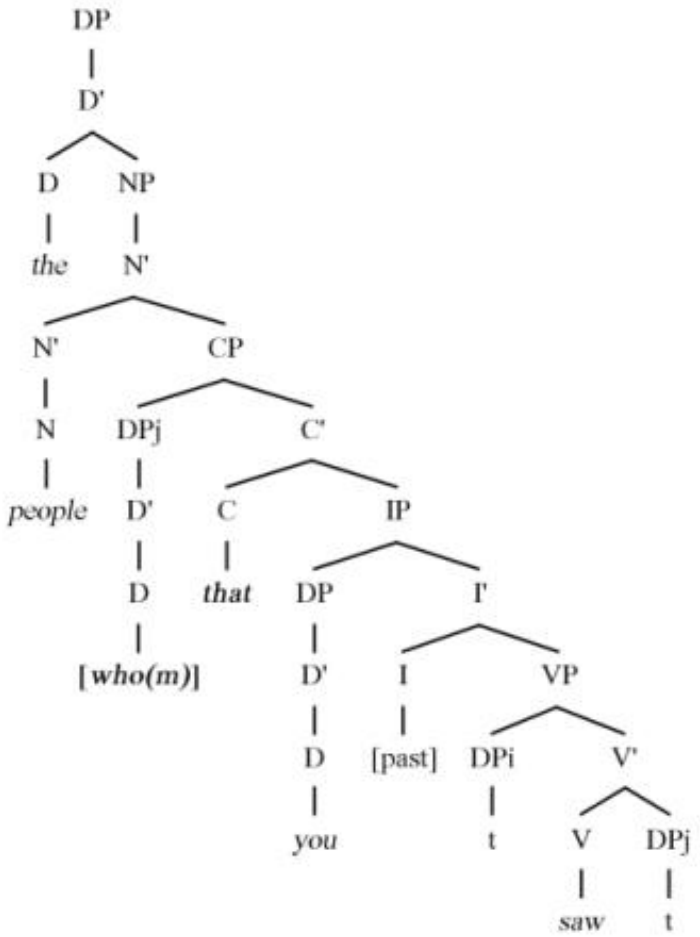

Figure 6: Syntax of *that* and *wh-* relatives

Of course, this is just one of many conceivable ways to represent the structure underlying an object RC like *the people* [*who*(*m*)|*that*] *you saw*. However, the important point here is to illustrate that if we assume *that* to be a complementizer (as opposed to a relative pronoun), we must accept the idea that the choice of relativizer has immediate repercussions on the overall syntactic pattern of the clause: *wh-* pronouns can appear in DP positions such as the specifier-CP position in Figure 6, which certainly is not true of complementizers, which occupy the head-CP position. Similarly, we are forced to postulate different semantic structures: while we may assume that pronouns are referring expressions, this is certainly infeasible for complementizers, which do only have grammatical meaning.

How to treat *that* in relative clauses is a controversial issue in linguistics. Rodney Huddleston and Geoffrey Pullum, who in their Herculean effort to describe the grammar of

modern English (Huddleston and Pullum 2002), present a number of arguments for the idea that *that* is not a pronoun:

I. **Argument from wide range of antecedent types**: If *that* were a pronoun, its use would be much wider than that of the uncontroversial relative pronouns.

II. **Argument from lack of upward perculation**: If *that* were a pronoun, theorists need to stipulate that it has no genitive form and that it never occurs as a complement of a preposition

III. **Argument from finiteness**: *That* cannot be inserted into non-*wh* relative infinitivals like the one presented in (11).

(11) *a knife that to cut with

IV. **Argument from omissibility:** In contrast to its *wh*- counterpart *which*, *that* is very largely omissible.

None of these arguments strike me as convincing: Argument I is flawed because it works only if one assumes there to be no additional variables to influence the distribution, for example register variation. Since the existence of such confounding variables cannot be excluded a priori, the claim is too strong and should in its present form be dismissed. In addition, Argument I appears to presuppose that *that* is less frequent than expected (by Huddleston and Pullum). Facts about the frequency of that (and other relativizers) are of course empirical facts and so we should give the argument a little "reality check". We will return such empirical issues shortly and present some corpus data of the frequencies of R-elements. Argument II is certainly stronger and admittedly involves an adequate characterization of Standard English. However, as Hudson (1990) observes, this statement is not generally true of English dialects. Hudson reports that in certain dialects of English *that*

allows a possessive form as in (12):

(12) The pencil that's lead is broken

On the basis of these dialectal data, we may ask ourselves how much force we are willing to Argument II. Argument III, the argument from finiteness, is downright obscure. By the same logic, we would have to say that *which* cannot be a relative pronoun either (for an acknowledgment of this implication cf. Huddleston and Pullum 2002:1057 - footnote 10). The last argument that Huddleston and Pullum mention, the argument from omissibility, strikes me as odd as it tacitly assumes that all zero relatives involve silent *that* forms. I fail to see the linguistic evidence for this assumption. Consider examples (13) to (16).

(13) The girl **who** you like so much unfortunately has a boyfriend.

(14) The girl **whom** you like so much unfortunately has a boyfriend.

(15) The girl **that** you like so much unfortunately has a boyfriend.

(16) The girl ∅ you like so much unfortunately has a boyfriend.

Assuming that all variants are well-formed, how do Huddleston and Pullum know that it was *that* and not *who* or *whom* that is phonologically empty? Even taken together the arguments brought to bear by Huddleston and Pullum do not seem to generate a lot of force.

But there are more arguments for the non-pronominal status of *that,* e.g. the argument from enclitics: In contrast to certain *wh*-variants, *that* is said to not undergo combine with reduced auxiliary verbs. This fact can be explained by the idea that auxiliary verbs to contract with (pro)nouns but certainly not with complementizers. If we do not observe auxiliaries attached to that, we may indeed take this as evidence for the non-pronominal status of *that*. However, if we have a look at real data and query a corpus of present day English, we can

easily show that the statement about clitics is empirically false: people do combine *that* with auxiliaries. Here are some examples from the ICE-GB R2:

(17) Nothing in **the road that's too short for its name** […] [S1A-023 #337]

(18) Anybody **that's got an eye each side of their nose** […] [S1A-020 #092]

(19) **The person that's affected** is me […] [S1A-026 #075]

If anything, we can in fact use the very same argument from enclitics to argue *for* the pronoun-status of *that*.

Another argument against the pronominal status of *that* involves pied-piping. Unlike its closest *wh*- counterpart *which*, *that* does not permit pied-piping. Consider (20) and (21):

(20) *The city in that we are living …

(21) *The person with that we were talking …

But as Sag (1997) notes this is true of *who* as well, whose pronominal status is beyond doubt.

(22) *The people in who we place our trust …

(23) *The person with who we are talking …

However, Sag's counterargument appears to be a little off target. It seems plausible that the ungrammaticality of (22) and (23) is due to the fact that the object of the prepositions are required to be case-marked, i.e. we can substitute *who* with *whom* (but not with *that*) to get well-formed constructions. So, the argument from pied-piping certainly retains it's a lot of its force.

And finally, a crucial difference between *that* and the *wh-* pronouns concerns case marking: we may distinguish the syntactic category underlying *that* from that of the *wh*-form on the basis of the observation that only the *wh-* variant can signal case information. Note that this case marking ability is a strong argument for the idea that English employs the relative pronoun strategy of relativization in the first place. However, the impact of this argument hinges a lot on the usage of case-marked relativizers. So let us have a quick look at the frequency distribution of relativizers in contemporary (British) English, i.e. let us turn to some illustrative data from the ICE-GB. Querying the corpus for all major relativizer types (*that*, *who*, *whom*, *which*, *whose*, *zero*) gives us the distribution presented in Figure 7.[9]

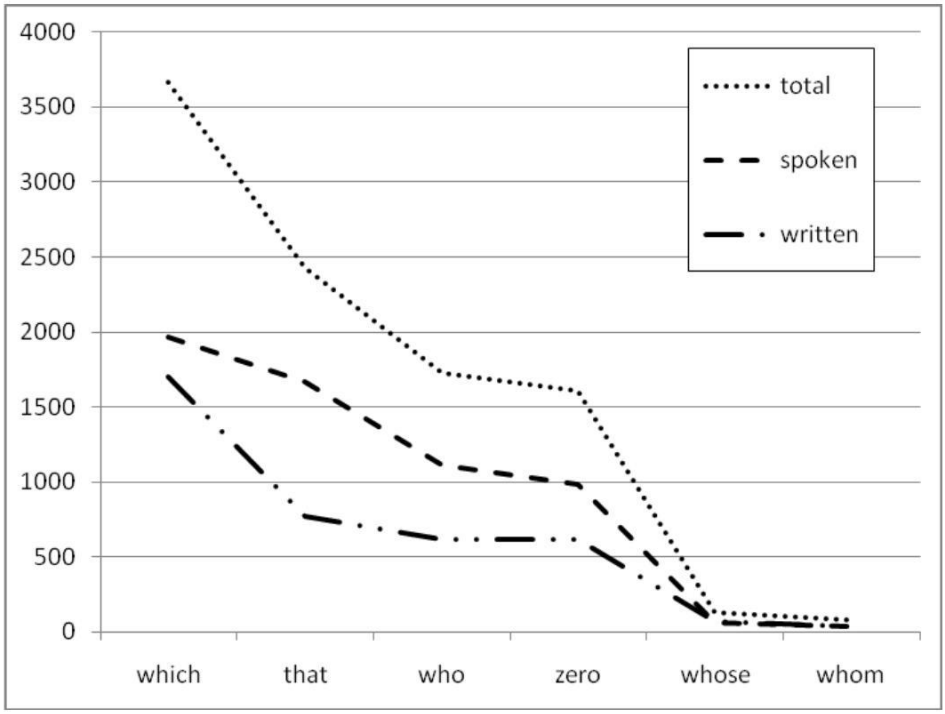

Figure 7: Relativizer usage in ICE-GB R2

On the basis of these results, we may conclude two things: first, the case-marked forms are

[9] The corpus was queried by way of matching the following strings: ((,PRON(rel))){which|that|who|whose|whom} matches all overt relativizers. The number of finite zero relatives was arrived at using ((CL(zrel, ¬edp, ¬ing, ¬infin))).

rather rare and may thus be taken to play only marginal roles in the system. The fact that *whose* is so rare might be explained by fact that that relativization on genitive objects is rather unlikely for pragmatic/communicative reasons. That is to say that one might propose that *whose* is infrequent because people hardly ever feel the need to access this low role. The low frequency of *whose* might also be due to the processing demand involved with relativization on low roles (as suggested by Keenan and Comrie 1977, cf. also Hawkins 2004). But such high processing demands certainly cannot account for the low frequency of *whom*. Object case-marking should be used not only with rather infrequent indirect objects but also with direct objects, which certainly are not rare and which rank high on the on the accessibility hierarchy. Many cases in which *whom* would be the "correct" form from the perspective of a prescriptive grammar, exhibit *who* instead (cf. e.g. Aarts 1994 for a treatment of the apparent clash between prescriptive norms and actual usage in this area of English Grammar).

A second lesson that we can learn from the results shown in Figure 7—specifically the low frequencies of both *whose* and *whom*—are very much in line with the idea that present day English continues to move away from morphological case-marking towards a system that employs a more rigid word order to signal grammatical relations. If we accept this alternative explanation, we can expect the boundaries between *wh-* and *that* relativizers to become more blurry over time. Consequently, *who* does not really signal nominative (or non-objective) case. If *whom* is indeed dispreferred in favor of *who* and the only other *wh*-item really marked for case is *whose*, which too is very infrequent, the argument from case-marking is rather weak, too.

Let us now have a look at some arguments that corroborate the idea that *that* is best viewed as a relative pronoun. The first two arguments come from the diachrony of English. Looking at the development of English we observe that (a) *that* in Middle English was the most common all-purpose relative pronoun and that (b) *that* could also introduce nominal relative clauses as in (24).[10]

[10] According to *http://www.hf.ntnu.no/engelsk/staff/johannesson/!oe/texts/eme/eme_gram.htm*

(24) *Lose that is vast in your hands*

lose what is fast in your hands

A proponent of the complementizer position would thus have to explain, why and how it might have changed its status from being the most productive relative pronoun to not being a pronoun at all.

Hudson (1990) does present yet another argument in defense the pronominal status of *that*, namely an argument from coordination: Hudson argues that *that*-relatives freely coordinate with *wh*-relatives (whereas coordination with zero|bare relatives is not possible). As an illustration, Hudson presents examples like those given in (25) to (27).

(25) *Every essay she's written and that/which I've read is on that pile.

(26) Every essay which she's written and that I've read is on that pile.

(27) Every essay that she's written and which I've read is on that pile.

In light of the evidence discussed here, it appears infelicitous to me to treat *that* as a complementizer (or relative particle or any other type of non-pronominal relativizer for that matter). If anything, there is more argumentative force pulling in the opposite direction.

Notice that the loss of case marked relativizers also affects the idea that English makes use of a pronoun strategy of relativization: neither *who* nor *which* (let alone *that*) can signal the internal role of the head. So let us return to the issue of relativization strategies and to some implications our (quite elaborated) discussion of the status of *that* has for the characterization of relativization strategies employed in English. Keenan and Comrie (1977) suggest that English uses two strategies for relativization: a primary one and a secondary one. As shown in Figure 8, the strategies that have been identified to be applied in the rich repertoire of systems that make up the languages of the world can be aligned on a scale

expressing the degree of explicitness, hence the name *explicitness hierarchy* of relativization strategies (we will focus on those strategies that can be argued to be are operative in English):

less explicit ---------- more explicit
gapping < relative pronoun < pronoun retention

Figure 8: Explicitness Hierarchy following Keenan and Comrie (1977)

The first strategy is termed the *gap strategy* (or *gapping* for short). This strategy involves cases where there is no overt case marking on the relativizer which could indicate the role played by the head inside the RC proper. The second strategy on the scale, the relative pronoun strategy, involves a pronominal relativizer that is case-marked so as to indicate the grammatical role of the head inside the RC. The pronoun retention strategy involves an explicit indication of the relativized role by means of a resumptive personal pronoun. The strategies are ordered in that particular way so as to reflect the informativeness of the forms employed. The gapping strategy is the least explicit or least informative signal type as the grammatical role played by the head inside the RC can only be recovered as soon as the RC syntax departs from the canonical form of declaratives and shows a gap, which has to be filled by the head noun so to become semantically saturated. With verbs of greater valency there may be more than one potential gap, i.e. more than one potential extraction site. An example is given in (28).

(28) This is [the book] (which|that) Ian read __ (site 1) to the children from __ (site 2) last night.

Readers (or hearers) of (28) are confronted with a local syntactic ambiguity: Until they get to the preposition *from*, they may suspect that the position from which the NP *the book* was extracted is the one labeled as "site 1", i.e. the direct object position of the verb *read*. As the rest of the structure unfolds, readers eventually receive all the necessary information about the correct structure and can identify site 2 as the correct gap position. If the relativized position were marked by a resumptive pronoun, such local ambiguities would not arise,

making the retention strategy the most explicit one. The relative pronoun strategy assumes an intermediate position on the explicitness hierarchy as it provides more information than the gap strategy, e.g. by marking the relativizer for object case (i.e. *whom*), but it is still less explicit than the pronoun retention strategy as it does not distinguish relativization on direct objects from relativization on indirect objects. Keenan and Comrie observed that in languages that use more than one strategy, the application of the respective type can be predicted if we assume a ranking based on the degree of explicitness. Keenan and Comrie hypothesize that the further one goes down the AH, the greater the processing load, and therefore the greater becomes the pressure to use a more explicit strategy.

In light of the somewhat problematic status of *that* and the rather infrequent use of case-marked *wh-* relativizers, the "either or"-characterization of English relativization strategies seems somewhat problematic. A more felicitous characterization should acknowledge the fact that relative clauses in English retain their gap even when the pronoun strategy is applied. Let me repeat the (re-ordered) set of object relatives from (13) to (16) here as (29) to (32):

(29) The girl$_i$ ∅ you like __$_i$ so much unfortunately has a boyfriend.

(30) The girl$_i$ **that**$_{(i)}$ you like __$_i$ so much unfortunately has a boyfriend.

(31) The girl$_i$ **who**$_i$ you like __$_i$ so much unfortunately has a boyfriend.

(32) The girl$_i$ **whom**$_i$ you like __$_i$ so much unfortunately has a boyfriend.

The perspective taken here presents the competing forms in (29) to (32) as possible means to express the same thought in English. The competing forms vary with respect to the amount of information that is provided to signal the internal role. (32) is the most informative variant: it exhibits an overt *wh-* relativizer that signals both (a) that the structure to be processed is a relative clause and (b) that the head plays an object role in that RC (by way of morphological case marking). No such signals are present in (29), which delays the disambiguation of the structure until the point the main clause verb, *has*, is presented (until that point the structure

could be a topicalized declarative). The two remaining forms, *who* and *that*, occupy middle positions. Both signal an upcoming RC, but we may consider *who* to be a little more informative due to the fact that it is biased towards signaling a subject relative (even though it here occurs with an object relative). Given the little difference in role signaling potential between *who* and *that*, however, it appears more sensible to assume that the most important difference between *that* and non-case marked *wh-* relativizers has to do with signaling the defining/non-defining function of the RC in question (with *that* signaling intended referential restriction). Once we help ourselves to such a characterization of linguistic choices, we acknowledge that nothing hinges on a categorical decision on the syntactic status of *that* and we may view linguistic categories as fuzzy and probabilistic. We will pick up this idea a number of times as we go through our analysis (but cf. Bod, Hay, and Jennedy 2003 for a discussion).

A natural position to go from here is to assume that speakers tend to employ those forms which are most cost-efficient for the communicative task at hand: if the internal role is highly predictable and easy to process—i.e. if it is not too low on the AH—an overt relativizer is not required as such RCs impose relatively low processing demands (by any measure of processing demand) and it may hence be dropped. The less typical and more difficult the patterns become, as in the case of relativization on a genitive object, the more is gained by an more explicit form (and/or strategy for that matter). Linguistic communication is after all a type of purposeful human activity and in being so is subject to certain conditions of use. Among the conditions that have a strong impact on the shape of language are conditions on efficiency of the linguistic means that are employed to reach certain communicative ends. Linguistic communities tend to conventionalize those forms that allow them to express their intentions fast. A straightforward example of this tendency is the positive correlation of the sheer amount of linguistic material (say measured in phonemes) that is used to express a concept: frequently used concepts tend to be short, while infrequent concepts tend to be long (cf. Zipf 1935, 1949). Ellis and Hitchcock (1986) report a similar adaptive behavior in other domains, specifically the usage of artificial language). It is, hence, highly unlikely to find languages that have polysyllabic words for the concept of self reference. As usual, there are exceptions to this tendency: in many East Asian languages first

and second person pronouns are derived from nouns and may very well be polysyllabic. For example in Japanese the (most usual) first person pronoun, *watashi*, and the (most usual) second person pronoun, *anata*, are trisyllabic (cf. Martin 2004). Such examples, however, only seem to be counterexamples as Japanese personal pronouns encode more information, specifically social (distance) information. It is expected that extra information correlates with extra length. Similarly—coming back to relative clauses and relativization strategies—we might hold economic reasons responsible for the fact that many languages entertain more than just one strategy: while being the more explicit strategy, (an application of) the pronoun strategy also consumes more energy than (an application of) the gapping strategy as speakers need to produce more material to achieve the very same communicative goal. Without going into the details of what exactly should be the currency in such a trade-off (e.g. sheer amount of material that has to be produced, or the working memory load of a structure, etc.), it seems to be a intriguing idea to think of languages as tools that get shaped by their users until they become optimal, i.e. optimal means for communication. This general line of thinking has been proposed by a number of linguists: a recent and quite explicit proposal has been put forth as the "performance-grammar correspondence hypothesis" (Hawkins 2004). This hypothesis holds that more cost-efficient forms are preferred by speakers (and hence used more often), which in turn acts on their longevity. Dahl (2004) presents another book-length discussion of such issues focusing on the growth and maintenance of linguistic structures. Research into the historical development of language has long reported cyclic state transitions and continues to stress such dynamics in grammars (cf., e.g., Jespersen 1917 on sentential negation, Brinton and Traugott 2006 on aspectual marking for a very early and a quite recent example of such ideas, Feldman (2006: 77f3). describes that similar dynamics can be observed for systems over a wide range of time and complexity scales—from molecules to individuals to societies to species).

As a last note on the characteristics of English RCC, we should point out that relative clauses, just like all other dependent clauses in English, can also be non-finite, i.e. they can take the form of participial, to infinitival or bare relatives. Consider (33) to (35)

(33)	He is talking to a girl [resembling Jane]	-*ing* participle
(34)	The only car [being repaired] is mine	-*ed* participle
(35)	The next train [to arrive] was from Berlin	*to* infinitival

In summary, we have noted that English allows relativization on all positions on the NP Accessibility Hierarchy (on all six with the pronoun strategy and on all except genitive with the gapping strategy using *that* and on all except subject and genitive with a zero-complementizer). If we cross the structural dimensions along which relative clauses can differ, i.e. their internal and external syntactic properties, the choice of relativizer, and their finiteness, we are already left with [7 x 5 x 3 x 2—x = k] formally different subtypes.

The next section is geared to provide the reader with a convenient transition to the theoretical framework to be used in this study. It is thus best viewed as a philosophical bridgeover to the psycholinguistic stance presented in the sections to follow. That is to say that it is the goal of this "overture" to make plausible why the particular psycholinguistic stance was taken in the first place. We shall therefore argue on a rather general level and contain ourselves to only the most elementary references to prior research. The goal is to motivate certain concepts and ideas and not to recapitulate their development.

1.2 Overture: Some precursors and some prerequisites

1.2.1 Symbolization and mental states

[O]ur minds can have many possible belief states, and interactions between minds […] allow their states to become correlated.

Jeffrey R. Hanson

Human beings have the astonishing ability to act on each other's mental states via the transmission of linguistic forms. This transmission is not restricted to a specific modality: we can after all communicate via the auditory channel (i.e. via phonological forms) or the visual channel (i.e. via some orthographic writing system) or even the tactile channel (e.g. in Braille). In fact, we usually combine some (or even all) of these channels in face-to-face communication. During their cognitive development, humans establish *semiotic systems*—systems of signs—that exploit statistical regularities in the environment to link perceived forms with meanings (e.g. Barlow 2001). Following Langacker (e.g. Langacker 2000), we may call this association of forms and meanings *symbolization*. Whereas forms are directly observable and thus relatively easy to describe, meaning certainly is a notorious notion: we may, however, work from the assumption that the meaning of a form can be determined by the causal effects a perception of that form has on a language user's mental state or web of beliefs—to use Quine's metaphor in a slightly different context. We may say that black clouds *mean* rain because perceiving clouds of a particular shape and color is likely to cause in the experiencer (the formation of) the belief *that it is going to rain soon*. Of course, this is not to say that black clouds *cause* rain but rather that a *perceived spatio-temporal proximity* of seeing black-clouds (= event A) and experiencing rain shortly after (= event B), eventually causes in the experiencer (the formation of) a particular type of belief, which can be expressed in English as an *if*-conditional, say "If there are black clouds in the sky, it will rain

soon". In short, in this view, the meaning of a form can be characterized as the set of causal effects that perceiving that form has on the experiencer's mental state, or web of beliefs, which in turn may correlate with particular behavioral dispositions.

The question what kinds of things meanings (or concepts) are—ontologically speaking—must be addressed very carefully. There are good reasons to view meanings as something that is publicly held, i.e. something that belongs to language as a social entity. This view is implicit in many lexicographical or computer science treatments in which meanings are abstracted from large corpora. There is, however, the equally sensible view to think of meaning as something that belongs to the individual, i.e. something that belongs to language as a cognitive entity. In this view, meanings may be conceived of as *structured mental representations* (or *conceptual structures*), i.e. mental objects with semantic properties (e.g. content, reference, truth-conditions, truth-values, etc.). These objects describe (or are in some sense about) situations and components of situations, most notably objects and relations. Let me illustrate the general picture using an example from early stages of language development: a language user's early linguistic output (Brown's stage I; typical entry 18-24 month) comprises of two-word utterances such as *daddy sit*, *crayon big*, or *my teddy* (Brown (1973). Obviously, the structure of such utterances is rather simple and restricted to a limited set of semantic relations that ascribe some property to a single entity. However simple these structures may be, the fact that we can observe them surely adds to the plausibility of the idea that the child has established mental representations of objects in its environment and also a number of properties that are applicable to these objects. From these simple observations, we may hypothesize that the complete knowledge system of an adult, a person's web of beliefs, can—at least as a first approximation—be expressed in terms of (localist) semantic networks like the one presented as Figure 9. The details of such a system are largely irrelevant at this point. The crucial assumption made here merely is that humans store their knowledge of the world in a systematized way leaving very few parts isolated from the coherent, integrated main body of knowledge. The notion of thinking about knowledge systems in terms of semantic networks has been popularized by Ross Quillian and colleagues in the sixties (Collins and Quillian 1969) even though the general idea can be traced back to Aristotle (cf. Anderson and Bower 1973:9).

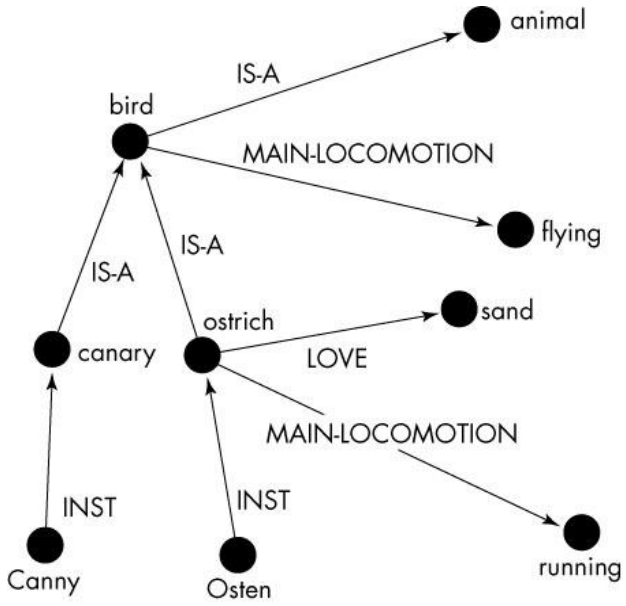

Figure 9: Partial semantic network (based on Collins & Quillian 1969)

The black dots, or nodes of the network, in Figure 9 stand for individual concepts, which, if this were a model of an individual's (encyclopedic) knowledge, represent things (or classes of things) that are part of his/her personal ontology. The arrows stand for relations or attributes that are true of the concept(s) involved. Hence, the network in Figure 9 encodes among other things the proposition 'that ostriches love sand' (LOVE(ostrich, sand)) and 'that ostriches are birds' (IS-A(ostrich, bird)). Of course, this personal ontology need not be restricted to things that are believed to actually exist in the "real world", but can very well also extend to other, fictional realities. So having the concept of PEGASUS does, of course, not imply that the thinker actually believes that the there is an entity in the real world that fits the description that defines the concept PEGASUS. Formally, these different planes of existence, or universes of discourse, can be modeled using possible world semantics (Hintikka 1962, Kripke 1963) or mental space theory (Fauconnier 1994, 1997). Of course, as humans have no direct access to anything like an objective reality, a person's concept of the real world is just another mental construct albeit a very important and particularly rich one, which is why I put the corresponding linguistic expression in quotation marks in the above.

Note that in addition to counting as evidence for a mental representation of objects and relations, we may take the very same fact, i.e. the fact that the child is able to produce

corresponding linguistic utterances in communicative contexts, as evidence for the idea that representations of perceived facts about the world have been successfully linked with representations of linguistic forms.

1.2.2 Linguistic units as processing instructions I: form to meaning

> *"The general idea is an old one, that any two cells or systems of cells that are repeatedly active at the same time will tend to become 'associated', so that activity in one facilitates activity in the other." (Hebb 1949: 70)*

The association of meanings (conceptual structures) with forms (linguistic structures), which is established and reinforced whenever humans experience and interact with their environment, eventually makes it possible to activate in the individual conceptual structures not only *directly*, i.e. via actual perception of a particular situation, but also *indirectly*, i.e. without any such direct experience of the objects that figure in a given situation. This indirect activation occurs when we categorize a stimulus/form to be symbolic[11], i.e. judge it to be an instance of a type of experience that stands for another type of experience. Linguistic forms certainly are associated with (or *symbolize*) particular types of situations (or components of situations), which is to say that linguistic stimuli certainly are at least potential symbolic stimuli.

From this standpoint we may say that learning a language necessarily involves learning all the conditions of usage of an expression, which is to say that learning a language is learning all symbolic links established by the conventions in a given speech community. This surely seems like an awfully complex task. But that—by itself—does not render it

[11] This hedged statement tries to account for the fact that nothing is a sign unless we interpret it so. This is just to say that I take it to be a fact that humans usually do not go out and assume that every single sensation they have stands in some kind of STAND-FOR relation to some other sensation.

empirically false. Logically speaking, nothing relevant—i.e. pointing in the direction of a required UG postulation—follows from the fact that this task seems hard to us. Whether or not we can actually get to the desired state of full linguistic competence by means of such envisaged inductive mechanisms clearly is an empirical question. And it requires much more specific explications of the learning mechanism than what I have presented here to even translate this statement into anything testable. At this point in our discussion, we may treat what was just stated in the above as a more or less plausible general picture. Fortunately, however, this is not the current state of the art: linguists, psychologist and computer scientists have already explicated and developed a lot of the necessary theoretical machinery and today we have good reasons to believe that grammar is in fact learnable without an innate linguistic module as envisaged by UG accounts (cf., e.g., Redington et al. 1993, Bates and Elman 1996, Jurafsky 1996, Seidenberg 1997, Christiansen and Chater 1999; Brants and Crocker 2000, Lapata et al. 2001, Pothos and Juola 2007). There are also good arguments in place that this task is not an intrinsically linguistic one but that it is rather much more general in nature: we exploit the redundancies and correlational structure of the perceived world in many cognitive tasks, first and foremost in vision. So, in some sense learning a language is not so special. In order to become competent language users, we need to detect a multitude of correlational relationships in the environment, in this case the ambient language. Linguistic expressions are usually not randomly uttered by the members of a speech community, but serve communicative purposes. It is this simple fact about language use that gives rise to the correlative structures in the ambient language, which serves as the body of data from which the learner extracts her knowledge of the communication system used by her speech community. To make things a little more tangible, let me illustrate the point on the basis of a simple example. We may for the sake of argument agree on the validity of the statement that the linguistic string "This is a cat" is more likely to occur in the presence of a cat (an actual cat or at least a depiction of a cat), then in the absence of a cat (or a depiction of a cat). If that were true—and this difference were pronounced enough, the experience of hearing an instance of the linguistic string *This is a cat* is likely to be positively correlated with the experience of, say, seeing a cat (or a depiction of a cat). Given our remarkable ability to recognize and store correlative information—which has been shown in many areas of human

cognition, above all in research into vision (cf., e.g., von Helmholtz 1925, Elder and Goldberg 1998, Rao, Olshausen, and Lewicki 2002, Geng and Behrmann 2006 and references therein), language learners will mentally represent the information encoding the tendency of the string *This is a cat* to occur in contexts where cat-like objects are present. It is this mechanism that allows us to eventually invoke in our addressees the concept of a cat-situation by merely presenting to our listeners the linguistic form associated with that type of situation.

In summation, perceiving a linguistic utterance (U) will cause in the hearer (H) the activation of a particular type of conceptual structure (M) that is associatively linked to a particular type of linguistic structure (F). In an ideal communicative scenario, the conceptual structure M constructed by H on the basis of her perception of U approximates the conceptual structure M' that S intended to invoke in H for purposes of communication.

This tentative approximation of the workings of linguistic communication must of course oversimplify the complexity of the phenomenon. But with the general ideas in place, we are now in a position to try and refine the picture. So far it suggests a rather direct activation of stored representations of situation types. One important point that we need to attend to in our attempt to refine our first sketch of linguistic communication is the creative aspect of language use, in particular the creative aspects of meaning construction. There cannot be any doubt in our minds that language users certainly are able to produce novel utterances that trigger novel conceptual structures. After all, we are able to develop new ideas, i.e. new types of conceptual structures, and communicate these ideas. In order to do so, a speaker's goal must be to select the set of linguistic forms best suited to cause in the hearer the desired effect, i.e. the construction of a conceptual structure of a particular type. Each unit within the linguistic structure can thus be seen as an instruction to the hearer's comprehension system of how to construe (or conceptualize) the situation described. It should be noted that understanding an utterance also involves grasping S attitude towards the expressed situation—roughly the intended speech act. For the present purposes, however, we may neglect these matters and focus on the (re-)constructions of syntactic and semantic/conceptual structures. Often there are various ways to describe a given situation that is presented to us in

consciousness and the choices of the linguistic forms contained in U, which S will eventually produce, is heavily constrained by the way S wants H to conceptualize the situation described. By choosing the appropriate linguistic forms, S can guide the way H is likely to construe the situation described. Note that the talk here about "wanting somebody to conceptualize something in a particular way" and similarly statements about "choices" should not be taken too literally here. It is very possible that many of the choices that are necessary for the production of a linguistic utterance are situated at a subconscious level. A speaker may not be aware of the fact that the situation she described with a ditransitive structure (*I gave Jon the letter*) could also be expressed using a prepositional dative (*I gave the letter to Jon*) with only slight changes in meaning. To the extent that one assumes that the term "choice" is synonymous to "an intentional act of choosing", which by definition has to be conscious, the talk of choices and similar notions, of course, becomes metaphorical. With this caveat out of the way, we may continue our discussion and compare the following examples in (36) – (39):

(36) Jack sprayed oil on the wall.
(37) Jack sprayed the wall with oil. # *spray-load* alternation
(38) I made a mistake.
(39) Mistakes were made (by me). # active-passive alternation

The examples in (36) to (39) present two pairs that involve a constructional choice. In both cases the members of a given pair of expressions have been claimed to be semantically equivalent (if only at the level of propositional content). If we conceive of meaning as something that goes beyond truth-conditional descriptions, we may describe fine-grained (maybe only probabilistic) differences that concern the preferred interpretations of the involved types. Examples (36) and (37) exemplify the *spray-load* alternation, in which one of the alleged alternants, namely (36), suggests that only parts of the wall are covered with oil. In contrast, linguists have proposed that (37) invites a holistic interpretation, in which the whole wall is covered with oil (cf., e.g., Levin 1993). Maybe more prominently—and

arguably also more heavily exploited on a conscious level—is the alternation between active and passive voice variants illustrated in (38) and (39). Again, even though these expressions may be analyzed as truth-conditionally equivalent, the variants are different with respect to their information structural properties to the effect that the active variant is dispreferred by speakers who wish to play down their role in whatever mistake was made.

Given these observations, we may say that a particular linguistic structure gets chosen by S if—metaphorically speaking—it promises to be a suitable means to invoke in H the conceptual elements necessary to construe the situation in the way intended by S.

As an interim conclusion, we may note at this point not only that linguistic forms are means that a speaker can employ to cause in the hearer the activation of a particular meaning in the sense above, i.e. the construction of a particular propositional mental object with certain semantic properties (content, reference, truth-conditions, truth-value, etc.), but S can also manipulate how the situation will be construed by choosing the appropriate linguistic stimuli. We may thus conceive of linguistic communication as an activity that involves designing (on the side of S) and interpreting (on the side of H) a series of processing instructions.

1.2.3 Linguistic units as processing instructions II: form to form

"Linguistic forms are themselves a part of the world within which the organism functions and to which it must adapt."

Bates & MacWhinney 1989

The examples in (36) to (39) were meant to illustrate that a speaker's linguistic choices influence the way in which an addressee will construe the situation talked about. While this is a very important idea, which has attracted a lot of attention in cognitive linguistic research, there is yet another—in some sense even more fundamental—way in which it is fruitful to

think of linguistic forms as processing instructions: not only does the choice of forms influence construal operations, i.e. the conceptual structures that will be built by H in language comprehension, but each (perceived) unit also provides H with information relevant for effective parsing, i.e. information needed for the construction of the syntactic structure of an utterance, which H arguably needs to (re-)construct in order to arrive at an adequate semantic interpretation of U.

Recent research into human sentence processing has amassed a growing body of evidence for the idea that speakers make use of knowledge of various types of associative relationships among linguistic units to anticipate what they are about to perceive and hence to speed up linguistic communication. The anticipatory character of language processing, which relies heavily on the usage of complex prefabricated units, is of critical importance for the approach taken here. A central goal of this study is to make a case for a relevant amount of variation in the formal choices made by language users to express a particular idea are predictable from the processing demand associated with these formal variants, to the effect that whenever possible speakers will prefer the low-cost variant. Let us illustrate this idea on the basis of a formal variation we have already mentioned earlier, namely the omission of an optional relativizer. Consider the examples in (40) and (41):

(40) The man **that** John likes hates Bill.
(41) The man Ø John likes hates Bill.

These sentences are not only truth-conditionally equivalent but there is also no reason to believe that the two forms evoke different conceptualizations of the asserted scenario. Even those linguists who subscribe to a strong Bolingerean position, which rejects semantic equivalence across different forms in principle, would certainly consider a pair like the one in (40)-(41) to be the closest approximation of true synonymy. There is, however, an obvious difference between the two expressions: utterances of the type exemplified in (40), i.e. with an overt relativizer, exhibit a lesser degree of ambiguity and can therefore be considered more informative. In contrast, the reduced version in (41) is locally syntactically ambiguous

at the time *likes* has been perceived (the sentence could instantiate a topicalized declarative, i.e. OSV order). So, in some sense—one that concerns the degree of informativeness of an expression type, (40) is clearly the "better" signal. It carries more information and thus imposes fewer problems on the comprehending system, viz. it requires no "guessing" in terms of structure building at *likes*. On the other hand, we must note that (40) also requires S to spend more energy on his utterance: after all it contains an additional linguistic element that (arguably) does not contribute to the conceptual structure on the side of the hearer. More generally, as far as communicative purposes are concerned, it introduces greater costs for the same gain.

1.2.4 Conventional patterns as routinized instructions

"[N]o one is able to persuade me that the correctness of names is determined by anything besides convention. No name belongs to a particular thing by nature, but only because of the rules and usages of those who establish the usage and call it by that name"

Plato, Cratylus

For trivial reasons, successful communication is more likely when S and H share the set of invoked associations of forms and meanings. This shared knowledge in turn is more likely when the forms employed by S to convey her communicative points are *conventionally* used to evoke a particular conceptual structure. The notion of conventionality is routinely invoked in linguistic discussions often without a clear explication of that term's meaning, which suggests that there is no need in defining or explicating its meaning. At closer inspection, however, it turns out to be a quite vexed notion, whose meaning is notoriously difficult to pin down. The notion of convention(ality) continues to puzzle sociologists and philosophers and questions that are immediately raised as soon as the notion is brought to bear on any phenomenon include: How do conventions arise? How are they sustained? How do we select between alternative conventions? Why should one conform to convention? What social good,

if any, do conventions serve? How does convention relate to such notions as rule, norm, custom, practice, institution, and social contract? For the purposes of this discussion, I will follow David Lewis (Lewis 1969), who analyzes convention as an arbitrary, self-perpetuation solution to a recurring coordination problem. The solution is self-perpetuating because no member of the (speech) community has reason to deviate from it, given all others conform. Lewis describes what it means for a regularity to count as a convention as follows (Lewis 1969: 76):

> "*A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a convention if and only if it is true that, and it is common knowledge in P that, in any instance of S among members of P*,
>
> *A. everyone conforms to R;*
> *B. everyone expects everyone else to conform to R;*
> *C. everyone has approximately the same preferences regarding all possible combinations of actions;*
> *D. everyone prefers that everyone conform to R, on condition that at least all but one conform to R;*
> *E. everyone would prefer that everyone conform to R', on condition that at least all but one conform to R',*
>
> *where R' is some possible regularity in the behavior of members of P in S, such that no one in any instance of S among members of P could conform both to R' and to R*."

As linguistic behavior is usually co-operative and given that we have characterized language users as trying to maximize the efficiency of their communication systems, we would expect that conventionally used forms have proved to be successful vehicles to bring about the desired mental states in the hearer. We may presume that for something to prove to be useful

means that it has been extensively tested, so that a conventional form, in addition to being a successful means to communicate some meaning, also is rather frequent. We will argue in subsequent chapters that there are good reasons to believe that frequent stimulus types, e.g. linguistic structures, are easier to process than infrequent ones. All we need to acknowledge at this point is the close connection between social convention, cognitive processing, and linguistic form.

1.3 Chapter summary

In summation, we have seen that in certain situations, speakers apparently have a choice between two semantically equivalent forms: one that contains less linguistic material but introduces more structural uncertainty and another one that contains more linguistic material but is more explicit with respect to structure building processes. From this perspective, it is natural to conceive of the speaker task, designing an utterance that leads to the felicitous reconstruction (on the side of the hearer) of the intended speech act, as something that requires S to weigh benefits against costs and decide on the basis of this decision what form to produce. Over time language users will conventionalize forms that have proved to be communicatively effective and those forms will become more frequent than less effective forms. Frequent forms are more typical and easier to anticipate, which has direct consequences for the processing demand associated with a linguistic structure.

This concludes our little overture and with it the precursors and prerequisites that I hope help make the general perspective taken here more palpable. In the part to follow we will provide a more detailed statement of the theoretical framework used for this study from both a linguistic and a psychological point of view.

2 Towards a theoretical framework of the right kind

The last twenty years have led to slow but steady changes regarding many core assumptions in the study of language, both theoretically and in terms of methodology. We have seen new answers to central questions including what language might be, how it should be studied, and what a linguistic theory should be able to account for. Specifically, instead of viewing language as an autonomous cognitive system, best studied by means of introspective data, and attempting to develop a minimal, i.e. most economical, description of what is referred to as 'core grammar' (cf., e.g. Chomsky 1995), many researchers today view language as deeply grounded in general cognition, best studied using empirical methodologies, and attempting to develop a maximalist grammar, i.e. a grammar describing "the full set of particular statements representing a speaker's grasp of linguistic convention, including those subsumed by general principles" (Langacker 1987: 46). Close inspection of actual language use has revealed the gradient nature of virtually all linguistic categories and eventually has led researchers to cast into doubt the usefulness of categoricity as a desirable property of linguistic theories (cf. Bod, Hay, and Jennedy 2003). Hence, most researchers in this approach highlight the probabilistic nature of linguistic knowledge and develop corresponding non-categorical theories.

This section will elaborate on two important trends in recent linguistic theorizing: the tendency of grammars to be *sign-based* (or *construction-based* in the sense explicated below) and the tendency to develop a perspective on natural language that is *usage-based*, i.e. a perspective that takes serious the idea that linguistic knowledge is shaped by language use and hence intimately tied to the cognitive operations governing the learning and online processing of symbolic structures. In addition to the cognitive aspects that act on the shape of grammars, there certainly is a growing trend to integrate more directly ideas from research into how languages change over historical time (cf., e.g., Hopper and Traugott 2003 and references therein). This last perspective, however, the development of language over historical time, is complex enough to deserve more than just a few side remarks and will be excluded from our discussion here for the simple reason that including it would break the

mold. So, while a diachronic perspective and an account of language change certainly is necessary for a full understanding of why grammars are the way they are, we will focus here on the synchronic state of an (idealized) individual grammar, i.e. a particular kind of implicit knowledge, and effects that language use has on that knowledge.

Questions regarding how humans process symbolic structures, most notably natural language, are investigated in many areas of cognitive science most notably maybe in cognitive psychology, cognitive neuroscience, computer science, the philosophies of language and mind, and of course linguistics. One of the most imperative and demanding challenges for the field of linguistics is the development of a theory that not only *describes the linguistic system* adequately but that is also *cognitively* or *psychologically plausible* so that it can be integrated into the theories of neighboring fields. In the view presented here, the linguistic contribution to the interdisciplinary attempt to understand human cognition should be a theory that presents an adequate characterization of the human linguistic system and how its elements and relations are learned, stored and processed in actual usage. This is by no means the only view that one might have regarding the goal of linguistic inquiry and the status of its theories. It may not even be the dominant one. To appreciate the diversity of conceptions regarding the goal of linguistic theorizing, we may divide the class of theories into ones that aim to be purely *descriptive* and ones that aim to be *explanatory*. Following Dryer (2006), we may say of *descriptive theories* that they are theories of "what languages are like". In contrast, *explanatory theories* aim at providing more than just a description of what languages are like and provide an answer to the question of *why languages are the way they are*. In contrast to American structuralism and also many contemporary typological theories (cf. Dixon 1997), generative theories in the Chomskyan tradition subscribe to the idea that a linguistic theory can—and indeed should—go beyond mere description and provide explanations for the attested forms of grammars. In Chomskyan linguistics, a linguistic phenomenon is explained when the observed linguistic facts have been related to the principles of *Universal Grammar*, which may be thought of as an innate mental capacity that comprises of a limited set of rules for organizing language.

In accordance with generative linguistics, the present account assumes that a linguistic

theory can—and in fact should—provide more than a mere description of what a language is like and should indeed provide some account of why it is the way it is. However, the present account departs from mainstream generative views in that it does not revert to Universal Grammar, i.e. a mental capacity postulated in addition to the mechanisms responsible for the production and comprehension of language. Instead, it aims at providing explanations of why languages are the way they are that follow rather directly from the way humans learn, represent and process symbolic structures. The merits of such an approach are rather obvious as they follow from very basic principles of scientific practice, most notably Occam's Razor, the principle of parsimony that requires us to prefer a simpler theory over one that is more complex, given that both can account for the phenomena under scrutiny. In this context, it is helpful to point out that the theoretical motivation for the postulation of innate knowledge, i.e. Universal Grammar. Proponents of the generative account (e.g. Chomsky 1980) have argued that the process of language acquisition is utterly mysterious if no innate knowledge were postulated as—so the argument goes—natural languages are far too complex to be learned on the poor basis of available evidence, i.e. the total amount of linguistic stimuli a child receives. The argument usually assumes the following form. Given the premises P1-P3, humans must have some form of innate linguistic capacity that provides relevant knowledge used in addition to the positive evidence they get from the linguistic input.

P1. There are patterns in all natural languages that cannot be learned by children using positive evidence alone.

P2. Children are only ever presented with positive evidence.

P3. Children do learn correct grammars for their native language.

Of course, the conclusion drawn does not deductively follow form the truth of the premises. In fact the UG hypothesis constitutes what philosophers of science call an *inference to the best explanation*, i.e. the result of an abductive reasoning process in which scientists elect that hypothesis which, if true, best explains the observational data. If we assume the truth of P1-

P3, there certainly is a problem (*Plato's problem*) and the UG hypothesis certainly offers a solution. There is, however, a growing skepticism regarding the empirical adequacy of P1 (and to some extent also P2, cf. Stefanowitsch 2006), which came to be known as the *poverty of the stimulus* hypothesis. Research into (artificial) neural networks and connectionist learning has generated a growing body of evidence that languages are in fact learnable without a postulated innate knowledge source guiding the process (explications of the arguments for UG-guidance can be found for example in Pinker 1979, Pinker and Prince 1988; Crain 1991; for a rebuttal cf. Rumelhart and McClelland 1986). Consequently, if the premise of insufficient input is at fault, the Chomskyan inference to the best explanation, i.e. the existence of Universal Grammar, lacks sufficient theoretical motivation and should, therefore, be discarded. This brings us back to the argument from theoretical parsimony: while there can be no doubt that we need a theory of language processing, i.e. of linguistic performance, simply because human processing of natural language obviously is an empirically real phenomenon, it appears to be less obvious whether in addition to this we need an account of linguistic knowledge, i.e. linguistic competence, as an integral part of our cognitive system. Ideally, our theory of linguistic performance would directly predict the facts described by a competence grammar. This, however, is not to say that we should not entertain a theoretical framework that provides an elegant, higher level description of the linguistic facts, if only not to get lost in the massive redundancy we can expect an explicit performance theory to exhibit. But such a theory would not describe any psychologically real capacity, much less a mental module.

In the view presented here, it is argued that a linguistic theory of the right kind should be able to adequately describe the linguistic phenomena under investigation and, if possible, it should be able to explain why it is the way it is without stipulating innate knowledge of the type envisaged in UG accounts. At this point, a little qualification is in order. It is not the aim of the present study to provide evidence against the existence of Universal Grammar. Such an endeavor would first of all require a precise explication of the meaning of the term itself. At any rate, a single, universally accepted extensional definition of the term is rather difficult to come by as various explications have been suggested in the literature—either explicitly or implicitly. Even if we follow influential proposals and take UG to denote some "specified

prespecification" or "innate structure" (cf. the discussion in Jackendoff 2002) or a set of "algorithms designed to acquire the grammatical rules and lexical entries of a human language" (cf. Pinker 1996), we would still need to spell out the details of that structure and/or these algorithms in order to decide on the issue of its cognitive reality. This, however, not only exceeds the scope of the discussion presented here it, but it would also certainly require more than observational data that can possibly be obtained from a corpus of present-day English. The discussion here is solely provided to make explicit the central theoretical assumptions made throughout this study. The present view is committed to the idea that the *why*-question can—and in fact needs to—be addressed from (at least) two angles: first, we need an account of how the grammar of a language develops, both as a cognitive phenomenon at the level of the individual, i.e. in language development, and also as a social phenomenon (or cultural artifact) within the language community, which changes over historical time. And second, we need an account of how the system—once it has been acquired—is used in actual communicative contexts. The view taken here argues that facts about language representation and processing will be relevant for both of these perspectives.

This remainder of this chapter will discuss the merits of the properties of being sign-based and being usage-based and in doing so present the theoretical framework that underlies the study. In need of a label I shall refer to this framework as a 'usage-based cognitive construction grammar'.

2.1 The merits of being sign-based

The family of construction grammars (CxG, Fillmore 1988, Fillmore and Kay 1993, Lakoff 1987, Goldberg 1995, 2006) is an emerging body of linguistic theories that are based on the notion of a *construction*, which essentially corresponds to the notion of a sign, i.e. a pairing of form and function. In her seminal monograph, Adele Goldberg has defined the notion of a construction as follows:

> *"C is a construction iff$_{def}$ C is a form-meaning pair <F_i, S_i> such that some aspect of F_i or some aspect of S_i is not strictly predictable from C's component parts or from other previously established constructions."* (Goldberg 1995:4)

More recently this definition has been extended to the effect that in order for something to count as a construction, the requirement of non-compositionality (=lack of predictability) no longer is necessary (while still sufficient). In her later explications of the constructionist view Goldberg writes:

> *"In addition, many constructionist approaches argue that patterns are stored even if they are fully predictable as long as they occur with sufficient frequency"* (Goldberg 2003: 220)

This extension is strongly embraced here for two reasons, first it de-emphasizes potential problems linguists/semanticists (including the author) may have with the notion of non-compositionality and what exactly is meant by some aspect of F_i or S_i to be "not strictly predictable from C's component parts" but secondly—and this aspect is more relevant for our discussion—it highlights the relation between such constructionist approaches and long-held ideas about chunking in memory research (cf. Miller 1956). Specifically, the notion of a partially-filled (or partially specified) construction will be adapted later in order to derive predictions about processing demands associated with particular sub-types of RCCs.

Although different versions of construction grammars differ in many details (cf. Cruse and Croft 2004 for a discussion), what they have in common is the idea that constructions make up the "stuff" grammars consist of, or as Goldberg puts it "the network of constructions captures our language *in toto*, i.e. it's constructions all the way down." (Goldberg 2006:18). It follows that in CxGs no principled distinction is made between lexicon and grammar and, consequently, there is no separate set of operations on the components of grammar.

Constructions do, however, vary in terms of their complexity and degree of abstractness. Examples are given in Table 1.

Table 1: Examples of constructions, varying in size and complexity (From Goldberg 2006:5)

Construction type (traditional name)	**Example**
Morpheme	*pre-*, *-ing*
Word	*avocado*. *anaconda*, *and*
Complex word (partially filled)	[N-*s*] (for regular plurals)
Idiom (filled)	*going great guns*, *give the Devil his due*
Idiom (partially filled)	*jog* <someone's> *memory*
Covarying Conditional	*The* X*er the* Y*er* (e.g. *the more you think about it the less you understand*)
Ditransitive le object)	Subj V Obj_1 Obj_2 (e.g. *he gave her fish a taco*)
Passive	Subj Aux VP (PP_{by}) (e.g. *the armadillo was hit by a car)*

The theoretical pressure motivating a construction-based view on language has been discussed in details elsewhere (cf. Fillmore et al. 1988, Lakoff 1987, Goldberg 1995) and shall not be repeated here. Suffice it to say that the merits of such a position have by now been recognized in virtually all domains of language research including early development (e.g., Tomasello 1992, Diessel 2004), second language learning (Ellis 1995, Haberzettl 2006), theoretical linguistics (e.g. Goldberg 1995, Müller 2006), historical linguistics and grammaticalization theory (e.g. Hopper and Traugott 2003, Diewald 2008), cross-linguistic

typology (Croft 2001, Stolz 2006), discourse analysis and interaction (Hopper 2001, Deppermann 2008), linguistic pragmatics (Stefanowitsch 2003), computational linguistics (Jurafsky 1996, Narayanan and Jurafsky 2001), and language processing (Bergen and Chang 2005, Wiechmann 2008a). From the ever-growing set of CxG variants that may serve the purpose of providing a linguistic background for the present study, it appears most felicitous to look out for a sign-based grammar that aims at psychological plausibility.

A sign-based theory that does emphasize its ambition to be strictly based-on general cognitive abilities is Ronald Langacker's Cognitive (Construction) Grammar (Langacker 1987, 1999, 2008). Langacker's proposal enjoys a privileged position not only because a) many of the key notions employed here—such as the notion of a *schema* and *entrenchment*—have found their way into linguistic theorizing through Langacker's expositions (Langacker 1987, 1999), but particularly because b) it is the designated goal of this approach to "characterize those psychological structures that constitute a speaker's linguistic ability, i.e. his grasp of linguistic convention" (Langacker 1990:263). To this end cognitive grammar is explicitly designed as a maximalist, non-reductive, and bottom-up description of linguistic knowledge and is also the framework that introduced the notion of a usage-based model, which we will have a closer look at in § 2.2.. Arguably more than any other framework mentioned so far, Langacker's cognitive grammar focuses on certain properties a desirable theory of grammar should bear. First, linguistic knowledge is conceived of as procedural rather than declarative in nature. Second, the units of grammar, i.e. constructions, are viewed as "thoroughly mastered structures" (Langacker 1999:15), which speakers can activate as preassembled wholes. This is especially relevant for our discussion of complex constructions like RCCs. In cognitive grammar, complex constructions are described as cognitive routines, which is an excellent term to express the anticipatory character of sentence processing. And third, it embraces a view on category structure, the network conception, which is another mainstay of the cognitive linguistic enterprise (cf. Lakoff 1987). The network conception draws from many ideas proposed by researchers that helped develop the prototype view on categorization, which brought to light so called prototype (or typicality) effects, i.e. effects that pertain to the subjective centrality of membership of particular examples for a given category (Rosch 1973, 1978). The resulting stance towards categorization departs from

classical (Aristotelian) conceptions in that it entertains a graded notion of category membership, in which instances of a category vary in terms of their representativity of that category so that some members are usually more central to the category than others. A category is defined in reference to a prototype, which can be understood as a schematized representation of typical instances. The more a given instance corresponds to this idealized schematic representation, the more likely it is to be judged to be central or typical for that category. As we will argue in subsequent sections, English RCC can be viewed as a set of (discourse-) functionally specified sub-constructions, which exhibit their own prototype effects and network structures. However, instead of using a prototype theoretical perspective on categorization, the present study will characterize all processes that involve categorization on the basis of an exemplar-based view, which can handle typicality effects just as well. Although we will later explicate how clusters of similar structures pertain to processing difficulty, it may be helpful at this point to foreshadow an important suggestion namely that the processing demands of a particular type of structure, say a RCC, can be conceived of as a function of that construction's similarity to a strong exemplar. We shall return to this issue in § 2.3 and continue our discussion with a closer look at some of the key properties of our constructionist theory of choice, viz. Langacker's cognitive grammar, and how regularities in language are handled in this account. This is best achieved by way of contrasting the here presented view with a more traditional conception, in which regularities are captured by means of rules. We will see that the cognitive grammar treatment, which solves the problem via the concept of schemas, is much friendlier to an empiricist stance towards the nature of linguistic knowledge.

2.1.1 Regularity in language: rules and schemas

Linguistic processes develop during language learning to embody

all but the most effete rules of grammar in automatic operations

Bock and Garnsey 1997

Regardless of whether we assume the epistemological viewpoint of rationalism, as Chomskyans do in their attempt to uncover innate linguistic knowledge, or empiricism, which rejects such innate knowledge, we still need to face the apparent fact that language use is regular, i.e. in some sense rule-governed or at least rule-describable. To be rule-governed (or rule guided) is for a system to have a representation before it (consciously or unconsciously) and try to match its behavior to that specified by that rule. In contrast, to be rule-describable is for a system to act in such a way that its behavior is describable by the statement of some relevant regularity. Obviously, the former implies the latter (as every system that is rule-governed ipso facto is also rule describable). So, strictly speaking we should have the contrast to be one between rule-governed and merely rule-describable systems. To the extent that this issue is relevant for the processing of language, it is relevant to be precise about what we mean by that, i.e. what—if anything—we take a linguistic rule to be.

In early proposals of the Chomskyan framework (Chomsky 1965), linguistic rules—such as phrase structure rules—clearly were not meant as procedures through which the human mind produces and understands linguistic utterances. Instead, linguistic rules, were viewed as elements of a mathematical representation of the innate knowledge that humans have about language. Specifically, they are productions in the sense used in computer science that specify a symbol substitution that can be recursively performed to generate new sequences of symbols. A grammar of a language, then, is the description of the set of all and only those strings that are grammatical in a particular language.

The details of an adequate mathematical representation of English grammar in the sense describe above need not be of further concern. What is important for the present purposes though is the fact that the mental grammar—i.e. competence or I-language (Chomsky 1986)—has also been portrayed as something that is cognitively real and that stands in some intimate relation to the mechanisms of language production, comprehension, and maybe most prominently language acquisition. For example, linguistic rules have been considered default operations involving abstract symbols. A default operation is "an operation that applies not to the particular sets of stored items or to their frequent patterns, but to any item whatsoever, as long as it does not already have a pre-computed output listed for it" (Marcus et al. 1995: 192). An alternative kind of linguistic rules are redundancy rules, which describe language regularities that are limited in scope, for example semi-regular inflections restricted to a small set of phonologically similar verbs in English past tense morphology (Jackendoff 1975). Rules and abstract symbols suggest an algebraic view of linguistic knowledge (Boole 1854, Marcus 2001). However, many current models of language (on the "implementational level" in David Marr's sense (cf. Marr 1982), viz. neural network models of language processing, "do not rely in any obvious way on rules" (cf. Plunkett & Marchman 1991:44). In light of these different conceptions of regularity—and their entailed theoretical commitments—, we need to be explicit about how the observed regularities in language can come about.

In the framework of Cognitive (Construction) Grammar (Langacker 1987, 2008, Goldberg 2006) regularities are captured by means of schemas, where the notion schema is explicated as

> *[…] a coherent, integrated structure comparable in most respects to those which support its extraction. A schema's internal organization is precisely parallel to that of the semantic, phonological, or symbolic structures it schematizes, thus reflecting whatever commonality they exhibit. It does however abstract away from their points of divergence, being neutral or less specific in regard to each; overall, then, it is characterized at a lower degree of precision and detail.* (Langacker 1990b:2)

The present account will follow Langacker and use this notion of a schema as a crucial element in the mechanisms used in language representation and processing.[12] It is assumed that the processing of a RCC is strongly influenced by top-down activation spreading from very general (coarse-grained) schemata that correspond to (abstract) configurations in the sense briefly discussed in introductory section. A more detailed discussion of the relationship of schemas in the sense of cognitive (construction) grammar and 'highly entrenched configurations' this study aims to uncover will be provided as the study unfolds. However, the example presented as Figure 10 should help getting hold of the notion.

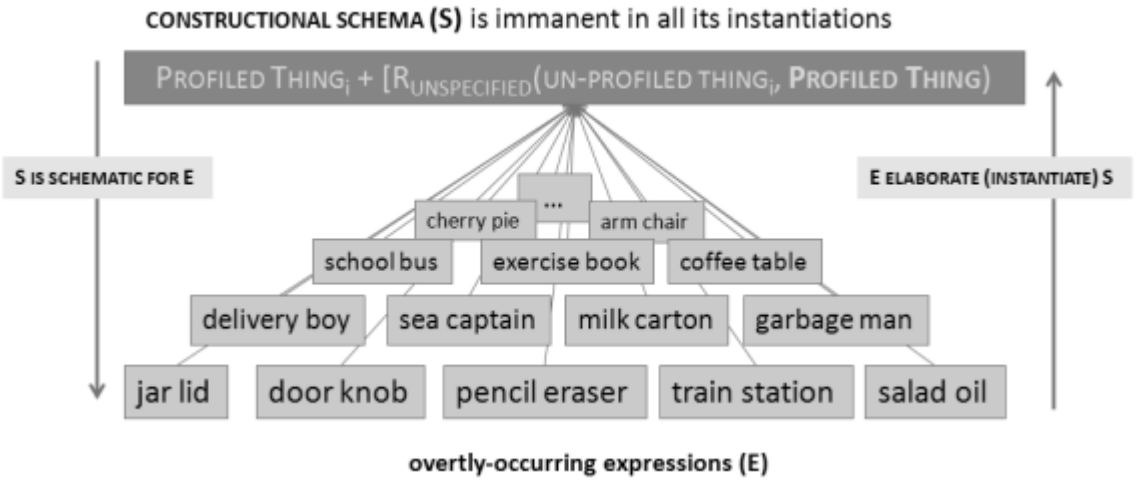

Figure 10: Example of a regularity (endocentric compound) in CCG

Figure 10 shows a description of a constructional schema of endocentric compounds in English on the very top (dark blue box) and a set of its instantiations below. As indicated by the arrows, the relationship between a schema S and its instances E can be described from both directions: we may equivalently say that S is schematic for E or E elaborates S. So a

[12] The present account is also sympathetic to the idea of fragment-tree (or subtree) as used in Data Oriented Parsing (DOP; Bod 1998). The DOP framework tries to integrate rule-based and exemplar-based approaches to language processing by viewing them as end-points on the same distribution.

schema is an abstraction of a set of expressions that share some set of properties, which in turn is the same as to say that the schema is immanent in all its instantiations. Note that even though this example shows a relationship between maximally specific constructions, overtly-occurring expressions, and a rather abstract semantic description, we should not infer that this necessarily has to be the case. As indicated above (cf. Table 1), we may very well have any number of schemas of intermediate degree of abstractness, such as partially lexically specified constructions.

So more generally speaking, we may formulate the cognitive linguistic view on regularity as follows:

> **REGULARITY IN COGNITIVE CONSTRUCTION GRAMMAR**
> A linguistic string E is well-formed (or licensed by the grammar), if it is an elaboration of a (set of) constructional schema(s) present in the grammar.

Note that this formulation does not depart in any obvious way from a formulation that makes reference to rules. The important difference is that in contrast to an account that makes reference to an inborn set of grammatical principles or rules, schemas, qua being abstractions of actual utterances, are established in a bottom-up (data-driven) fashion.

Under the assumption that linguistic knowledge is best characterized as a system of signs, which is structured with respect to relations of elaboration/instantiation so that more schematic construction subsume more specific ones, we can—using the same set of postulated mechanisms—describe the processes involved in language comprehension regardless of whether we are dealing with lexical items, idiomatic structures or syntactic structures: the ease of processing of a linguistic expression will—no matter what type of expression we are dealing with—always be a function of the ease of accessing the construction from the mental "construct-icon" to borrow Dan Jurafsky's term (Jurafsky 1993). This ease of accessing units from a structured inventory of signs is in turn is

predictable from the construction's frequency. As a first approximation of how frequency may act on the processes of mental, we can assume that each successful access to a construction makes future accessing to that construction a little easier than it has been prior to that access. Similarly, if a pattern to be processed is complex, i.e. consist of a sequence of units, it will become easier to process with each successful processing event of the same type. If processed often, the sequence will eventually become more and more unit-like and processing the structure will become easier the more unit-like the complex pattern becomes. Langacker describes this automation (or automatization) process, which he calls *entrenchment*, as follows

> **ENTRENCHMENT (EXPLICATION 1)**
>
> *Automatization is the process observed in learning to tie a shoe or recite the alphabet: through repetition or rehearsal, a complex structure is thoroughly mastered to the point that using it is virtually automatic and requires little conscious monitoring. In [cognitive grammar] parlance, a structure undergoes progressive entrenchment and eventually becomes established as a unit.*
>
> (Langacker 2008:16)

We will return to the effects of frequency on representation very shortly in § 2.2 when we discuss in some more detail what it means for a model of language to be usage-based and why this is desirable property in the first place. But before we focus on usage and frequency, a few more words should be said about the constructionist perspective and why it is so attractive for accounts that attempts to bring together strands of linguistic and psycholinguistic theorizing.

2.1.2 Constructions and the uniform representation of linguistic knowledge

One of the things that make construction grammars particularly interesting as a class of linguistic theories to be integrated into psychological accounts of language is its compatibility

with recent treatments in psycholinguistic theories regarding the nature of linguistic knowledge. Recent research into language representation and processing has continuously adjusted their characterizations of lexical and syntactic representation and disambiguation to the effect that there is now a growing consensus on the idea that those representations and processes are in essence very much the same (cf. MacDonald et al. 1994 for an overview of the discussion, Jurafsky 1996 for a computational perspective that take advantage of the merits of the constructionist perspective). So, construction-based linguistic theories and psychological accounts of language representation share the view that the elements of lexicon and grammar (if we wish to keep this distinction as a manner of speaking) are not so different after all. Let us briefly have a look at what may be referred to as the standard picture, i.e. the pre- constructionist view, and how it relates to the view advocated here.

The standard view on language processing, which dominated psycholinguistic research until the mid-nineties, treated the lexical and syntactic processing as orthogonally different (cf. Table 2).

Table 2: Standard view on ambiguity resolution

Properties of Ambiguity Resolution Systems	
Lexical Ambiguity	Syntactic Ambiguity
Multiple alternatives initially considered	Single alternative initially considered
Parallel processing	Serial processing
Capacity free	Capacity limited
Context used to select appropriate meaning	Context used to confirm analysis and guide reanalysis
Meanings are stored and accessed	Syntactic structures constructed by rule

Much of the research—in fact still is—based on the idea to focus on situations where the human processing system is expected to "have trouble", study the systems behavior in these situations and deduce from the observations made its general architecture and mechanism. Situations that meant "trouble" for the system were essentially evoked by means of having subject perform two tasks simultaneously (to see if they tap into the same resource and if so how much of that resource was spent on a given task) or feed the system with (locally) ambiguous structures and study how it would resolve these ambiguities. We shall restrict ourselves here to a brief glimpse at the latter, i.e. ambiguity resolution. Starting in the late seventies (Swinney 1979, Tanenhaus et al. 1979, Altmann 1998), the process of lexical ambiguity resolution was described on the basis of multiple access models, i.e. models in which common (or even all) meanings of an ambiguous word are looked up in the lexicon in parallel. Contextual information was assumed to be used very early in the process to select the appropriate meaning form a set alternatives (and suppress these alternatives). In contrast, syntactically ambiguous structures were resolved using a multi-step procedure that had different sources of information enter the processing at different times. These different views about lexical and syntactic ambiguity resolution are motivated and derived from assumptions about the types of knowledge involved in each domain (Frazier and Fodor 1978), and the notion that language processing is accomplished via a serial operation of autonomous processing modules (Fodor 1983). So, in the standard view, lexical ambiguity involves accessing items stored in some mental lexicon, whereas syntactic ambiguity resolution involves constructing structures on the basis of a rule governed system (cf. § 2.1.1). Operations within that system place demands on working memory and attentional resources, which presumably are capacity limited. Even though the details may of course differ across accounts, the resolution process is typically characterized as follows: the comprehension process involves (at least) two stages. During the first stage a phrase structure representation of the input is constructed based on syntactic category information (POS information) only. Very general—and hence potentially highly automatic—parsing heuristics like *minimal attachment* and *late closure* (Frazier 1987) are used to identify and generate minimally complex structures, which are initially preferred. After the completion of the initial phase of

parsing, the result of the structural description is passed on to other processing systems which confirm or disconfirm the interpretation and, if necessary, initiate a stage of reanalysis.

Historically, there has always been a strong link between models of parsing and language comprehension and linguistic theorizing at that time. As syntactic theories put more and more emphasis on lexical representations (cf. Chomsky 1970; Jackendoff 1975), psycholinguistic research, too, supplied more and more evidence for a parsing mechanism that is guided by lexically specific information (cf. Jurafsky 1996 for an overview). Interestingly, the alleged difference between lexical items and grammatical rules was also called into question from another angle. Not only was grammar becoming more lexicon-like. Units in the lexicon, in some sense, were also becoming more grammar-like in that their mental activation involved more computation. This is to say that with the advent of parallel distributed processing and spreading activation accounts (Rumelhart and McClelland 1986), it has been suggested that word meanings are not just accessed from a static storage, a mental lexicon, but rather are computed as part of the recognition process (Barsalou 1987). The meaning of a word in this view is represented as a pattern of activation over a set of units representing semantic primitives and these units may participate in the representation of many words (cf. Nilsson 1998 for a discussion of the shift to sub-symbolic approaches). Hence, meaning of a word is not accessed but constructed (i.e. in some sense computed) and different patterns are activated in different contexts. While the present view is certainly well-disposed to the general idea of dynamic lexical meanings, we should at this point not dismiss the idea of accessing a pattern from memory altogether. After all, the kind of computation in (artificial or actual) neural networks is different enough to anything that corresponds to the processes described in what I have labeled the standard view. Characterizing the process involved in entertaining a lexical concept in an activation spreading account as a case of computation may thus be misleading. It certainly is misleading, if computation is conceived as symbol manipulation. The literature on issues pertaining to the question to what extend particular types of connectionist/neural networks can be described as implementing abstract algebraic rules is both rich and gets quite technical very soon (cf. Fodor and Pylyshyn (1988), Marcus (2001) for good entry-points into the discussion). What is important here, however, is that psycholinguistic theorizing about the processes involved in language comprehension has

been changed very noticeably. The dominant view in linguistics until very recently has been that lexicon and grammar are two very different types of entities with the latter operating on the former was shared by psycholinguists and shaped their ideas about lexical and syntactic disambiguation. The growing body of experimental evidence that the disambiguation processes were not so different after all (cf. MacDonald 1994, 1999, MacDonald and Seidenberg 2006) was of course quite puzzling for proponents of a view that treats grammar and lexicon to be ontologically distinct. It is, however, very natural from the perspective of a sign-based theory of language, which views language to be a repository of signs of varying degree of specificity.

2.2 The merits of being usage-based

"[A] linguistic pattern is recognized as a construction if some aspect of its form and function is not strictly predictable from its component parts, but a linguistic pattern can also receive unit-status, i.e. be recognized as a construction, if it occurs with sufficient frequency."

Goldberg (2003)

The preceding section has presented arguments for the idea that an adequate linguistic theory should be one that recognizes the sign, i.e. a conventional association of a form and a semantics/function, as its central building block. This section elaborates on the second property that is presented here as essential for an adequate view on natural language, namely the property of being usage-based. The expression *usage-based (model)* has been introduced into the discussion by Ronald Langacker in the late 80ies (Langacker 1987, 1988) and has been kept close to its original formulation in later explications (Langacker 1999, 2008). One of the most important properties of a usage-based approach to language is that it sets out to induce all the properties of a postulated grammar from actual language data, which ideally approximate an individual's experience with language (Bybee 1999, 2006). Grammar, in this view, just *is* the cognitive organization of one's experience with language. In his original introduction of the term, Langacker emphasizes from the very beginning the inductive,

bottom-up character of linguistic knowledge. He writes:

> "*Substantial importance is given to the actual use of the linguistic system and a speaker's knowledge of this use; the grammar is held responsible for a speaker's knowledge of the full range of linguistic conventions, regardless of whether these conventions can be subsumed under more general statements. [It is a] non-reductive approach to linguistic structure that employs fully articulated schematic networks and emphasizes the importance of low-level schemas*"
>
> (Langacker 1987: 494; repeated in Langacker 2008)

An important finding that linguists who pay attention to usage have emphasized again and again—and that is very well captured in spirit by construction grammars—is that language users rely heavily on conventionalized word sequences (*chunks*, *prefabs*) that vary in complexity. These can consist of simple bi-gram such as *pull strings* but may also be more complex so as to contain many units. The relationship between the recurrent use of sequences of signs and their representational status in the mental grammar has already been described in the context of automatization/entrenchment in § 2.1. The next section will return to this issue and spell out in some more detail the effects that frequent usage has on mental representation.

2.2.1 Effects of frequency

The effects that the frequency with which a particular stimulus type is processed by an individual have intrigued functionally oriented linguists for at least thirty years (e.g. Fidelholtz 1975) but can be traced back to much earlier times (e.g. Schuchardt 1885, Zipf 1935). Even though frequency effects were continuously introduced into linguistic discussions these were often restricted to arguably more remote fields of linguistic inquiry such as phonetics/phonology or sociolinguistics and have been largely ignored (i.e. treated as performance phenomena) by researchers that set out to explain linguistic competence in the sense of Chomskyan linguistics. In recent years, however, the interest in frequency effects and distributional information has regained much of its popularity in connection with

developing ideas of language as a dynamic and emergent system (e.g. Hopper 1987) and in connection to research into grammaticalisation (cf. Bybee 2007, Diessel 2007 for an overview of frequency effect in language acquisition, use, an diachronic change). The (re)growing body of frequency sensitive research has disclosed a number of different effects that frequencies of use may have on language, which have been disclosed, collected and systematized in recent years most notably by Joan Bybee and her colleagues (Hooper 1976, Bybee 1998, 1999, 2001, 2007, Bybee and Scheibman 1999). A recent formulation from that work is well suited here to illustrate the general position, which is adhered to throughout the present work.

> *"Language can be viewed as a complex system in which the processes that occur in individual usage events [...] with high levels of repetition, not only lead to the establishment of a system within the individual, but also lead to the creation of grammar, its change, and its maintenance within a society"*
>
> (Bybee 2007:23)

Bybee distinguishes three types of (high token) frequency effects:

1. The *phonetic reduction effect*, i.e. an effect observed for high frequency words and phrases in which the units undergoing the effects undergo a phonetic reduction faster than units of lower degrees of frequency. Bybee and Scheibman (1999) discuss a—by now classic—example of *don't* reduction, which is observed to be most pronounce in high frequent environments such as *I don't know*.
2. The *conserving effect*, i.e. an effect observed for high frequency sequences of morphosyntactic strings, such as word strings. Such high frequent sequences become entrenched and so resist changes on the basis of more productive patterns. Examples would be verbs, such as *go* or *keep*, whose high token frequency shields them against regularization of their respective past tense forms (Hooper 1976).

3. The *autonomy effect*, i.e. the effect observed for high frequency morphologically complex forms to lose their internal structure and become dissociated from their etymological context. An example of that effect is the semantic opacity of words like *dislocate*, which is argued to be due to the fact that the complex derived form is more frequent than its base (Hay 2001).

Bybee and colleagues propose that these effects result from mechanisms of human cognitive processing, albeit not necessarily the very same processes. The reduction effect is often discussed in contexts of (historical) language and is closely connected to the kind of consonant mutation known as *lenition*. As a result of this phonetic process, which has been observed for many languages, a consonant will continuously change so as to become "softer", metaphorically speaking. This may involve a change in voicing, e.g. form [f] -> [v], debuccalization (loss of place), e.g. [s] - > [h], or deglottalization, e.g. [k'] -> [k]. Eventually, such processes may lead to a consonant being lost altogether, which may give rise to a chain-reaction of changes in the language system (Fidelholtz 1975, Johnson 1983, Pierrehumbert 2001, Aylett and Turk 2004, Boersma 2005). It is important to note that the phonological changes that started with some frequent words often carry over to other analogous lexical items in a process named *lexical diffusion* (Hooper 1976, Phillips 1984, 1999, Bybee 2000, 2002). These processes are very closely related to the concept of entrenchment and we may use them to motivate a complement to our first explication of the concept in § 2.2.1.

ENTRENCHMENT (EXPLICATION 2)

Articulating language involves the execution of neuromotor routines. When sequences of neuromotor routines are repeated, their execution becomes more fluent (Anderson 1993). Repeated sequences become unit-like, which means that after being accessed in rapid succession for a critical amount of times, they eventually can be accessed as a single unit. In addition to becoming unit-like, the memory trace corresponding to the now established unit gets strengthened by extensive activation (high frequency of use), making it easier to access/activate.

We shall come back to a more detailed discussion of the cognitive underpinnings of these processes in section §2.3. At this point, however, we should briefly pause and think about what this suggests for complex patterns like relative clause constructions. When we transfer the general logic behind the proposed explanation of these frequency effects to the processing of complex sentence, we can immediately derive the line of argumentation employed in the present study:

> If (**p**) processing difficulty is identified with the difficulty of mental activate a stored representation and (**q**) activation difficulty is a function of usage frequency, and (**r**) frequencies in corpora can yield approximations of an individual's experience, then (**s**) we can utilize corpus-based methods to infer properties of the cognitive system.

We have now provided the theoretical underpinnings from linguistic thinking arguing for the idea it is desirable for a linguistic theory to be first sign-based (or constructionist) and second to be usage-based. Grammar, in this view, is shaped by language in so far as usage frequencies act on the ease of acquiring and processing linguistic structures in the individual and also play a key role in the change of language undergo over historical time.

With regard to how to read these explications, I would like to remind the reader that the assumptions and commitments presented in this section are best viewed as restrictions on linguistic theories that are suited to serve as the conceptual backbone of the present study. This is to say that even though the conceptual apparatus of construction grammar will continue to be used for many illustrative purposes as it provides a very natural way of representing the phenomena to be examined, the overall theoretical embedding will be open to a larger body of approaches. This relatively open framing is due to two reasons. First, it is not my contention to suggest that the present study is primarily geared to corroborate a particular linguistic theory as it presently exists. Rather the goal is to produce results that are

compatible with a variety of approaches that share a certain vantage point (while still others will not have any obvious connections to the present study). In this respect, I follow the approach taken by John Hawkins (2004) and others who try to ensure that their suggestions are translatable into a number of theories provided they meet certain minimal requirements.

In short, what we are interested in at this point is the identification of properties that are desirable from a linguistic and psychological perspective. A fully adequate theory contains—or at least has explicated interfaces to—accounts of how such a system can possibly be learned, how it is mentally represented in the individual, and how it is processed under real-time in actual communicative contexts. Because of the probabilistic character of cognitive processing in general (cf. Rao, Olshausen, and Lewicki 2002 for an overview), these desiderata categorically rule out proposals that do not incorporate information about frequencies of use. It leaves, however, still enough room for a number of approaches that share a crucial set of assumptions and meet the theoretical desiderata. These include certain versions of Lexical Functional Grammar (e.g. Bod and Kaplan 1998), certain versions of Head-Driven-Phrase-Structure-Grammar (e.g. Arnold and Lindarski 2007), and certain versions of Construction Grammar (e.g. Bergen and Chang 2003, Steels and De Beule 2006) to name but a few.

In consequence, instead of deriving hypotheses from a specific linguistic theory, the present work is grounded in a theoretical environment that embraces the central ideas of usage- and sign-based linguistic theorizing married to an exemplar-based conception of representation and processing. The next section shall present an outline of the latter.

2.3 Construction-driven memory-based language processing

Having provided the general motivation for framing this study within a usage-based cognitive constructional paradigm, this chapter is dedicated to providing the remaining necessary background from cognitive psychology and computational approaches to language learning and processing. Specifically, we shall now provide an outline of exemplar/memory-based approaches to language processing. As for many of the notions employed here, a single agreed upon definition of *the* exemplar-based approach has not yet been reached in the

psychological fields that study them and the goal of this section is not to answer any of the unresolved issues in those fields. Rather the goal here is to provide a psychologically plausible basis for the central hypothesis, i.e. a framework which is general enough so as to include different approaches unified by certain high level commitment, namely the role of memory and analogy in language processing. It should however be specific enough so as to allow the derivation of testable predictions about processing difficulty.

We have argued in the above for the idea that the processing difficulty of a linguistic structure can be understood as a function of that structure's degree of entrenchment as envisaged in Cognitive Grammar. By now we have already worked our way to accepting the idea that degrees of entrenchment can be identified with the cost associated with the activation of a set of memory traces, which in turn is heavily influenced by the amount of times (i.e. the frequency) a stimulus of a particular type has been processed in past experience. Let us turn to a sketch of how an exemplar- or memory-based model of language processing might look like and how this may add to the plausibility of the present approach.

2.3.1 Memory-based language processing

One of the most promising types of model in contemporary cognitive science is the family of *memory-based* or *exemplar-based models*. Memory/exemplar-based models have been very successful in psychological research into categorization (cf. Smith & Medin 1981, Nosofsky 1986), they constitute one of the current mainstream approaches to modeling memory (Baddeley 1997, Neath and Surprenant 2003), and they are becoming more and more successful in domains pertaining to language as well (cf., e.g., Chandler 1993 for a alternative to the *dual route* models of language processing). In fact, the area of application of such models is actually a lot broader ranging from problem-solving (e.g. Rodriguez et al. 2000) to computer vision (e.g. Ong et al. 2006) to musical processing (Bod 2001) and, very ambitiously, even to science in general (Bod 2006). Memory-based approaches to language processing share the assumption that linguistic behavior is guided by the language user's prior experience with language and is hence very much in line with what has been said about the usage-based approach above. For the purposes of this discussion, I take the expression "memory-based language processing" to label a class of approaches that are unified by their

higher level assumptions about the nature of linguistic processes, and not as a label for a particular variant of such an approach or even a particular incarnation. Hence, the term is supposed to include memory-based language processing accounts in the narrow sense (cf. Daelemans et al. 1997, Daelemans 1999) but also related approaches such as analogical reasoning approaches (Skousen 1989, 2002) and the work on data oriented parsing (Bod, Scha, and Sima'an 2003). For a discussion of some theoretical, algorithmic and empirical differences among these approaches the reader is referred to Daelemans (2002). In a way, usage-/memory-based models of language can be viewed as a revival and refinement of ideas about analogy and induction in language already present in the work of de Saussure and Bloomfied, which at the time were specified only vaguely and were eventually replaced by the clearer and more rigid notion of a rule-based grammar in generative theories following Chomsky (cf. Skousen 1989 for a discussion of this point).

The central idea of memory-based approaches is that language learning and processing involves the "direct re-use of memory traces of earlier language" (Daelemans 1999:1). The approach incorporates two principles: first, that learning is the simple storage of experience and, second, that solving a new problem is achieved by reusing solutions from similar previously solved problems. Typically, memory-based models assume that people store individual exemplars in memory and categorize new stimuli relative to already stored exemplars on the basis of their similarity. From this point of view, all linguistic tasks are conceived of as classification (or categorization) tasks, which can be described informally as the process that solves this problem:

> Given a set of features F_s detected for stimulus s and given a set of categories C already represented, which category c exhibits a feature structure F_c most similar to s (or: which c allows for the easiest integration of s into C).

Obviously, looking at this description, a lot depends on the way similarity is measured in such accounts. We will return to this point later when we introduce the statistical procedure used here to measure inter-constructional (dis)similarities (§ 4.2.3).

The other notion central to the account is, of course, the notion of an exemplar itself. So let us start with a very innocent question: What exactly is an exemplar? There is, of course, some disagreement (or "variability") in the literature as to what the notion exemplar means exactly. Specifically, there is disagreement with regard to the question of whether an exemplar is a type or token representation. For example, McClelland and Elman (1986) propose a model in which words are represented in the lexicon in the form of abstract phonological representations, which would correspond to a "type representation"-conception. Others, e.g. Goldinger et al. (1992), Goldinger (1996) or Pisoni (1996), assume that word-forms are stored in the form of detailed acoustic traces, corresponding to a "token representation"-conception. Following Bod (2006), we will assume that

> "[…] *an exemplar is a categorization, classification or analysis of a token […] while a token is an instance of use. […] Thus an exemplar in syntax can be a tree structure of an utterance, a feature structure or whatever syntactic representation one wishes to use to convey the syntactic analysis of a particular utterance*."
>
> Bod (2006:2)

Exemplar-based models usually keep a store of representations of all previous language experience with each representation, i.e. an exemplar, corresponding to the analysis of a particular usage event. In addition to storing exemplars, some models make use of more abstract categories as well: so called *instance families* are variable-sized sets of same-class nearest neighbors and are helpful here conceptually as they correspond very closely to the idea of a schema as introduced earlier. Even though there are results from computational experiments showing that abstractions of representations are unnecessary and maybe even harmful (Daelemans 1998b), we shall still make use of the notions of an instance family and schema, if only for reasons of argumentative convenience and exposition. We may think of schemas and instance families as higher level descriptions, which—even though they may not be constituents of an adequate computational model of language processing—may be useful when it comes to talking about sets of exemplars. Similarly, Bybee (2002) employs the

notion of an *exemplar cluster*, which corresponds to the idea of an instance family. Figure 11 illustrates the concept of an instance family in a two dimensional instance space.

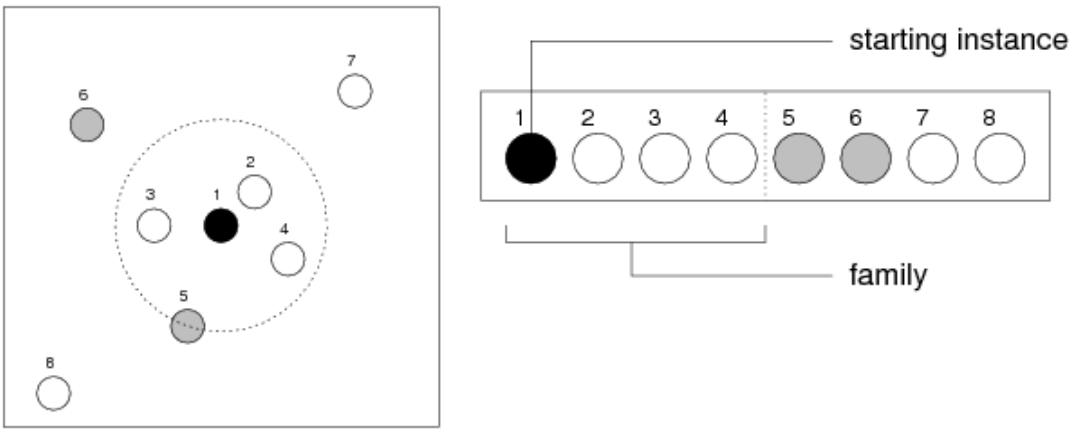

Figure 11: Illustration of instance family in two-dimensional instance space (Van den Bosch 1999)

Starting from some instance 1 (the big black dot), we can define an area (denoted by the dotted line) marking the boundary of that family. This area denotes an area in instance space in which instance exhibit a user-defined degree of similarity to instance 1 and hence are considered as belonging to the same family as instance 1. In our example the instance family would be the set {instance 1, instance 2, instance 3, instance 4}. Family membership is discrete: an instance either is a member of the family or it is not. Elements not fully satisfying the properties specified in the definition (instance 5) are treated as non-members. The difference between a classification based on instances only versus a classification based on instance families is that in the latter approach it is possible to match new instances against value combinations that have not been observed (and stored) before. We shall, however, not ponder about this any further as differences at this level of detail are most relevant in the comparison of different variants of memory-/exemplar-based models (for an overview and discussion cf. Van den Bosch 1999). What is most important here is the idea that processing a linguistic stimulus is its categorization relative to a set of already stored cases and/or collections of such cases.

Now that we have helped ourselves to a working definition of some of its key notions, we may have a look at the use of exemplar-/memory-based models in linguistics and how

they relate to English RCCs. Following Hay and Bresnan (2006), we may start from the observation that exemplar-based models have been developed (quite independently) in different areas of linguistics to the effect that it is sometimes helpful to distinguish the "phonetic exemplar theory" (PET) from the "syntactic exemplar theory" (SET). We shall focus our discussion on the latter (for the simple reason that this study is interested in the syntactic forms of English RCC, but will be silent on the phonetic forms, which pertain to an even more fine-grained level of description). The PET has been employed in both speech perception and production (cf. Johnson 2007 for an overview of exemplar-based phonology). In this approach we store every word we encounter in actual usage events. Whenever a new word is encountered, it is compared and categorized relative to already stored exemplars. This categorization is done on the basis of the degree of similarity and is usually computed in terms of the items distance in some parameter space (Pierehumbert 2001). The syntactic exemplar theory holds that there are no explicit rules in syntax. Instead the regular nature of linguistic behavior is a product of analogical generalizations that are ubiquitous in human cognitive processing. Analogical processes surely are not restricted to form-form matching but figure in processes at the conceptual level as well. It can be argued that analogical reasoning is the basic style of human thought. Grammar arises through analogical processes over stored chunks of previous language experience, i.e. actual usage events. As language is used for the purpose of communicating ideas, these experiences tend to be more complex than a single word and may thus be viewed as sequences of units. Recurring sequences may be as complex as phrases, e.g. the VP *jog* <someone's> *memory* but—as people grow older and collect more and more data—soon become even more complex so as result in the establishment of high level patterns that correspond to argument serialization constructions. This way of thinking about language and grammar is thus closely connected to the concept of schematization and routinization postulated in cognitive construction grammars.

In summation, memory- or exemplar-based models assume that all linguistic experiences are stored in memory and that they are structured on the basis of their degrees of similarity. The notion of an exemplar allows us to treat complex and only partially specified structures, i.e. particular types of schematic RCC, as exemplar clusters, i.e. categories whose members are more or less central to the category, resulting in prototype effects. Exemplar

representation also allows specific information about instances of use (most notably frequency information) to be retained in representations and thus provides a natural way to allow frequency of use to determine the strength of exemplars.

2.3.2 Categorizing complex constructions

With this general conception in place, let us now turn to a sketch of how a memory-based view can handle abstract patterns, such as English RCC. Figure 12 presents a sketch of the architecture of an exemplar-based theory of language (taken from Daelemans 1998a). It depicts the relation between linguistic experience and linguistic knowledge as it is conceived in a Chomskyan view (⇒ italicized components of grammar on the left) and an exemplar-based view (⇒ italicized & bold components on the right).

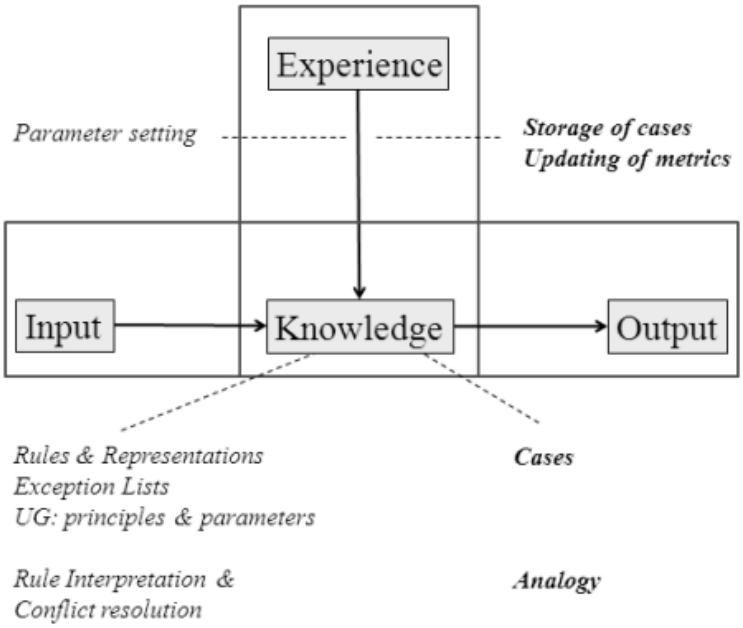

Figure 12: Sketch of an exemplar-based linguistic theory (Daelemans 1998a)

While vertical arrow in Figure 12 denotes acquisition, which in the exemplar-based case is reduced to the incremental, data-oriented storage of experiential patterns, the horizontal ones pertain to performance, which—in this view—is modeled as extrapolation of information on the basis of a language-independent similarity metric. As our main interest here lies in language processing, we may focus on the horizontal relations enclosed by the red box (for

exemplar-based views on acquisition, see Abbott-Smith and Tomasello 2006, Batali 2002, Bod 2009.). At any given point in time, an adult grammar consists of a huge set of established constructions which serve as models (or templates) for new cases. Each exemplar consists of an input representation and an output representation. Input representations always take the form of a vector of symbolic features, whereas outputs could be any type of classification result. To give a typical example: if the task is past tense formation, the input would be a feature vector containing information about segmental and syllable structure information about the stem, and the output would be a past tense form. If the verb is known to the system, the associated past tense form is retrieved from memory. In case the verb is new (=unknown), the past tense is formed on the basis of an analogical matching procedure which uses as a model for the new type the stored exemplar that is most similar to the current case. The similarity assessment in Daelemans (1998a) uses a distance metric with feature relevance weighting. The weighting provides for each feature in the vector describing the input representation a numerical value indicating its information gain, i.e. "a number expressing the relevance of the feature in terms of the average amount of reduction of information entropy in memory when knowing that feature" (cf. Daelemans & van den Bosch 1992). The method is just one of many possible to weigh the relative importance of features used in statistical pattern recognition and machine learning. We may content ourselves with an intuitive understanding of the approach here, but shall return to issues in assessing similarities among linguistic object in § 4.2.3. The general idea behind the exemplar-based approach to processing is that it essentially characterizes all linguistic tasks as classification tasks of some sort. Correspondingly, syntactic processing is characterized as a complex process involving a number of constitutive processes (tagging, constituent detection and labeling etc.). But again, the intriguing thing about the approach is that all these sub-problems are all solved by the very same universally applicable mechanism. Figure 13 presents a hierarchical organization of such task as they pertain to syntactic processing:

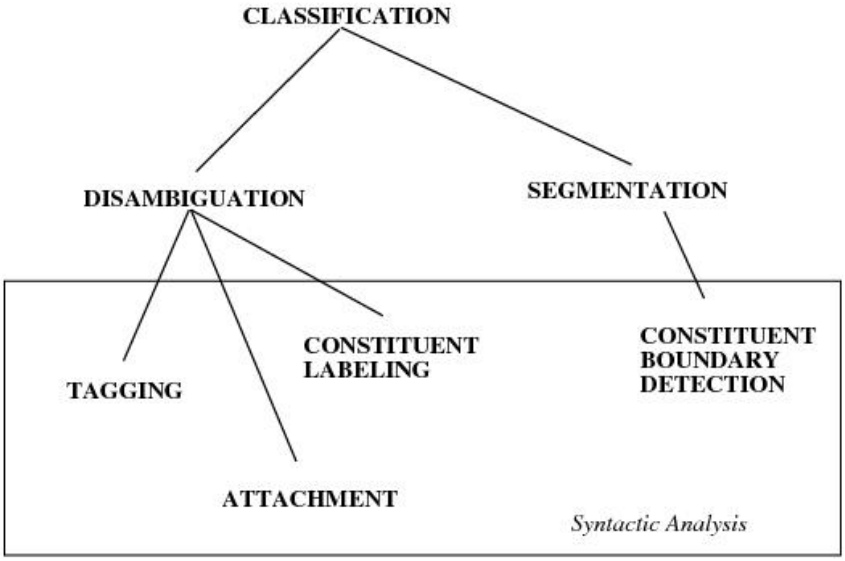

Figure 10: Linguistic tasks as classification tasks (Daelemans 1998a)

For all the specific tasks (sub-problems) involved in syntactic analysis, we may assume a process taking us from a feature vector (input representation) to a classification result (output representation).

So, how can we apply this general model to the processing of complex construction such as RCC? One of the most developed syntactic exemplar models has been worked out in the grammatical framework of Data-Oriented-Parsing (Bod 1998, Bod 2000) and so we are well advised to have a look at this work to get an idea of how the problem is tackled there. In this model, exemplars correspond to syntactic structures of previously processed utterances and a new utterance can be comprehended by matching it against the largest possible and most frequent chunk of stored units. Figure 14 can help us grasp the relationship between frequencies of a structure and processing difficulty in such a model.

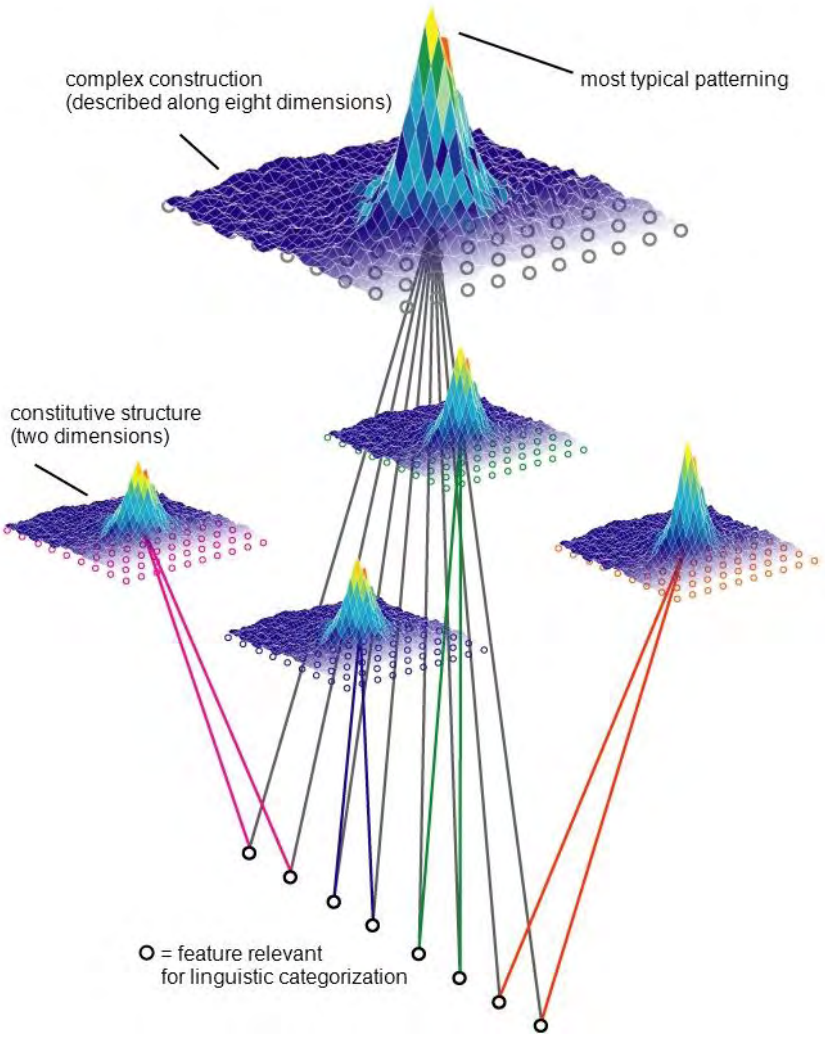

Figure 14: (Joint) attractor strength of linguistic structures

Let us say that each circle at the very bottom may represent a feature (or feature bundle) that figures in the description of a subset of the symbolic units (=constructions) that make up our linguistic knowledge. A given feature may figure in the description of many symbolic units if these units share that feature. For example, the feature +CONCRETE is shared by the representation of a large set of entities, say SCHOOL BUS, HUMAN BEING, or LASER PRINTER. The same feature may also figure in the description of signs that are not situated on the same level, if they were related within a hierarchy. That is to say that the feature +CONCRETE might figure in the description of the string $_{NP}$ [*the guy*] but also in the description of its dominating phrase $_{NP}$ [$_{NP}$ [the guy] $_{NP}$ [*on the roof*]]. The structure label corresponding to the string *the*

guy in *the guy on the roof* need of course not be analyzed as an NP and might as well be considered to be an N-bar constituent. This, however, is not relevant for our present purposes. Similarly, it will also figure in the complete description of a sentence level representation of, say, the string *The guy on the roof is about to jump*. The complexity of an utterance usually has a rather direct impact on the number of features that figure in its description such that instances of more complex constructions—such as relative clause constructions—incorporate by necessity a number of simpler constituent constructions of different types: minimally, i.e. in a scenario in which both clausal constituents exhibit minimal valency values, these include a single referring expression syntactically realized as a NP constituent that functions as the head of the RC proper and two predicating expressions, e.g. two VP constituents. As this is rather abstract, we may for expository purposes think of a sentence like "$_{NP}$ [*The guy*] *who* $_{VP1}$ [*never slept*] $_{VP2}$ [*died yesterday*]". These necessary constituents in turn consist of a certain number of constituent constructions (i.e. the set of phrasal constituents dominated by these constituents), which often are still divisible into even smaller constituents (i.e. the set of terminal nodes). Eventually we will arrive at the lowest symbolic level, say a set of integrated morpheme-sized constructions, which can be described on the basis of particular sets of properties. Of course, verbal constructions require different sets of properties for their description than nominal constructions as their respective semantics are quite different. The overall, integrating structure, however, will incorporate all dimensions of contrast used in the description of its constituents. Now, a constitutive structure of type NP can vary along a number of dimensions, say, definiteness and syntactic realization (lexical—pronominal) and some of the many possible types are likely to be more prominent than others. We may presume for example that a typical speaker of English has perceived a far greater number of tokens of pronominal NPs than lexical NPs, and may assume furthermore that certain types of lexical NPs, say *the guy*, have been perceived more often than others, say *the orthodontics*. The difference in token-frequency of possible instantiations (or elaborations) of a schema are represented in Figure 14 by the height of the hump, so that the highest point in a given map corresponds to the most frequent state in that state space. It is important to note that it does not follow from the fact that a particular instantiation/elaboration is most frequent across all possible linguistic scenarios, that it is also most frequent in a particular syntactic

environment. It is very well possible that for the set of all NPs ever perceived by a language user, the pronominal *I* is the most frequent head overall, but a rather rare one post-modified by a relative clause. The conditional probability of *I* being the head of a dominating NP, P(*I*|NP_{dom}), presumably is far greater than P(*I*|NP_{dom} with RC). Consequently, it is possible that the peaks in the maps representing the constitutive constructions (i.e. the four small maps in Figure 14) to differ from those in more complex constructions (i.e. the top-most map in Figure 14). Consequently, it is the goal of this approach to identify patterns of RCC that exhibit an above chance co-occurrence frequency. These patterns are predicted to have the greatest impact on processing difficulty as they serve as dominant instance families, i.e. they represent salient schemas relative to which an incoming instance is categorized.

2.4 Chapter summary

This chapter has provided the theoretical background against which the study is meant to be understood. It was argued that it is advantageous to describe linguistic knowledge and the regularities that we can observe in the structure of that knowledge in terms of relationships among signs, i.e. conventional associations of forms and meaning/functions. A unified conception of linguistic knowledge is particularly useful when we take it as the goal of grammatical theorizing to provide accounts of language that are psychologically plausible so as to allow a more direct exchange of ideas of theoretical linguistics and research into language acquisition and processing. Recent developments in psycholinguistics strongly suggest that language processing is not fundamentally different at the lexical and syntactic level respectively and thus it appears sensible to reflect these insights in the way these types of knowledge are represented in a theory of grammar. The importance of frequency information has long been acknowledged not only in psychological domains of language but also in the context historical development and accounts of language change. If the factor frequency is so important in language change, in language acquisition and in on-line processing, it appears reasonable to include it into grammatical description as well. And finally, this section has described the basic assumptions of exemplar-/ memory-based models of language processing and representation. It was argued that these models combine nicely with the sign-based assumptions of the nature of linguistic knowledge and also provide a

mechanistic underpinning of the observed frequency effects and the usage-based character of language. A usage-based cognitive construction grammar married to an exemplar-based model of language processing thus promises to help us bridge the gap between psycholinguistic and grammatical theorizing. Having provided all the necessary conceptual background, we are finally in the position to approach the empirical part of the study.

3 Describing English RCCs: Methods, data, and beyond

This chapter will present the corpus data used in the study and the variables used in their description. Special emphasis is put on the contrast between written and spoken language and so we will provide discussions of potential distributional differences of some key variables in a modality-specific contrastive fashion. We will discuss a subset of these variables. These variables are grouped into four coherent groups:

I.	Variables that are encoded on the head nominal	(Section 3.2)
II.	Variables that concern the relative clause proper	(Section 3.3)
III.	Variables that concern the dominating main clause	(Section 3.4)
IV.	Variables that relate the clausal constituents	(Section 3.5)

Each factor that has been included in any of these groups will be contextualized in the sense that its impact on the presumed overall processing demand will be explicated. The discussions of the psychological relevance of these variables, however, will not always be comprehensive. In some cases, a given variable does not fully express a quantity targeted in a theoretical treatment but is only part of a larger factor. These larger factors are variable-bundles that pertain to the *complexity* of a RC or the *predictability* of a RC. Both complexity and predictability are complex notions that may consist of variables discussed here in different groups. As a result of these considerations, we will pick up some of the variables in later sections, when we discuss the results of the multivariate procedures in Chapter 4.

We will start the empirical part of this paper with the description of the data set that was used in the analysis. The description addresses the following issues:

A. Exactly what types of constructions were targeted

B. Why the data were extracted from the ICE-GB corpus
C. Exactly what types of data were excluded from the data set and why.

The primary data set used in this study consists of 1000 bi-clausal relative clause constructions (RCC), i.e. complex sentential patterns that comprise of exactly one relative clause and one corresponding main clause.[13] The data set was so restricted in order to control for the formal variation of relative clause constructions that may be due to the complexity of the linguistic environment. For example, it is conceivable that the overall constructional complexity induced by effects of a larger number of clauses, has an effect on the preferred patterning of the relative clause. It is for this reason that the number of clausal constituents was held constant so as to minimize its potential effects on the outcome. To get a better idea of how more complex sentences can distort the picture consider the example in (42):

(42) I mean out of everyone [**that** I know] **that** I went to college with …

[S1A-034 #164]

The example in (42) suggests that a preferred patterning can be overridden in cases where the overall complexity of the sentence is increased by an additional clause. As we will see in more detail later the presence of an overt relativizer particularly in the first RC is rather surprising given the elaborations of the variable slots. Without going into the details of possible explanations of R-element omission, suffice it to say at this point that optional relativizers, i.e. R-elements that may but need not occur with finite non-subject relatives, are likely to be omitted when (a) the RC modifies the direct object (i.e. when the RC is right embedded), (b) when that object is formally realized as an indefinite pronoun (*everyone*) and

[13] It says "primary data set here" because for some specific purposes additional samples were extracted. These will be described along the way as they are introduced in the study.

(c) when the subject of the RC proper refers to the speaker of that utterance (*I*). As a first rough approximation of this distributional property we may apply a very crude technique: a quick-and-dirty Google-query through only the English sites of the WWW. Searching for occurrences of the string *everyone I know* and *everyone that I know* yields a ratio of 57.000 to 638.00 in favor of the shorter variant corresponding to a value of the odds of 0.098. These numbers certainly suggest that the *that*-less variant is the preferred one. In fact, the patterning in (42) is very close to what may be considered the prototype of a reduced (*that*-less) relative clause. A possible explanation for why it nevertheless occurs with an overt relativizer may very well be the presence of an additional subordinate clause that (for trivial reasons) adds to the overall complexity of the RCC. While it is argued in the present study that a revealing investigation of relative clauses has to take into account the properties of an obligatory main clause, it certainly is helpful to delimit the scope of the constructions that are submitted to the analysis so that more local factors influencing preferred linguistic patterning can be identified. Restricting the data set to what is minimally required is considered to be the most principled way to demarcate the object under investigation.

3.1 Corpus and data used in the analysis

The data for the present study were extracted from the British component of the International Corpus of English (ICE-GB R2). Unless stated otherwise, all queries were performed using the ICECUP III software package. Correspondingly, all syntactic specifications of search queries refer to the logics employed in that program. The ICE-GB R2 was chosen as a data source for the following reasons.

First, the corpus is adequately sized for the phenomenon of interest here: The corpus consists of roughly 1 million words of contemporary British English collected in the 1990ies. It subsumes 200 written and 300 spoken texts each of which is grammatically annotated and fully parsed resulting in a total of 83,394 parse trees, of which 59,640 are from the spoken part. The rich grammatical information allows for a systematic and sound extraction of the target constructions that is virtually impossible to achieve from unparsed corpora. In total, the ICE-GB corpus encompasses 8,248 dependent relative clauses, which was assessed by way of querying the corpus for strings matching the pattern [,CL[rel, ¬indrel]], which excludes

independent relative clauses that do not have a head word. An example of such a construction, which is not part of the present analysis, is the string *what I want* in a sentence like *what I want is a new car*. Such constructions are often labeled 'nominal relative clauses', which alludes to their functional similarity to nominal clauses and if we target them in the corpus, we observe that they are far from being rare ([, CL[¬rel, indrel]] ⌒ 3,505 hits). Such independent relatives are formally and functionally different enough to be excluded here. All targeted constructions serve the function of argument modification.

The crucial subset of bi-clausal constructions was arrived at using a rather larger number of queries that targeted specific types of RCC individually. Taken together these separate queries cover all logically possible target constructions. To give an example: Right embedded constructions of the desired type were extracted querying for the regular structure [(PU, CL) ((,NP) (,VP) (,NP) ((,) (CL(depend, rel))…)…)], where "PU" specifies the "parsing unit", which is a necessary feature of any RCC-main clause. "CL" specifies the syntactic unit "clause". As indicated by the bracketing, the next lower level describes a serial order of NP, VP, NP constituents, with the relative clause being a constituent of the latter NP. The ICE syntax does not treat direct objects as VP internal arguments but aligns SVO elements on the same level. This query is the most general description of right embedded RCC in which the RC modifies an argument of the MC (as opposed to an argument in yet another subordinated clausal constituent). Figure 15 illustrates the corresponding output:

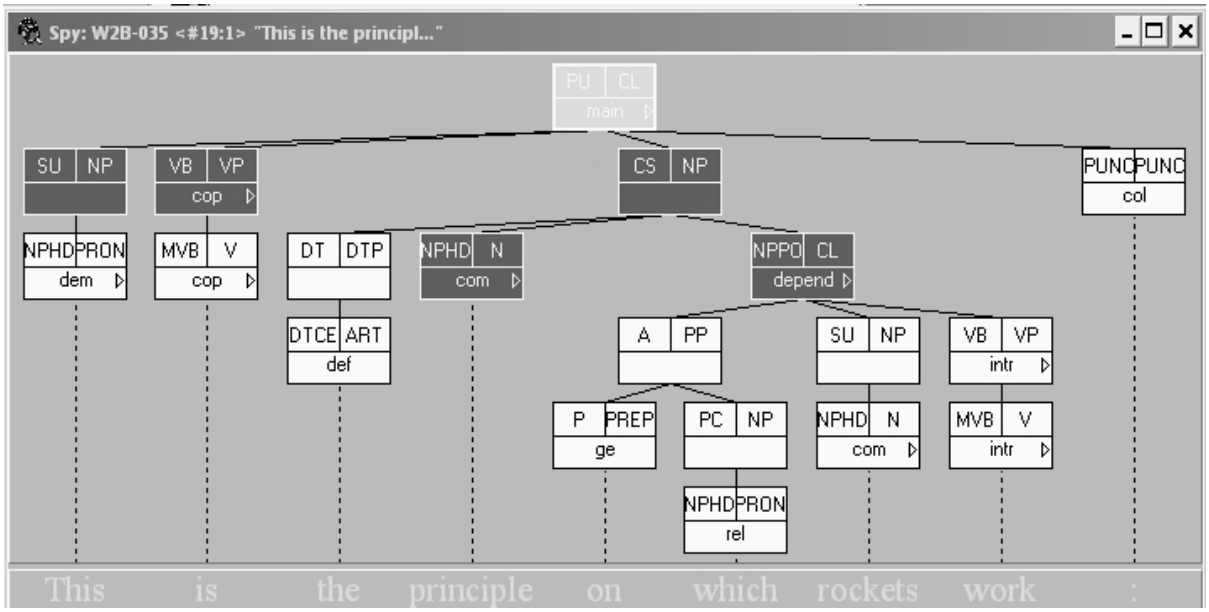

Figure 15: Example ICECUP output: right embedded RCC [W2B-035 #19:1]

Figure 15 presents a corpus example matched by the search pattern above, namely the structure assigned to the string *This is the principle on which rockets work.* The dark boxes represent elements specified by the search pattern. In the actual output, these boxes contain a little more information than was actually specified in the search pattern rendering the relation of input and output somehow opaque. This, however, was judged to be unproblematic for the purposes of the study and so will not be discussed any further. White boxes specify the units not specified by the search string.

For the present study, two complementary approaches to the extraction of the desired data were used. First, the data extraction procedure was conducted at a very general (coarse-grained) level that consisted of strings like the one presented in the example above. In addition to this, the study also employed search patters at more fine-grained levels to target more specific constructions by means of imposing more constraints on the output. This included for example the setting of ICECUP parameters such as TRANSITIVITY of the clauses involved to a specific value (i.e. complex transitive) or setting the GRAMMATICAL ROLE of the element modified by the RC. This dual approach, using very general and more specific queries in a complementary fashion, was employed to increase the reliability of the output. Ideally the sum of all specific constructions should equal the number of constructions

subsumed under a more general description and thus function as a checksum. It turned out that only rarely the general and more specific extraction procedures arrive at exactly the same numbers. However, the deviations tend to be rather small (rarely exceeding more than 10 cases) and may be due to different treatments of ignored cases (cf. Nelson et al. 1996 for details on the annotation procedure). This small deviation was not considered to be problematic especially since all outputted constructions were checked manually by the author to minimize the number of false positives (=Type I errors). After the first inspection of the data 2,388 points were consistent with the specification introduced so far, i.e. instantiating a bi-clausal RCC. In order to minimize the amount of noise in the data set, further constrains were subsequently imposed on the data. These constraints include the following: First, all main clauses had to be in the declarative mood. This restriction was introduced because in English (*wh*-) interrogatives introduce an additional gap and can therefore—all other things being equal—be viewed as being structurally more complex than their corresponding declaratives. This contrast in complexity is of course not a contrast that is directly related to relative clauses per se and was hence excluded from the investigation. In other words, sentential mood of the main clause was another controlled variable. Also, all relative clauses that made it into the analysis are clausal modifiers of a main clause nominal. Sentential relatives were weeded out by hand. This is another example of the merits of manual data inspection as it is impossible to automatically identify unwanted cases of that kind. Consider the following examples:

(43) And then I had the vegetarian option *which was a wonderful spinach cheese thing* (…). [S1A-011 #261]

(44) I've been able to use some French in Romania *which was useful.* [S1A-014 #113]

If we classify the RC in the examples above on purely structural grounds, we will be forced to treat all of them as subject relatives introduced by a particular variant of *wh*- relativizer, namely the form *which*. We cannot automatically retrieve the information of what the logical

subject of the RC effectively is, because the structures alone do not carry this information. The logical subject must be recovered on the basis of semantic plausibility and overall pragmatic coherence. The RC in (43) modifies the nominal *option* (or *vegetarian option* depending on whether or not one considers *vegetarian* to be an adjectival premodifier) and is thus clearly an instance of the target construction. In contrast, (44) was excluded because it does not modify the entity denoted by the head noun (i.e. either *French* or *Romania*) but comments on the proposition expressed by the superordinate main clause. A linguistic test that can help us determine what type of RC we are looking at is rephrasing the sentence in question as an *it*-cleft construction (as in 45) or a sentential subject construction (as in 46):

(45) It was useful that I've been able to use some French in Romania.
(46) That I've been able to use some French in Romania was useful.

Notice that—unless we postulate ad hoc meanings that differ from the conventional semantics of the terms involved—such reorderings do not yield semantically acceptable sequences in the case of (43) as shown in examples (47) and (48). The symbol '??' is used here to indicate semantic incongruity.

(47) ?? It was a wonderful spinach cheese thing that I had the vegetarian option.
(48) ?? That I had the vegetarian option was a wonderful spinach cheese thing.

It is important to note that this cannot be due to the fact that the string *a wonderful spinach cheese thing* is of the syntactic type NP (whereas *useful* in (44) instantiates an AP). It would clearly be possible to modify the example in (44) slightly without changing the fact that the RC expresses a comment on the proposition described by the MC. Consider (49).

(49) I've been able to use some French in Romania which was $_{NP}$ [a great thing].

These examples demonstrate that if we wish to maximize data quality, we cannot do away with manual data inspection and linguistic judgments. As long as machines do not understand natural language these judgments have to be made by humans. However, even with manual inspection of the data, it was not always possible to determine on syntactic grounds alone whether not a given example should be included in the data set. The example in (50) represents a case of syntactic ambiguity which could not be resolved without committing oneself to a particular semantic interpretation, which is highly problematic, given that we cannot know what the speaker of that utterance wanted to express.

(50) And then on Sunday <uh> we did a third wood in the morning *which was different*. [S1A-036 #215]

Strictly speaking, the RC *which was different* could modify the nominal *wood*, even though the sentential modification interpretation is far more plausible. The point is that plausibility judgments are always probabilistic and it was the first objective of this study to minimize the use of subjective criteria to the extent that this is possible. Quite generally, whenever the question arose as to whether or not to include a given data point in the analysis, it was systematically answered in the negative so as to minimize replication difficulties associated with subjective annotation choices. The same heuristics also requires the exclusion of all cases for which it cannot be decided exactly what nominal of the main clause was modified by the RC. Such a situation, a so called *RC attachment ambiguity*, arises in the context of periphrastic (analytic) possessive constructions, i.e. 'N1 *of* N2'-constructions as exemplified in (51).

(51) Here we have the latest **version** (N1) of the **car** (N2) that virtually killed of Land Rover in Africa. [S2A-055 #087]

In this example it is impossible to determine the head of the relative clause on purely structural grounds. English permits an RC to modify either the nominal inside the PP (=N2), which in the example above would be *car*, or the head of the dominating NP that functions as the object of *have* in the MC (=N1), which would be *version* in this case. If we represent the structure of the sentence hierarchically, it is natural to refer to these competing structures as being a case of 'low' and 'high' attachment, respectively. Such relative clause attachment ambiguities have been studied extensively in the sentence processing literature and are important for theoretical accounts that rely on some kind of locality principle, e.g. Frazier's principle of *Late Closure* (Frazier 1987), the *Recency* principle discussed in Gibson et al. (1996), *Most Recent Head* suggested in Kozniecznу et al. (1997), or *Locality* in Gibson (1998). Interestingly, it appears that different languages exhibit different preferences (or default interpretations) for one of the two possibilities. For example, Cuetos and Mitchell (1988) show a N2-over-N1-preference for English, but a N1-over-N2 preference for Spanish. More recent studies tend to deemphasize the idea of universal parsing principles which are then accompanied by language specific preferences, because there is so much intra-language variation. These approaches put more emphasis on explanations that relate attachment preferences to individual learning histories (cf. Cuetos, Mitchell, & Corley 1996), which is well in line with the usage-based account opted for in the present work.

Also, in order to minimize the effects associated with structural priming and therefore structural repetition based on resting activation (cf. e.g. Bock 1986; Bock & Loebell 1990; Pickering & Branigan 1999), I did not include more than one example from the same speech situation. This was considered an important factor in the attempt to maximize the overall data quality. It is, however, far from being the norm in corpus-driven work. In many cases linguists tend to emphasize sheer sample size without noticing (at least not commenting on) the effects that following this dubious maxim has on sample quality.

Finally, in addition to the points discussed above I have added a few minor constraints on the output, which I shall present in list-form.

I No RC INTERNAL VP ELLIPSIS as in

"I would take rather the same view that Joe Haines **did** […]"

[S1B-040 #031]

Reason for exclusion: Complexity estimation of such construction varies with particular theoretical decisions regarding the psychological status of the omitted (or at least unobservable) material.

II No PSEUDO EMBEDDING, i.e. the presence of linguistic material that appears to constitute an additional embedded structure, as in

"Those details that **you think** are most exciting"

[S1B-020 #074]

Reason for exclusion: Syntactic structure is controversial. There are good reasons to believe that strings like *you think* have grammaticalized into discourse (/pragmatic) markers that do no longer influence the hierarchical structure of the sentence (for a discussion cf., e.g., Romero-Trillo 2006)

III No COMPLEX ANALYTIC PREDICATES as in

"any foreseeable time-limit that you **are likely to propose**"

[S1A-024 #125]

Reason for exclusion: Theta role assignment is problematic. Depending on the apparatus used to describe the semantics of such predicates, we will assume one or more argument slots in which the logical subject. If we assume more than one slot, we may have situation in which the logical subject assumes different semantic roles, which prevents an uncontroversial decision of the role played by the head.

IV No 'FACT-S' CONSTRUCTIONS, which bear a close relationship to relative clauses and are sometimes treated as such. However, if anything they are similar to the sentential relatives discussed before. In English, these clauses behave rather differently in terms of their syntax as they do not contain a gapped role but are syntactically complete at the surface level. An example would be *I like the idea that students can become independent learners*.

V GRAMMATICAL ANOMALIES as in
"[…] and then there are the really bland ones that I think oh come on."
[S1A037 #186]
"[…] I started off applying for jobs that I was kind of like <…> in architecture […]"
[S1A-034 #160]

VI No data points from HIGHLY ATYPICAL TEXTS (e.g. staged speech from spoken part of the corpus as this involves only minor real time planning; dialogues in novels that mimic spoken language)

Filtering out all cases that exhibit any one (or more) of these properties reduces the data set to 1,188 case. These cases were then randomized and the final set of 1,000 cases comprises 500 randomly selected written and 500 randomly selected spoken examples. The remaining 188 cases were dropped. This last step was introduced for sheer convenience of interpretation of (relative) frequencies.

3.1.1 A roadmap for the analysis of the corpus data

The analysis that will be presented in the following is rather complex and so a good way to start is to provide a general overview of the steps to come. Figure 16 provides a schematic representation of how the analysis is structured, which I hope allows for a better overall orientation.

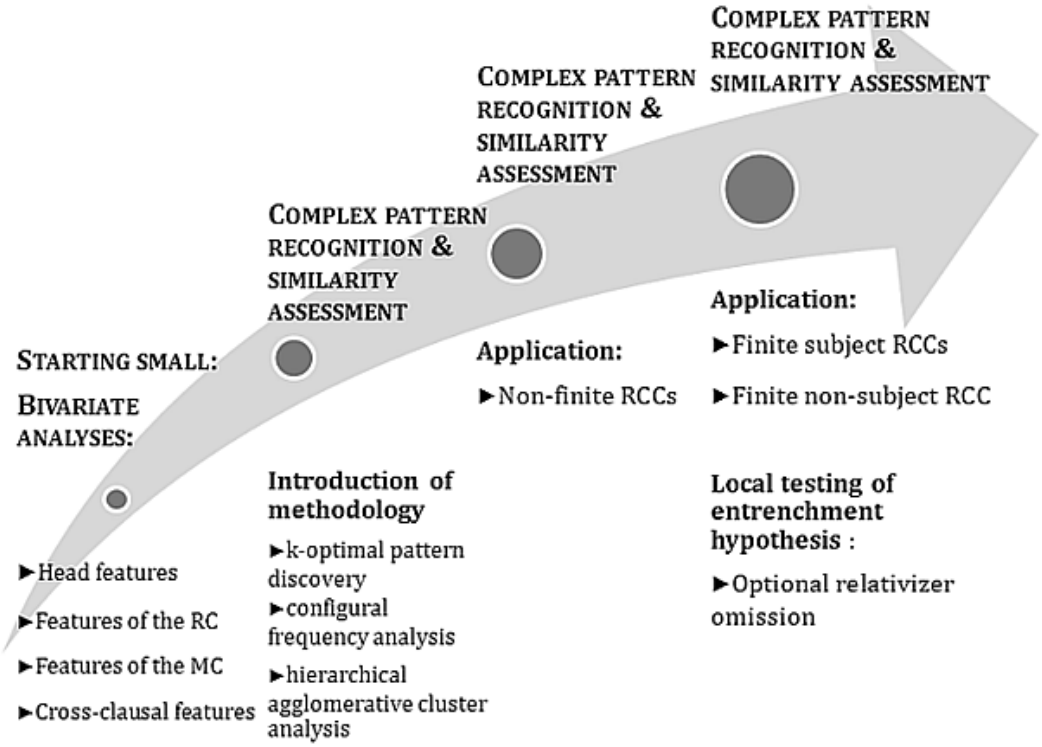

Figure 16: Roadmap of the study

As already indicated, the rest of this chapter will present a number of discussions of some the most salient variables that have been argued to modulate the processing difficulty of RCCs. The factors in this general overview are grouped into four coherent classes that describe a number of features encoded on the head (e.g. animacy, definiteness), a set of features describing the clausal constituents, i.e. the RC proper and the dominating MC, and finally a set of features that mediate the clausal constituents. It is in this last sub-section that we will tackle such issues as syntactic and thematic parallelism. In Chapter 4, we will introduce the

multivariate statistical machinery employed in the analysis, namely *k*-optimal pattern analysis (ARM), (hierarchical) configural frequency analysis (hCFA), and hierarchical agglomerative clustering (HACA). This triple of techniques is geared to serve two goals: first, to disclose interesting associative relationships underlying the data that are both complex and abstract. In other words the first goal is to detect schematic constructions in a statistically sound way. Once such entrenched patterns have been identified, the second goal is to relate these constructions to each other on the basis of their degrees of similarity. This grouping of prominent RCC types on the basis of their structural similarity allows us to identify salient exemplar clusters. A usage event of an RCC that falls within such a cluster is then predicted to be relatively easier to process/categorize than an RCC that departs strongly from such schemas. We will apply these techniques first to non-finite RCCs and continue our discussion with characterizations of finite RCCs, the focus being clearly on the latter constructions. This is partly due to the nature of the structures themselves. Non-finite relative clauses are less variable—and in a way less clausal than finite RCs—making them less interesting from the configurational view presented here. However, the main motivation for focusing on finite RCs here is grounded in the fact that the psycholinguistic research has clearly focused on finite structures. Consequently, there is considerably more experimental data on finite RCCs which we can compare to the present findings. Throughout the analysis, we shall always try to motivate the privileged representational status of the detected clusters by way of identifying the functions that these structures may serve in the discourse. Finally, the last step in the analysis is set up to provide corpus-based evidence for the idea that there is a close relationship between preferred patternings as identified by way of corpus analysis and the processing difficulty associated with a RCC-type. In this step, we will follow a dominant view in the processing literature, namely that overt optional relativizers signal processing difficulty, and exploit this idea to assess the general plausibility of the proposed account.

So generally speaking, our discussion of the corpus data will proceed from the general to the specific. It will begin with a description from a bird's eye view that allows for a first grasp on the most coarse-grained contrasts. With such an overview in place, the discussion will then "zoom in"—metaphorically speaking—, focus on more intricate details and inspect the interrelationships of the variables under investigation. We may entertain this "lens

metaphor" a little longer and think of the general procedure of this study as something very much like a photographer's search for a sharp image: in order to find the right adjustment it is helpful to play around with the lens so as to finally see the fine-tuned image.

3.1.2 Variables investigated in this study

This section presents the variables that were used in the description of the data. The set of 1,000 bi-clausal RCCs that met all the desiderata explicated in the preceding section were annotated with grammatical and conceptual information captured by some 30+ descriptors.

Some of the grammatical variable specifications (e.g. for tense, voice, finiteness, and transitivity of the two clauses involved) were directly imported from the ICE markup language and are indicated with an '*' in the left-most-column of Table 3 below. All remaining variables were annotated manually. Most of the variables were selected for the characterization of English RCCs because they have been argued to modulate the processing difficulty of linguistic expressions in some way. For the sake of expository convenience, we will explicate how a given variable pertains to the processing demand of a structure when we discuss the distribution of that variable in the corpus data. In addition to these factors, some variables were added by the author to test specific hypotheses derived from these proposals. Table 3 presents in alphabetical order the labels of the variables, a short description that should help the reader identify the nature of a given variable, and finally—in the right-most column—a relative assessment of the presumed processing demand associated with the respective factor levels that were distinguished in this investigation.

Table 3: Investigated factors (overview)

	Factor name	Factor description	Relative processing demand of factor levels
1	*add.mod*	presence of an additional modifier	add.mod << no.mod
2	*animacy.head*	animacy of the head noun	animate << inanimate
3	*animacy.SRC*	animacy of subject of RC	animate << inanimate
4	*concrete.SRC*	concreteness of subject of RC	concrete << abstract
5	*concreteness.head*	concreteness of the head noun	concrete << abstract
6	*content.head*	contentfulness of the head	general << specific
7	*definite.SRC*	definiteness of subject of RC	definite << indefinite
8	*definiteness.head*	definiteness of the head noun	definite << indefinite
9	*embedding*	type of embedding of RC	right << center
10	*finite.type**	finiteness of RC	fin << ing << to << ed*
11	*head.type*	syntactic type of the head	pronominal << lexical
12	*medium**	modality of case	NA
13	*rel.type*	type of relativizer	prn << that << zero*
14	*relativizer*	presence of relativizer	presence << absence*
15	*SRC.type*	type of relative clause	no a priori assessment
16	*subject.RC*	type of subject of relative clause	pronominal << lexical
17	*syn.parallelity*	Parallelism of grammatical roles	parallel << not parallel
18	*synR.ext*	external syntactic role of the head	S << PN/O << adjunct
19	*synR.int*	internal syntactic role of the head	S << O << adjunct
20	*synR.order*	serial order of syntactic roles	no a priori assessment
21	*tense.main**	tense of MC	present << non-present
22	*tense.rel**	tense of RC	present << non-present
23	*text.type*	specification of genre	NA
24	*theta.ext*	thematic role of head in MC	high roles << low roles
25	*theta.int*	thematic role of head in RC	high roles << low roles
26	*theta.order*	serial order of thematic roles	no a priori assessment
27	*theta.parallel*	parallelism of thematic roles	parallel << not parallel
28	*theta.SRC*	thematic role of subject of RC	high roles << low roles
29	*trans.main**	transitivity of MC	low trans << high trans
30	*trans.rel**	transitivity of RC	low trans << high trans
31	*unique.A*	presence of an uniqueness adjective	present << non-present*
32	*voice.main**	voice of MC	active << passive
33	*voice.rel**	voice of RC	active << passive

The symbol '<<' in right-most column is used to denote the assumed ranking of the factor levels, such that "X << Y" should be read as "the processing demand of a structure X is lower

than that of structure Y". In many cases these presumed orders are fairly uncontroversial. It seems safe to assume that high transitivity values add to the complexity of linguistic patterns and so we may assume that complex transitive clauses are harder to process than intransitive or copular constructions. In other cases, the relative processing demand associated with the contrasted factor levels may seem intuitively plausible, but at second glance turns out to be trickier than expected. An example for such a parameter is the factor VOICE. We may assume that active constructions are easier to process than their passive counterparts, because the serialization of the involved logical arguments (or thematic roles) departs from the preferred ordering in English (cf. Bates and MacWhinney 1989). While in active clauses, which arguably are a lot more frequent in the experience of a language user, more agent-like roles tend to occur before less agent-like roles, the reverse is true for passives. The first argument in a clause is likely to receive an agentive interpretation as the basic word order in English is SVO (c.f., e.g., Halliday and Matthiessen 2004 for a discussion of the relationship between grammatical and thematic roles). If hearers have less trouble with typical patterns, passives should thus be harder. However, there are certainly more factors that act on interpretative processes, e.g. the animacy values of the referent(s) of the NP(s) in question. Inanimate entities are more likely to play less agent-like roles in the situation described by the clause. Hence, a clause-initial NP that denotes an inanimate object can be seen as signaling an upcoming passive structure. If the structure under processing exhibits more cues that point to a passive construction, e.g. a verb that preferentially occurs with the passive (cf. Pinker 1989, Gries and Stefanowitsch 2004), such default interpretations are likely to be overridden fast. In a nutshell, it is certainly possible that certain passives are easier to process than certain actives. As similar comments may apply other variables as well, the a priori ordering in Table 3 must be taken with caution.

This brief discussion also illustrates the theory dependency of such a priori assessments. In multi-stage models of parsing (e.g. Frazier and Fodor 1978), the parser initially builds syntactic representations on the basis of default heuristics and uses non-syntactic information only later, in case these default operations have led to an ungrammatical (or inacceptable) result. In most constraint-based approaches to sentence comprehension however (cf. Trueswell and Tanenhaus 1994), it is assumed that the

interpretation process is influenced by information from many different sources from the get-go, including semantic and pragmatic information. If the latter view is taken, a general statement like "active << passive" loses a lot of its intuitive appeal.

In light of these considerations, I should add some words in defense of the decisions presented in Table 3. The present work is very sympathetic to the constraint-based perspective. However, the very same assumption that adds to its attractiveness, i.e. the emphasis of the interactive nature of the processing apparatus, in some sense also hinders it from making general and categorical predictions of the type exemplified in the right-most column of Table 3. To reiterate, all these statements about relative processing demands need to be understood on a ceteris paribus basis and must—in the absence of any knowledge about possible interactions with other factors—be treated with caution. But if we wish to make such general and directed statements, it appears to be more felicitous to assume an "active << passive" order than to assume a reverse order, i.e. "passive << active".

As a last remark on Table 3, let me comment briefly on the variables whose values for relative processing demand are annotated with an '*'. The asterisk in this column indicates that the relative processing demand of the levels of those factors is dependent on the perspective taken on the issue. That is to say, if one assumes a more local perspective, i.e. one that focuses on the factor-levels contrasts per se, or a more global perspective, i.e. one that goes beyond this paradigmatic contrast. Let me present an example to illustrate this point. Consider the variable UNIQUE.A. From a local perspective the presence of a uniqueness adjective is more costly than its alternative, i.e. the absence of such an element, because of the very simple fact that an additional word imposes additional processing demand. However, from a more global perspective, which is assumed here, the extra element is evaluated on the basis of its role in the structure and its function in discourse. From this perspective it can be viewed as a signal of an upcoming RC and so is viewed as an element that actually helps the hearer anticipate the structure to come. The same reasoning led to the ordering of the values of RELATIVIZER and REL.TYPE. Non-obligatory *that* relativizers provide information that is beneficial to structure building processes as *that* signals the onset of an RC. *Wh*- relativizers were assigned an even higher position on the ranking as they—in addition to signaling an

RC—may provide information about the internal role played by the head. The last variable marked with an asterisk, FINITE.TYPE*,* is similar in spirit, yet different enough so as to invite a little elaboration. The ranking "fin << ing << to << ed" was assumed because more finite structures introduce less uncertainty than non-finite ones. Only non-finite RCs involve implicit arguments, which gives rise to questions about control. Within the set of non-finite variants *–ed* was judged to be most difficult for reasons of internal consistency (recall that we assume "active << passives") and, finally, *to* infinitival RC are presumably harder than *–ing* relatives, because the to be recovered argument can be either an implicit subject or object, while in *-ing* relatives the implicit argument necessarily is the subject. The structures in (52) to (54) provide examples.

(52) This is a computer system$_{i\ RC}$[__$_i$ to map its waterworks].

(53) This is the book$_{i\ RC}$[to read __$_i$].

(54) Who is the guy$_{i\ RC}$[__$_i$ sitting in the corner]

Finally, it should be noted that not all variables will receive the same amount of attention. This is in part due to the simple fact that not all variables are of equal interest for the purposes of this investigation and the amount of ink spent on a given variable iconically represents this interest. But there is also a pragmatic reason for this treatment: discussing 33 variables in rapid succession certainly is not exactly reader-friendly. Be that as it may, those variables that did not earn their own sub-sections will be discussed either in the context of another related variable or will make its appearance when we arrive at the multivariate perspectives on RCC in Chapter 4.

3.1.2.1 Grouping descriptors

Instead of presenting the distributional statistics as an unstructured list, the variables are grouped to form coherent sets of descriptors that capture a particular dimension of contrast. In doing so, we secure a more manageable set of axes of comparison illustrated in Figure 17:

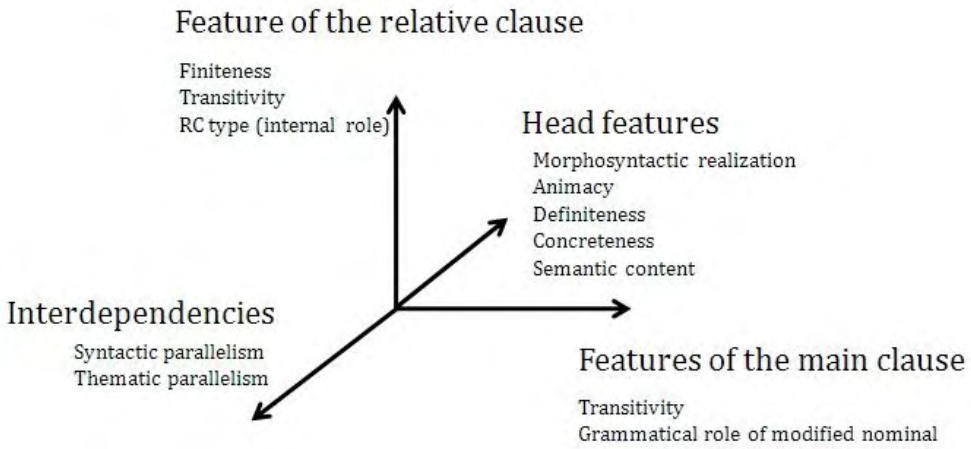

Figure 17: Axes of RCC description

Figure 17 presents four dimensions or axes on which the variables discussed in this chapter can be aligned. Two of these dimensions concern certain grammatical features of the clausal constituents of RCCs, i.e. the main clause and the relative clause proper. These include the grammatical categories such as voice, tense, transitivity, and finiteness of the clause. From the perspective of a cognitive construction grammar, the variables in these groups can be seen as grammatical reflexes of particular aspects of human conceptualization. We will elaborate on this idea when we turn to the respective variable. The other two axes in some sense relate the two clausal constituents: the head of the relative clause plays a role in the relative clause but also, of course, in the main clause. Following recent suggestions in the processing literature we will have a look at certain morphosyntactic and conceptual properties of the head. The other axis that in some sense links the clausal constituents is labeled 'interdependencies'. The corresponding section is concerned with syntactic and thematic parallelisms, which have been proposed to factor into the overall processing demand of a complex pattern.

3.1.2.2 Language Processing and Distributional Differences across Modalities

At each stage in the analysis, special emphasis will be put on the contrast between spoken and written discourse. The reason for why register contrasts receive so much attention here is closely related to the most general goal of this study, namely the demonstration of the

intricate relationship between language processing and shapes of grammars. Following the most fundamental assumption in usage-based linguistics, it is argued that grammars develop in accordance with principles of language use. As all usage events are the product of a language user's processing system, it seems almost trivial to state that the properties of that system play an important role in the shaping of grammar. Given that spoken language has to be processed in real time—while the processing of written language usually is self-paced—we would expect there to be differences in the distributional patterning across modalities. Even though this general idea may seem fairly obvious, I consider it worthwhile to delve into the issue a bit further.

Processing language auditorily or visually (i.e. through reading) clearly involves solving different kinds of problems. Whereas readers can control their rates of intake—they can simply slow down in case the prose is hard and speed up when it is easy—listeners are of course not so privileged but have to cope with whatever rate of speech is presented to them. Also, should the need arise, readers can easily go back and re-analyze the input as the linguistic material is externally represented—say on a piece of paper or on a computer screen. In contrast, the input in spoken language is fleeting and is available to the comprehender only for a limited amount of time. We can substantiate these considerations by relating them to the underlying architecture of human working memory. Maybe the single most influential model of that system has been developed by Alan Baddeley and colleagues (Baddeley and Hitch 1974, Baddeley 2000, Baddeley 2007). Figure 18 presents the model diagrammatically.

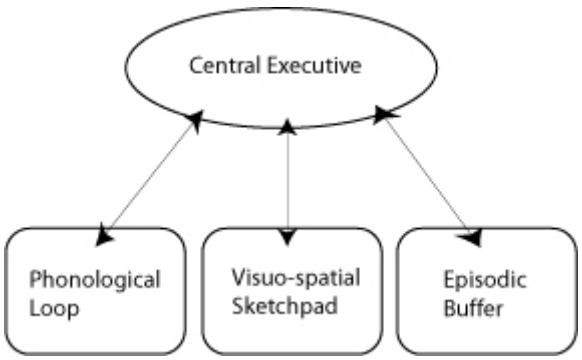

Figure 11: Multi-component working memory model (Baddeley 2000)

The term *working memory* here refers to "a limited capacity system allowing the temporary

storage and manipulation of information necessary for such complex tasks as comprehension, learning and reasoning" (Baddeley 2000: 417). As illustrated in Figure 18, the model presumes that system to comprise of four components: a supervisory system termed c*entral* e*xecutive*, which is responsible for the control and regulation of cognitive processes, and three slave systems, namely the *phonological* l*oop*, the v*isuo-spatial* s*ketchpad*, and the *episodic buffer*. The episodic buffer, which is the latest addition to the model, is dedicated to linking information across domains in order to form integrated units as well as organizing the chronological ordering of these units. It is also assumed to have links to long-term memory (cf. Baddeley 2000 for a discussion). As the name already suggests, the visuo-spatial sketchpad is dedicated to hold visually perceived information such as the shape, color, location and the speed of an object (cf. Baddeley 2007: Ch. 4). The exact nature of these components need not bother us here as it is the remaining component—the phonological loop—that may help us illustrate the differences between visual language processing (reading) and auditory language processing. The phonological loop (or *articulatory loop*) is dedicated to holding phonological information and in turn consists of two functionally distinct components (not shown in Figure 18): a *short term buffer*, which is capable of storing auditory memory traces, and an *articulatory rehearsal component*. Auditory verbal information will enter the short term buffer creating an auditory memory trace, which is subject to rapid decay. The function of the rehearsal component is to revive such traces, but there are limits to such revival processes. Now, it is important to note that visually perceived linguistic information (information from reading) is assumed to be transformed into a phonological code. Visually perceived linguistic information is thus thought to enter the phonological loop as well and is kept in memory by the very same machinery. So while the processing of visually and auditorily perceived linguistic information is strikingly similar (at least in this model), there are of course tremendous differences in the way this information is presented to the individual. Readers can "re-fill" their phonological loop at will due to the fact that the input is externally represented and thus is constantly available. That is to say they can just go back to the relevant material in case the memory traces from the first-pass reading cannot be revived. The self-paced nature of reading also allows for a well-timed transformation and sequencing of the information. These differences in the initial conditions

across the modalities (+/- controllable rate of input and +/- external representation of that input) can be seen as a strong cognitive motivation for the modality specific utilization of linguistic structures as well as differences in the usage frequency of structures that are used in both modalities. The processing perspective advocated here predicts that whenever there are notable differences across registers, the easier forms should be observed with spoken language.

It is interesting to note that different variables have been investigated in different research contexts in the attempt to answer different questions. Most prominently these questions either concern the resolution of local syntactic ambiguities or revolve around issues concerning the role of (syntactic) complexity for language processing. Relative clauses play a pivotal role in both domains: they are relevant for the former issue as they figure in one of the best studied of local ambiguities, the *MV/RR ambiguity*, which is demonstrated in (55), which arguably is one of the most famous sentences in recent linguistic history.

(55) The horse raced past the barn fell

The expression in (55) is locally syntactically ambiguous because at the time the word *raced* is perceived it is yet unclear whether it is the main verb (MV) of the subject, *the horse*, or a participle in a reduced relative (RR) clause, which—reading on—turns out to be the correct interpretation. The importance of relative clauses is even more pronounced when we turn to research targeting linguistic (or syntactic) complexity. In this domain, psycholinguistic research has tried to disclose complexity related parameters that can account for the observed processing differences between subject and object relatives, the difficulties associated with center embedded structures (as opposed to right embedded ones), and the conditions under which non-obligatory R-elements can be omitted. The present study shares the perspective presented in Traxler et al. (2002) and Gennari and MacDonald (2008) that there are reasons to combine the insights gained in these domains of research and describe all these phenomena from a unified theoretical perspective. Specifically, it appears very promising to investigate both, syntactic ambiguity resolution and complexity issues, from the viewpoint of constraint-

satisfaction accounts of language processing (MacDonald 1994, 1999). We will return to constraint-satisfaction models and connectionist language processing in § 3.4.2.1, when we discuss the view on working memory that follows from the general architecture of such models. Gennari and MacDonald (2008) propose a view from constraint-satisfaction models and argue for

> "[…] *an approach to relative clause comprehension within a constraint-based approach to ambiguity resolution, in which comprehension difficulty is a function of the amount of indeterminacy in the sentence at various points in time. This indeterminacy is itself a function of the extent to which lexical and other constraints converge to promote a single, ultimately correct interpretation.*"
>
> (Gennari and MacDonald 2008: 18)

As the present study tries to add to the force of this line of argumentation and complement the experimental findings with evidence from ecologically more valid corpus data, we will return to the theoretical position, its predictions as they pertain to corpus observations and a more detailed discussion of the relation between experimentally and observational data in succeeding sections, when we turn to more specific contrasts and their investigation.

Before we start our presentation of the results, a last preliminary comment is in order. Following a growing trend in statistics, the present study emphasizes the role of visualization in data analysis as graphical representations are generally much more efficient in the communication of ideas than numeric ones. For interval-scaled data, graphical methods are well-developed and widely used; magnitudes and relationships among variables can be represented using scatterplots with trendlines and the like. For categorical data, however, graphical methods are not so well-developed and researchers working with such data types rarely ever go beyond using barplots and piecharts. However, there is a growing body of research into data visualization (as an integral part of data analysis) trying to develop the graphical repertoire following design principles of perception, detection and comparison.

Most notably, it has been shown that frequencies are most usefully represented as areas (Friendly 1994, 1999). Given their scarce deployment in the field of linguistics, some of the representational techniques used in this study (e.g. extended mosaic plots for *n*-way contingency tables or association plots for 2 x *k* contingency tables) are likely to be unfamiliar to the reader—in fact many of them have not yet found their way to commercial statistical software solutions. These techniques will be explained as they are introduced.

3.2 Head features

We will begin our discussion with a descriptive overview of a number of syntactic and conceptual features encoded on the head. The idea that the processing demand of a sentence is influenced by the type(s) of NP it incorporates has been investigated by many researchers in various contexts and we will discuss these in some detail in § 3.5.4. We may assume that the complexity of a sentence is influenced a) by a set of properties that can be read off from the linguistic form of a given NP and b) by possible interdependency effects of consecutive NPs with certain RCC types, i.e. non-subject relatives. The intrinsic properties of NPs have of course logical priority over properties that involve minimally two NPs, so we may start our discussion with a look at the element of the main clause that receives clausal post-modification, i.e. the head NP.

Linguists have long pointed out a close connection between the linguistic form of an NP and the salience of its referent in the ongoing discourse (e.g., Du Bois 1980, Sanford and Garrod 1981, Givon 1983, Gundel et al. 1993, Chafe 1994, inter alia). While individual accounts may differ at some level of detail, there certainly is considerable agreement on a more general level. We shall use the *Givenness Hierarchy* (Gundel et al. 1993) for illustration.

in focus	> activated	> familiar	> uniquely identifiable	> referential	> type identifiable
{*it*}	{*this*, *that*, *this* N}	{*that* N}	{*the* N}	{indefinite *this* N}	{*a* N}

Figure 12: Givenness Hierarchy (Gundel et al. 1993)

As illustrated in Figure 19, the use of a pronominal NP constitutes strong linguistic evidence

for the idea that the corresponding referent is *given*, i.e. already established in the discourse. The further we move down the hierarchy, the less given is the corresponding entity in the discourse. In the terminology of language processing, this translates into saying that the information expressed by that NP is active in the interlocutors' consciousness and hence easy to access (Sperber and Wilson 1995, Ariel 1990).

Before we can apply these findings in the context of our discussion of RCC, we should note that such generalizations are robust if and only if the NP is question is a referring expression (like the subject NP of the RC for example). The head in a RCC, however, is not by itself a referring expression, at least not in those cases in which the RC is restrictive/defining. In such cases the actual referring expression is the NP dominating the head and the RC and includes the property ascription(s) within the RC proper. This is arguably true even for cases where the head is a personal pronoun as in *he who is without sin*. Semantically speaking, such expressions do not qualify as definite descriptions because they do not presuppose a referent that uniquely meets all the conditions asserted in the complex expression. Rather, we may treat these expressions as (intensional) definitions of a class of objects, i.e. a type. Consequently, our discussion of the accessibility status of an NP (referent) is not fully applicable to the head nominal. But the information encoded on the head is nonetheless very important for an assessment of the overall processing demand of a RCC: The experience-based approaches to language processing that serve as a theoretical backbone of this study assume that language users store and make use of many correlational structures in their knowledge of the language. As heads precede their clausal modifiers in English, we may presume that language users take advantage of certain regularities in the language and use the information encoded on the head to predict likely structural continuations. As we have mentioned earlier, one such piece of helpful information involves the animacy of the (referent of the head). Recall that we have discussed the idea that an inanimate head is more like to be followed by an object relative because a) inanimate objects are more likely to play patient roles than agent roles and b) lower thematic roles tend to occupy lower grammatical roles/functions. These and similar cues have been argued to influence the degree to which hearers can anticipate an upcoming relative clause. Some researchers have argued that the processing difficulty of a RCC can be understood as a function of the *predictability* of the RC

(cf. Jaeger et al. 2005). This first discussion of current views on processing difficulty is best taken as a precursor of a more thorough discussion. We shall return to these and other issues in later sections, when we look at the results of the multivariate procedures. At this stage, we are mainly concerned with a description of the data set and certain distributional facts. Let us now turn to the head feature discussed in this section.

The following variables will be considered in the order of presentation in the list below (with corresponding variable labels given in SMALL CAPS):

- Morphosyntactic realization of the head (HEAD.TYPE)
- Definiteness of the head NP (DEFINITENESS.HEAD)
- Semantic specificity/generality of the head (CONTENT.HEAD)
- Conceptual animacy (ANIMACY.HEAD)
- Concreteness of the head (CONCRETENESS.HEAD)[14]

3.2.1 Morphosyntactic realization of the head

The first distribution to be shown concerns the morphosyntactic type of the head. The contrast of interest for this factor was the distinction between lexical nouns (including proper and common nouns), which—semantically speaking—label individuals or classes of objects, and pronominal heads, which are semantically rather vague and thus more flexible in their application. Two sub-types of pronouns that can function as heads of an RC were distinguished: indefinite pronouns as in $_{NP}$[*anyone* $_{RC}$[*who*…]] and demonstrative pronouns such as in $_{NP}$[*those* $_{RC}$[*who*…]]. This distinction does not capture all possible types of pronominal heads: in principle English also allows personal pronouns to function as heads as

[14] The variable ANIMACY and CONCRETENESS concern properties of the corresponding referent representations (as opposed to linguistic properties).

in constructions like *He who does not understand your silence will probably not understand*. However, such constructions are rather infrequent in contemporary English and there was in fact not a single example of such a pattern in my data set. Nevertheless, a specific query for these patterns in the ICE-GB corpus, i.e. searching for {*I*|*you*|*he*|*she*|*it*|*we*|*they*} {*who*|*that*|*which*}, does in fact match a handful of examples. At any rate, the pronominal types in the data set, indefinite and demonstrative pronouns, certainly do not by themselves allow for a direct identification of their referents as they encode only very general semantic information. Basically, what the they encode is restricted to whether it is a single entity that is referred to or more than a single entity and whether or not the entity referred to is human.

With these general considerations in place, we may now turn to the distribution of these morphosyntactic syntactic types across the complete data set (Figure 20) and the modality specific sub-sets (Figure 21).

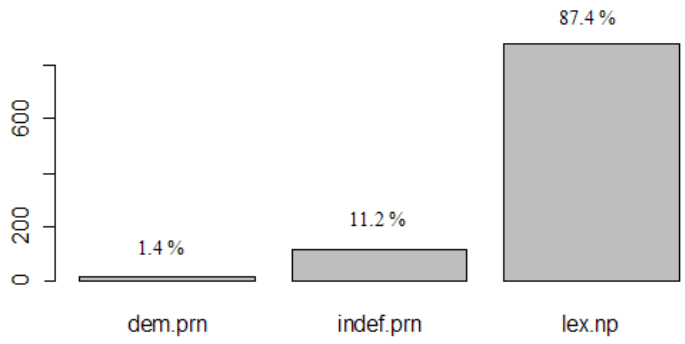

Figure 20: Syntactic type of head noun

It should come as no surprise that the lion's share of heads are lexical nouns (874/1000); only (1.4 % + 11.2 % =) 12.6% of the heads in the total data set (n=1000) are pronominal. The majority of pronominal heads are indefinite pronouns, which belong to a class of expressions referring to sets of unspecified inanimate objects (most notably *any-*| *some-*| *everything*) or persons (most notably *any-*| *some-*| *everybody* or *any-*| *some-*| *everyone*). Demonstrative pronominal heads—though possible and attested—are rather rare (14/1000). Figure 21 presents a graphical display of these distributions for both registers separately.

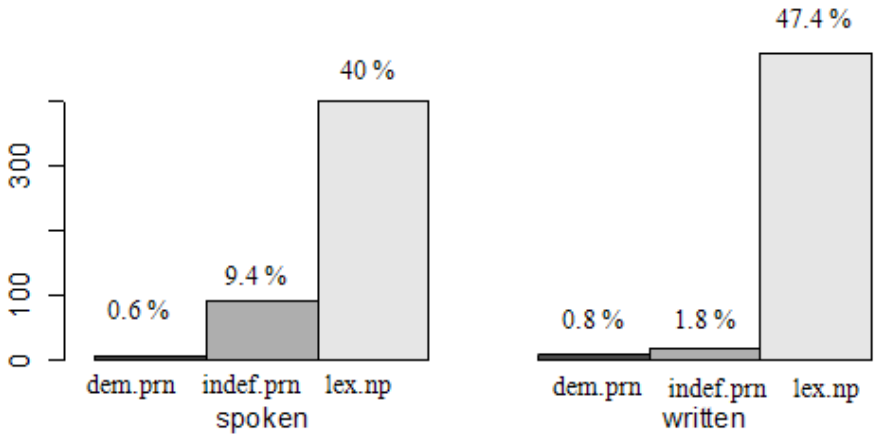

Figure 21: Syntactic type of the head across modalities

A global chi-square test of independence, which assesses the relationship between the variables HEAD.TYPE and MEDIUM yields a probability p = 2.392e-13 ($\chi 2 = 58.1226$, df = 2).

When we look at individual cells, i.e. at the individual chi-squared components (or residuals), to see which factor levels combinations contribute most to the χ^2 sum, we learn that the greatest contribution results from the difference in the number of indefinite pronouns. As shown in Table 4, the null hypothesis predicts that we should find 56 cases of indefinites in both modalities, but what we find is that 94 of the 112 cases are from the spoken part of the data set.

Table 4: Expected counts (HEAD.TYPE X MEDIUM)

		Medium	
		Spoken	written
head.type	dem.prn	7	7
	indef.prn	56	56
	lex.np	437	437

The deviations from the other cells are far less pronounced: as can be seen in Table 5, the modality specific difference of occurrences of indefinite pronominal heads accounts for 2 x 25.79 =51.58 of $\chi 2$.

Table 5: Chi-square components (sum = 58.12)

		Medium	
		Spoken	written
head.type	dem.prn	0.14	0.14
	indef.prn	**25.79**	**25.79**
	lex.np	3.13	3.13

A convenient graphical display of the distributional differences is shown in Figure 22.

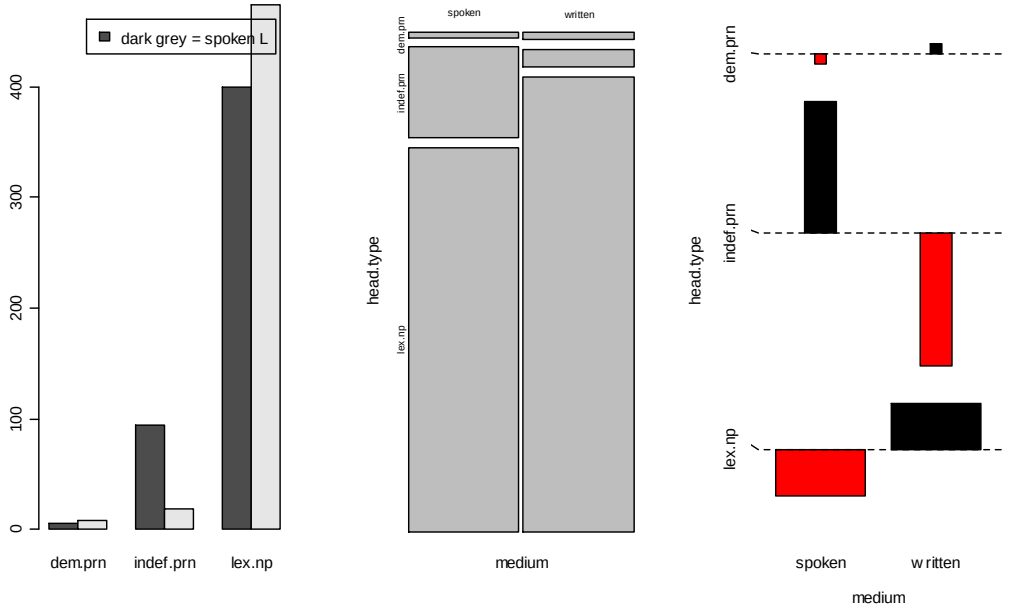

Figure 22: Syntactic type of head across modalities (plot-triple)

Figure 22 presents from left to right a triple of plots consisting of:

I a BARPLOT presenting the observed frequencies of the contrasted factor levels across modalities (with dark bars representing the spoken medium)

II a MOSAIC PLOT, which iconically represents the values of the cells of a contingency table, and

III a COHEN-FRIENDLY ASSOCIATION PLOT, which indicates deviations from independence

Whereas the barplot offers a quick and familiar visualization of the frequency contrast in question, the mosaic plot presents this information proportionally. The third plot, i.e. the association plot, helps us see possible interactions between the factors crossed. For a two-way contingency table, the signed contribution to Pearson's χ^2 for cell *i, j* is *d_{ij} = (f_{ij} -*

e_{ij}) / sqrt(e_{ij}), where *f_{ij}* and *e_{ij}* are the observed and expected counts corresponding to the cell. In the Cohen-Friendly association plot, each cell is represented by a rectangle that has a (signed) height proportional to *d_{ij}* and width proportional to *sqrt(e_{ij})*, so that the area of the box is proportional to the difference in observed and expected frequencies. The rectangles in each row are positioned relative to a baseline indicating independence (*d_{ij} = 0*). In a nutshell: if the observed frequency of a cell is greater than the expected one, the box rises above the baseline. The plot thus presents all local differences separately providing immediate access to the information regarding a) which factor levels are most distinctive and b) the direction of the difference due to the signed height. In the example above, the plot triple helps us understand very quickly

a. the absolute (▷BARPLOT) and relative (▷MOSAICPLOT) frequencies of all factor level combinations
b. that the distribution is far from even, so that a global chi-square test is likely to yield significant results)
c. that the difference across modalities is most pronounced for indefinite pronouns (▷assocplot)
d. that indefinite pronominal heads are more common in spoken discourse than in written discourse and also more common than expected (under H0), whereas the inverse is true for the other factor levels (▷assocplot)

More generally, looking at the graphical models, we have good reason to suspect an interaction between the variable HEAD.TYPE and MEDIUM. We can test for such interactions using an ensuing significance test of interaction in 2 x *k* designs with proportions (cf. Marascuilo 1970). The Marascuilo procedure tests the hypothesis of equal proportions and proportion differences in 2 x *k* designs. The resulting test statistic Q' (= Q bar) corresponds to the sum of proportion differences and—very much like Q in a $\chi 2$ test—has a $\chi 2$

distribution with (k-1) degrees of freedom. The value for Q' is thus used to determine corresponding p-value. The procedure can be used for any sample size and even when the data contain extreme proportions. When we apply the Marascuilo procedure to the HEAD.TYPE x MEDIUM data, we learn that from the three comparisons (INDEF.PRN-DEM.PRN, INDEF.-PRN-LEX.NP, DEM.PRN-LEX.NP), only the difference between indefinite pronominal and lexical heads is statistically significant (at $\alpha = .01$) and this is exactly what we suspected just looking at the graphical models. This data visualization, which employs a combination of bar-, mosaic- and association plots, thus permits a convenient quick-and-easy way to analyze such data types and will be used for a number of similar scenarios throughout this study.

Recall that investigating modality specific differences not only is considered to be important for descriptive purposes, but has direct implications for psycholinguistic theorizing and experimental protocols: the greater the difference across modalities, the more problematic become course-grained estimations of frequency based expectations, which lie at the very heart of usage-based accounts of linguistic knowledge and processing. Given the importance of this issue, we will continue to address it in all subsequent sections. We can immediately see an important difference across registers when we look at the lexical realization of the head. As shown in the pie chart in Figures 23, there is a high degree of heterogeneity in lexical choices in written language.

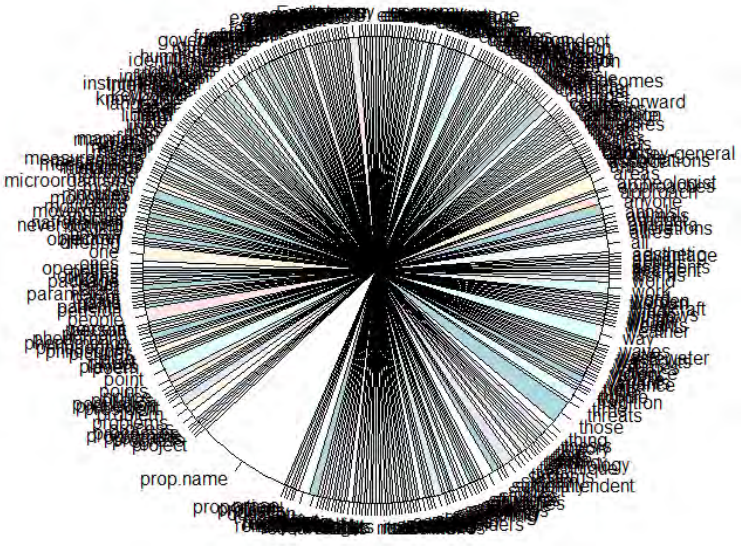

Figure 23: Lexical choices at the head position (written discourse)

The only type of head that clearly sticks out due to its high token-frequency is the type PROPER NAME, which is of course actually a set of types as it subsumes numerous lexical units. When we "zoom in" on the set of lexical heads that are proper names and describe their referents in terms of concreteness and animacy, we get the results presented as Table 6:

Table 6: Animacy and concreteness of referents of 'proper name'-heads in written discourse

		concrete		
		no	yes	
animate	yes	0	15	15
	no	14	3	17
		14	18	32

While 15 out of 32 heads refer to animate objects, which by implication are concrete, we also find that in more than 50% of the cases the name refers to an inanimate object (17/32) and that in most of these cases (14/17) the referent is an abstract entity. Typically, the RCs modifying these heads are employed to describe the respective object(s) allowing the hearer to construct a concept of the intended referent on the basis of that description. Describing the referent object by way of a subordinate clause does, of course, still allow saying something about this object in the main clause prediction. The result is a semantically quite dense bi-

clausal construction. An example is given in (56):

(56) Automated **NDT using specialized mechanical scanners or ROVs** has a number of attractions for land-based and offshore applications. (W2A-038 #029:1)

If we take away heads that are realized as proper names, we observe a highly variable set of lexemes occupying the head position. The remaining 472 usage events in the written medium distribute across 354 types and the second most frequent type, *those*, occurs only 8 times in the data.

In contrast to these characteristics, spoken discourse presents a completely different picture. Spoken head realization is dominated by a small set of types that occur with rather high token frequencies. Figure 24 presents an overview.

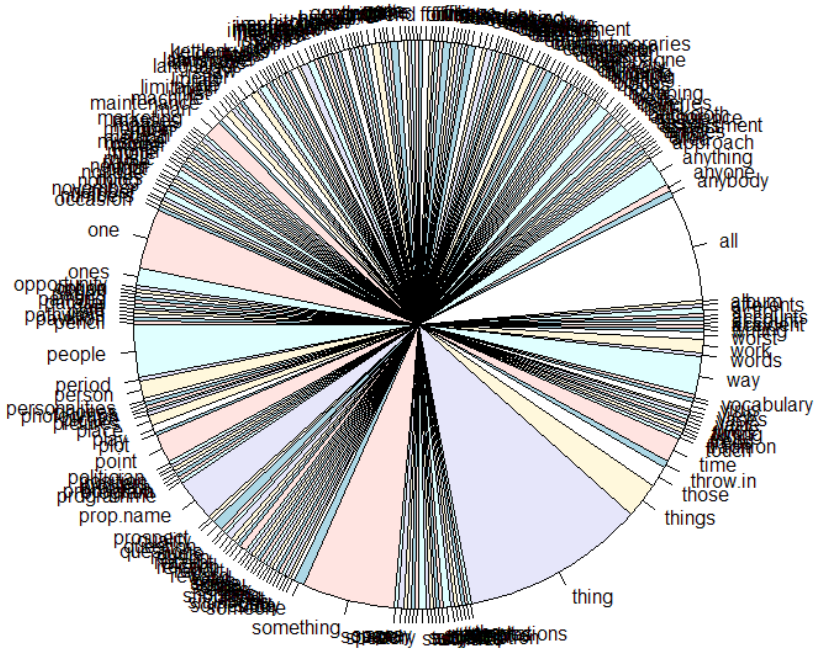

Figure 24: Lexical head choices (spoken discourse)

There are 249 different types of lexical realizations in the spoken part, which making no assumptions about lexical frequencies and given that we are looking at 500 instances, would

have us expect there to be roughly two tokens for each type. Figure 25 shows the 25 most frequent types.

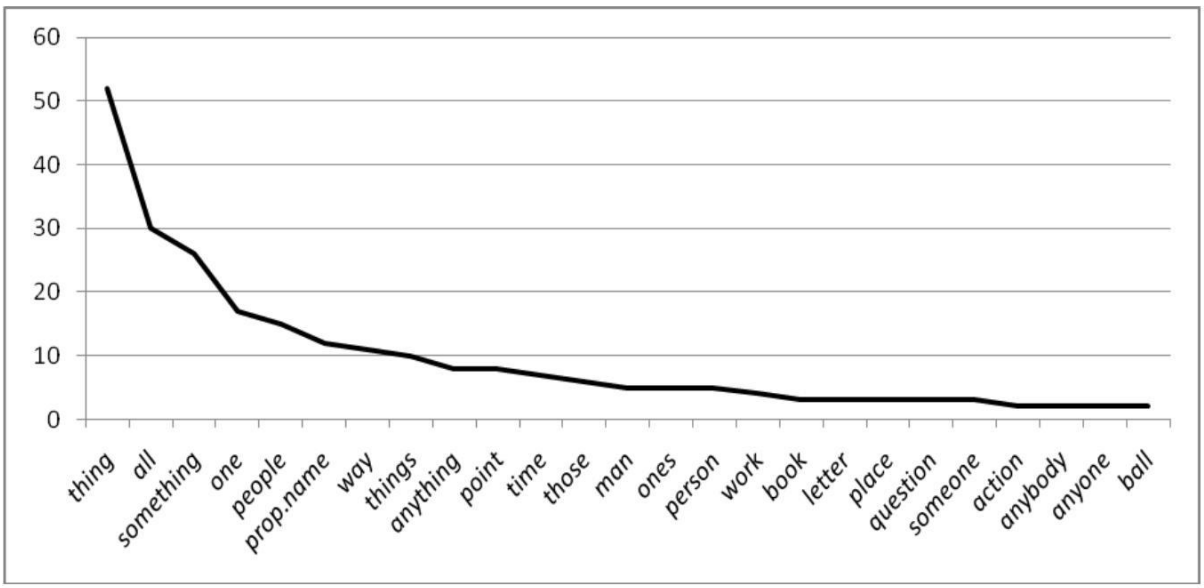

Figure 25: Graph of the 25 most frequent lexical realizations of the head (spoken discourse)

The lemma *thing* alone occurs 62 times in the data, which of course is clearly above our expectation. There is no need to discuss the exact numbers here; suffice it to say that that those heads whose labels we can read in Figure 25 apparently enjoy a privileged status in so far that they occur with rather high token frequencies. Notice that these frequent types (*one, ones, people, prop.name, something, thing, things, those, time, all, anything, anyone*) denote very general classes of objects, i.e. they are semantically rather unconstrained. If we combine the token frequencies of the abovementioned types, we learn that they jointly account for (206/500 =) 41.2 % of all attested cases. Hence, we must note that referring to objects by way of restricting the denotation of a general term with a clausal modifier apparently enjoys a quite respectable popularity in spoken discourse.

3.2.2 Definiteness of the head

The next variable in our discussion concerns the definiteness of the head. A head was treated as definite if the entity referred to was specific and identifiable in a given context of utterance. Formally this is often marked either by the presence of a definite determiner or by a head that is realized as a demonstrative/personal pronoun or a proper name. Strictly

speaking, these formal criteria are not sufficient to determine the value for definiteness as a semantic property. Definite NPs are indeed often used to single out an actual instance of a certain type denoted by the head noun, but this need not be the case. We can easily see that the presence of a definite DPs (or lexical NP introduced by a definite determiner for that matter) is not sufficient to determine the value for definiteness, when we consider a simple example. In a context where you are over at your friend's place and she shows you her cat, you might utter something like $_{NP}$[*The cat*] *is very cute*. In this context, the expression *cat* can be said to denote a certain class of objects, i.e. the set of all cats, and the definite article quantifies this set so that the expression *the cat* can be used to refer to an actual member of the class of cats. This, however, is not necessarily true of all usages of definite NPs. Langacker—following Fauconnier—has discussed a class of counterexamples under the label of "virtual" or "fictive instance" (Langacker 2007:55). An example would be an expression like *The general's limousine keeps getting longer*. In this usage there need not be an object that actually changes its shape. In fact this interpretation is not very likely to be the preferred one. Such examples were not counted as definite because they differ from true definite descriptions in that they do not presuppose there to be one and only one object that satisfies all predications φ in that description, symbolically ($\exists x \, \forall y \, (\varphi \, (y) \leftrightarrow y = x$). This distinction is important because a definite article can be said to presuppose uniqueness in a context of utterance, if and only if it is part of a true definite description and it is exactly this presupposition that raises the likelihood of a restrictive (clausal) modifier in the first place. We will return to the issue of asserted uniqueness and RC likelihood, when we discuss the presence of a *uniqueness adjective*. At this point we may appreciate this subtle difference in usage of the definite article and acknowledge the value of linguistically informed manual inspection of the data.

Having settled on a treatment of definiteness, the same statistical procedures used in the discussion of the variable HEAD.TYPE was applied to evaluate the distribution of this second head variable. We arrive at the results presented in Figure 26.

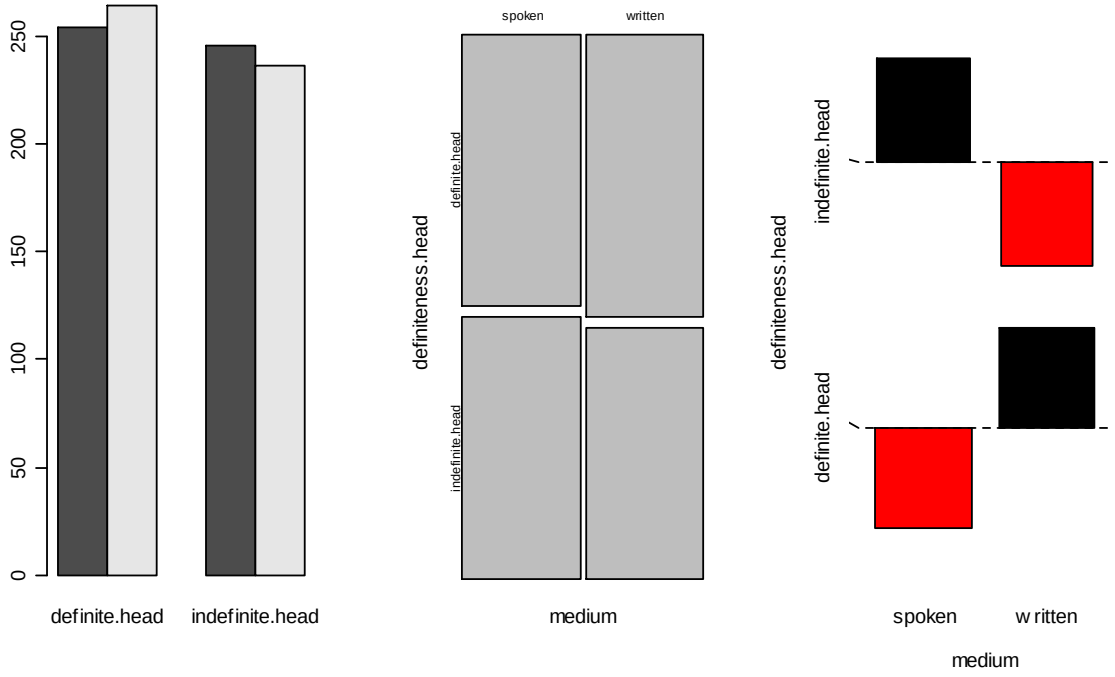

Figure 26: Definiteness of head

For the variable definiteness of the head, we observe that the data provide no evidence against the null hypothesis, which denies any meaningful relationship between DEFINITE.HEAD and MEDIUM ($\chi 2 = 0.4005$, df = 1, p = 0.5268, all comparisons of proportional differences are ns). So the definiteness of the head does definitely not discriminate between the modalities.

We may conclude that speakers use definiteness marking to signal presupposed availability of the NP referents in the discourse irrespective of whether the discourse is conducted via the auditory or visual channel. The information structural constraints on definiteness marking apply to both modalities.

3.2.3 Contentfulness of the head

The next variable of interests concerns the 'contentfulness' or semantic richness of the head. It has been suggested (e.g. Jaeger et al. 2005) that the semantics of the head have an effect on the likelihood of a clausal modifier to follow the head. Specifically, it has been argued that

the more general the term is, the greater is the need (and hence the likelihood) of additional material which allows the hearer to identify the intended referent of the complex NP. In other words, the more general the applicability of the term, the greater is the need to narrow down its scope of predication by way of a relative clause.

An expression was taken to be contentful, if it entails a large number of semantic features. The more properties are entailed by a term, the greater is its contentfulness. Even though intuitively this distinction appears tenable, it is of course difficult to rigidly determine the number of features implied by a given term. As a first approximation, we may use entailment tests to identify good candidates of meaning components. Regardless of what stance is taken towards (lexical) meaning, there seems to be little room for denying that the truth of the statement x *is a Honda Civic* necessitates the truth of the more general proposition x *is a car* and hence x *is a concrete object*. At least if we exclude clear cases of highly non-conventional usages of the expression *Honda Civic*. Given the delicate nature of such semantic decisions, contentfulness was judged very conservatively so as to include only the clearest cases. Negative values (=low content) were assigned only if the head was judged to be two general as to possibly fix reference. This led to the following extensional definition: If the term in question is in the following list, it received the value *low.content/generic*. If the term in question is not in that list, it received the value *high.content/contentful*:

List of attested types judged as generic:

adult, all, anybody, anyone, anything, best, chap, everybody, everybody, everything, father, girl, guy, information, man, masses, matters, means, men, more, no one, nothing, one, ones, people, person, persons, place, point, road, somebody, someone, something, son, stuff, that, thing, things, those, time, two, uncle, way, what, woman, women, worst

Figure 27 presents the resulting distributions.

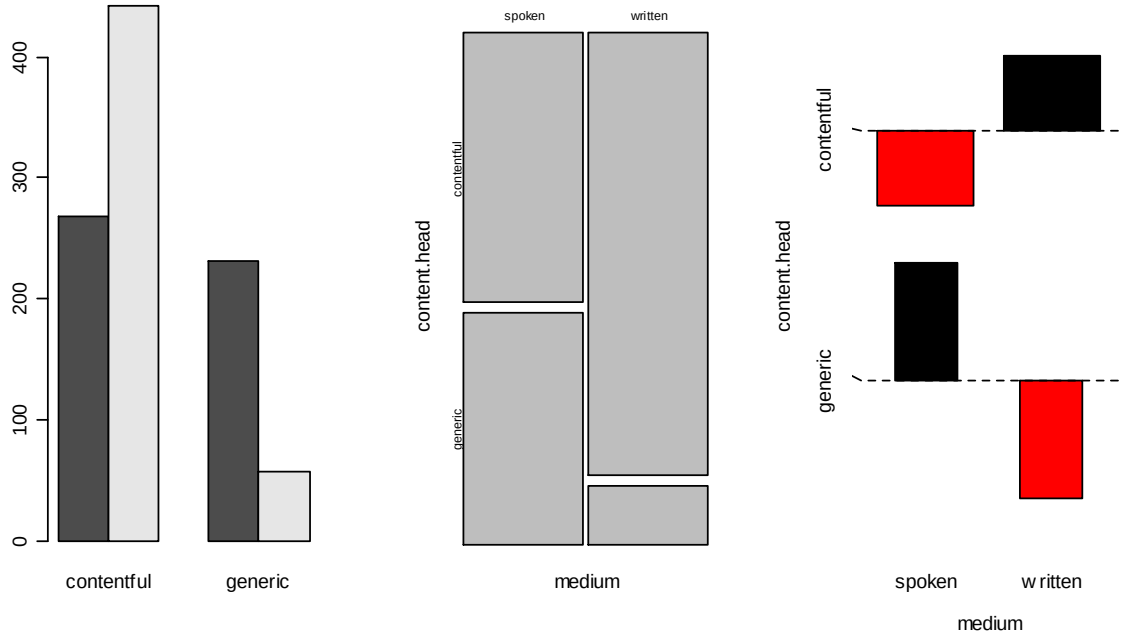

Figure 27: Contentfulness of head

We can see that there is a strong difference between spoken and written relative clauses constructions with respect to this variable: 231 out of 289 generic heads (=tokens) are from the spoken part of the data set suggesting a very strong association between those heads and that modality ($\chi 2 = 145.6548$, df = 1, $p < 2.2e\text{-}16$, proportional difference (Q') is significant at $\alpha = 0.01$).

The reason for this striking difference may be grounded in the fact that the definitional character of relative clause construction—general term followed by scope restriction—constitutes a very productive way of introducing new discourse referents. It may also be grounded in the fact that speakers pay more attention to the information structural properties of their linguistic output in spoken discourse than they (need to) do in written language. We will return to this important issue in Chapter 4, when we get to see a more detailed picture of the preferred patterns in the respective modalities.

3.2.4 Animacy of the head

This variable concerns the animacy of the head as a property of the referent of the complex

NP. Strictly speaking, it is not a property of the referent per se, i.e. some non-mental entity—say a real-world object, but a *referent representation*, i.e. another mental construct of a ontologically different type (cf. Pecher and Zwaan 2005: Ch. 1). However, at this point of the discussion nothing hinges on this difference. Figure 28 presents the distributional differences.

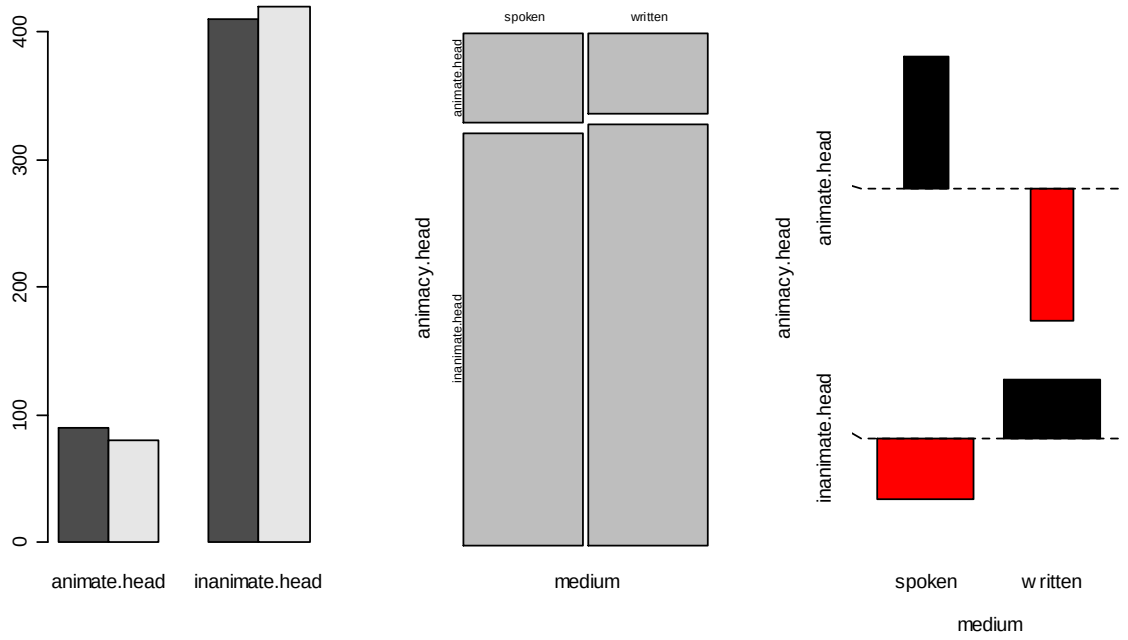

Figure 28: Animacy of the head

Again the variable was measured as a binary factor and again the plot-triple turns out to be very useful: while the association plot on the right suggests that there might be a difference across modalities and indicates a potential direction of the effect, both the barplot and the mosaicplot rather suggest that the difference is not very pronounced. Calculating the Marascuilo test reveals that the difference of proportions is in fact not significant (Q' = ns; also: X-squared = 0.5768, df = 1, p-value = 0.4476).

In both modalities the modified nominal typically is inanimate. Animate referents are slightly more frequent in spoken discourse, which is probably more a fact about what people like to talk about than a fact about human processing preferences.

3.2.5 Concreteness of the head

The final variable regarding the head describes the property of the NP referent being a concrete or abstract entity. Coding again was binary such that entities that are perceivable by the human senses were treated as CONCRETE and entities not so perceivable were considered ABSTRACT. Figure 29 presents the resulting distribution across modalities.

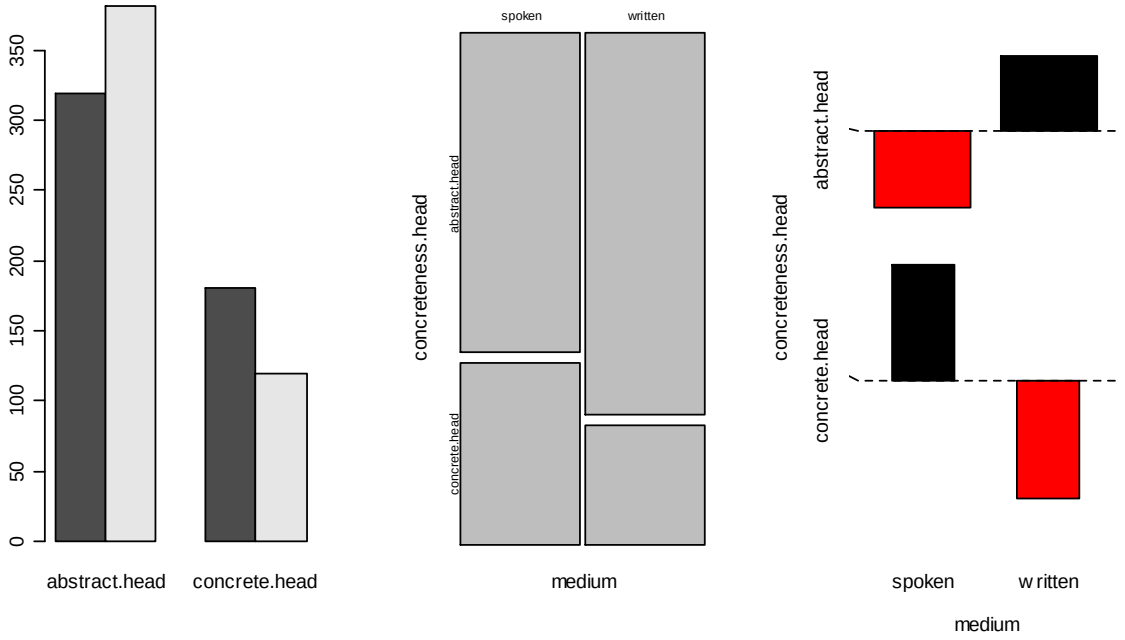

Figure 29: Concreteness of the head

Again, we observe a significant difference across registers to the effect that there is a tendency to prefer concrete heads in spoken discourse and abstract heads in written discourse ($\chi 2 = 18.3048$, df = 1, p-value = 1.882e-05, Q' < 0.01).

The information expressed in mosaic plot and the association plot can be unified into a single graphics using so called 'extended (shaded) mosaic plot' (Figure 30).

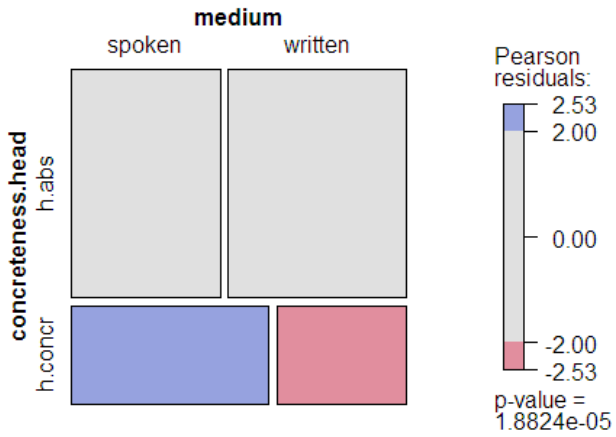

Figure 30: Concreteness of the head across modalities (shaded mosaic plot)

The coloring (or shading) of individual cells allows for a graphical representation of the degree of deviation of an observed cell from its expected value (under H0 = complete statistical independence). As indicated by the legend to the right of the plot, it is required that the contribution to chi square (Pearson residuals) has to be greater than 2.0 for these data in order to be marked as significant.[15]

Because of the implicational relationship between the factors animacy and concreteness, we may consider the difference at this level to be a more coarse-grained expression of a conceptually similar contrast. And it is at this more general level that we observe an important difference between what is typical for spoken and written discourse respectively. As with the (statistically insignificant) greater proportion of animate heads in spoken language, it appears plausible that the greater proportion of concrete heads in this modality can also be traced back to typical topics in these registers. We may suspect that the observed differences reduce to differences in formality of the genre. Figure 31 presents an overview of the donor genre of the usage events in this analysis. This information was automatically retrieved from the ICE annotation and will thus not be justified here (cf. Nelson et al. 1996 for an explication of the genre classification).

[15] The Pearson residuals for each cell are ($F_{Obs} - F_{exp}$/ sqrt(F_{exp})).

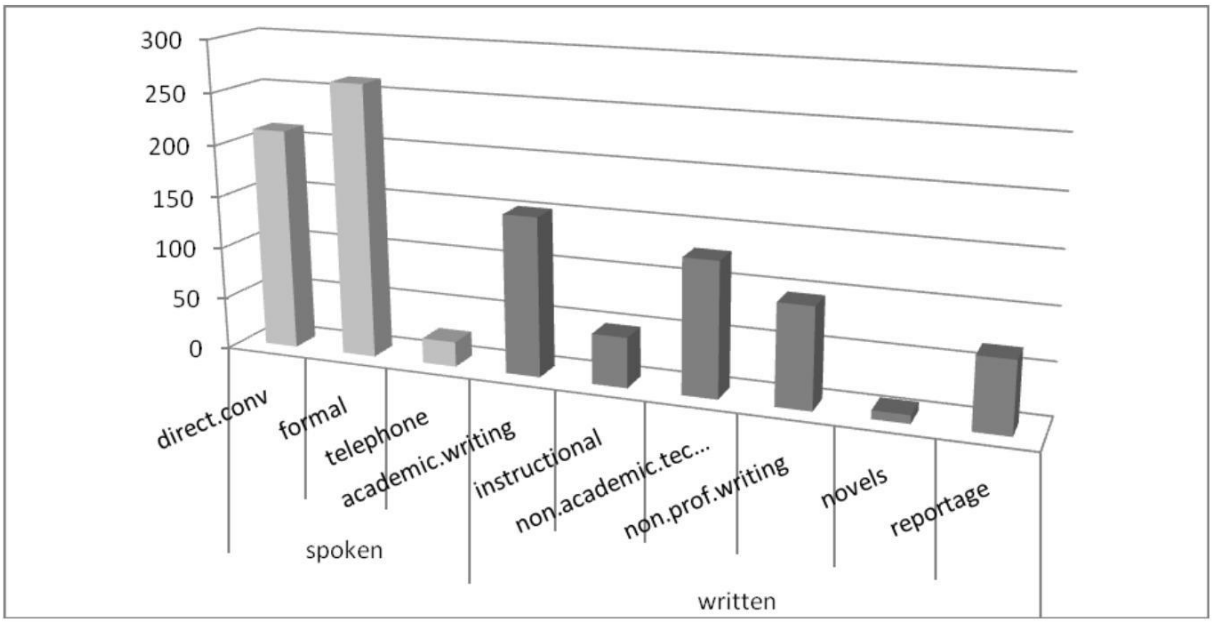

Figure 31: Composition of in spoken and written data (genre)

For the sake of the argument, we may derive the variable FORMALITY from make-up of the registers. For spoken language (columns in light grey), we may contrast formal texts from more informal texts by way of lumping together the types DIRECT CONVERSATION and TELEPHONE CONVERSATION so and contrast them with the type FORMAL. For the written modality, the dichotomization was achieved by collapsing on the one hand the types ACADEMIC WRITING and non-academic technical writing, which are very formal, and on the other hand the types NON-PROFESSIONAL WRITING, INSTRUCTIONAL, REPORTAGE and NOVELS, which arguably are less formal. Figure 32 presents the results:

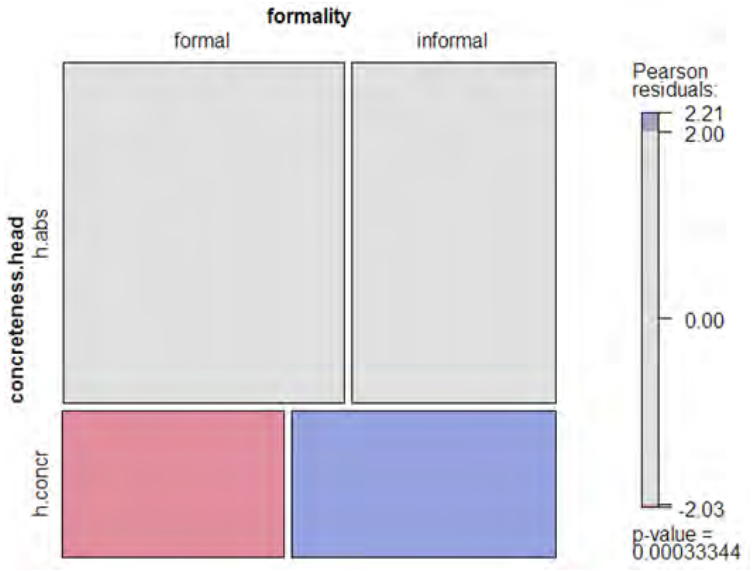

Figure 32: Concreteness of the head versus formality of discourse

We observe that there is indeed a statistically significant association between the type of head and the formality of the discourse ($\chi 2 = 12.873$, df = 1, $p < 0.000333$). Note, however, that the association between spoken and concrete is more pronounced than the association between informal and concrete. Because the sample sizes are identical (n=1000) and we are interested only in the relative strength of association, we may use the respective p-values as a direct expression of association strength and need not bother about effect sizes (cf. Evert 2004, Wiechmann 2008b for a discussion). While it cannot be excluded that the approximation of degree of formality was too coarse to disclose true effects, it appears that there is more to the greater proportion of concrete heads in spoken discourse than what can be accounted for by genre effects. At this point we have reached the limits of what a can be inferred from a corpus study like the present one and we will confine ourselves with what we were able disclose. Discourse-pragmatic analyses with denser data bases and experimental designs promise to uncover more intricate details on the issue.

At this point we may summarize the results reported in this section. On a methodological level we have motivated the employment of extended (shaded) mosaic plots, which allows us to conveniently read off interesting associative relationships implicit in 2 x 2 tables. On a more content-oriented level, we have seen that

1. spoken languages employs significantly more indefinite pronouns as a morphosyntactic realization of the head
2. there is no statistically significant difference with respect to the definiteness of the head across modalities
3. there is a strong association between the lack of contentfulness of the head and spoken register
4. there is no statistically significant association between the animacy of the (referent of the) head and register
5. there is a positive associative relationship between concreteness of the (referent of the) head

The pair plot in Figure 33 summarizes these results in a single graphical display and also allows us to read of all possible pairwise comparisons of the factors investigated.

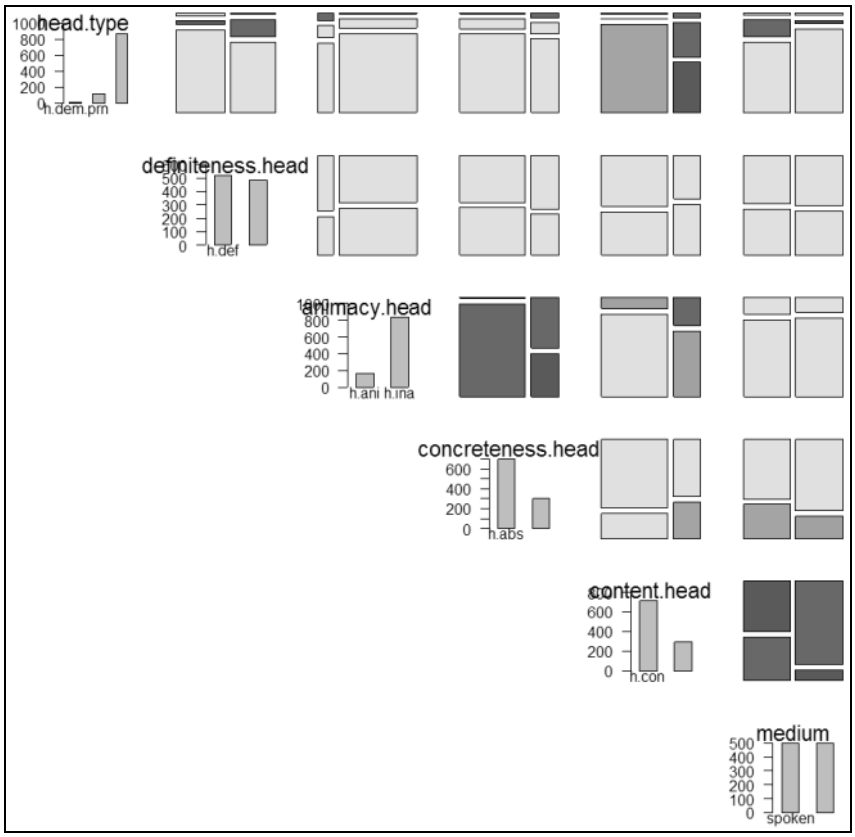

Figure 33: Pairs plot of head feature

A pairs plot for contingency tables provides a matrix of mosaic plot displays. On the diagonal we find all variables discussed so far and—in the lower right corner—the variable MEDIUM, which allows for a direct comparison of registers. Along with the labels of the respective variables, the diagonal also gives us the frequencies of the levels of a given factor. So, for MEDIUM we observe an even distribution, 500-500, which was of course controlled for. The mosaic plots above the diagonal represent the results of all logically possible factor crossings, so that a given plot represents the results of a crossing of the corresponding factors to the left and below that plot. Statistically significant associations are indicated by the shading in the cells. Hence, if we go through the right-most column, we can inspect the distributions of all five factors across registers.

3.3 Features of the relative clause

The next set of features that we will have a closer at concerns the relative clause proper. We will begin our overview with an assessment of the distributions of different types of relative clauses as defined along the parameter finiteness, i.e. we will look at the frequencies of finite and non-finite RCs. As we have already indicated in § 3.1.2, the relative processing difficulty is hard to assess a priori and the relative ranking is more dependent on the way the quantity processing demand is measured. The fact that the four different types of RC distinguished here, i.e. finite, *-ing* and *–ed* participial, and *to* infinitival relatives, do not really compete in their application, motivates that we also have a look at their distribution across different genres (text types). In doing so, we may infer some of the discourse functional potential of a given pattern from the usage patterns within a particular stylistic domain.

The second variable of interest aims at the complexity of the clausal constituents. The complexity of a clause was approximated by way of the transitivity value of that clause. Transitivity was treated as a grammatical reflex of the arity of the predicate. All other things being equal, we should expect verbs that express ternary relations, e.g. *give*, to impose greater demands on the processing system than unary ones, e.g. *sleep*. The processing hypothesis predicts that more complex patterns should be observed in the written modality.

The third variable presented in this section concerns the type of relative clause as

defined by the grammatical role played by the modified element within the RC. Again the processing perspective proposed here predicts an overall preference for the relativization of higher roles, which, however, is expected to be more pronounced in spoken discourse.

The final subsection in this domain will present a comparison of the here presented data with the results of a recent large-scale corpus analysis conducted by Doug Roland and colleagues (Roland et al. 2007). The main reason for this comparison concerns the representativity of the present data set. A high degree of isomorphism across samples is viewed as an expression of high degrees of representativity of the present data.

3.3.1 Grammatical features of RC: Finiteness

Like all subordinate clauses in English, relative clauses can assume different values concerning their finiteness: they can be finite or non-finite and if non-finite, RCs can assume participial (*-ing* participle and *–ed* participle) or *to*-infinitival forms. Let us start our discussion with an overview of the distributions of these sub-types across modalities. Figure 34 presents such an overview:

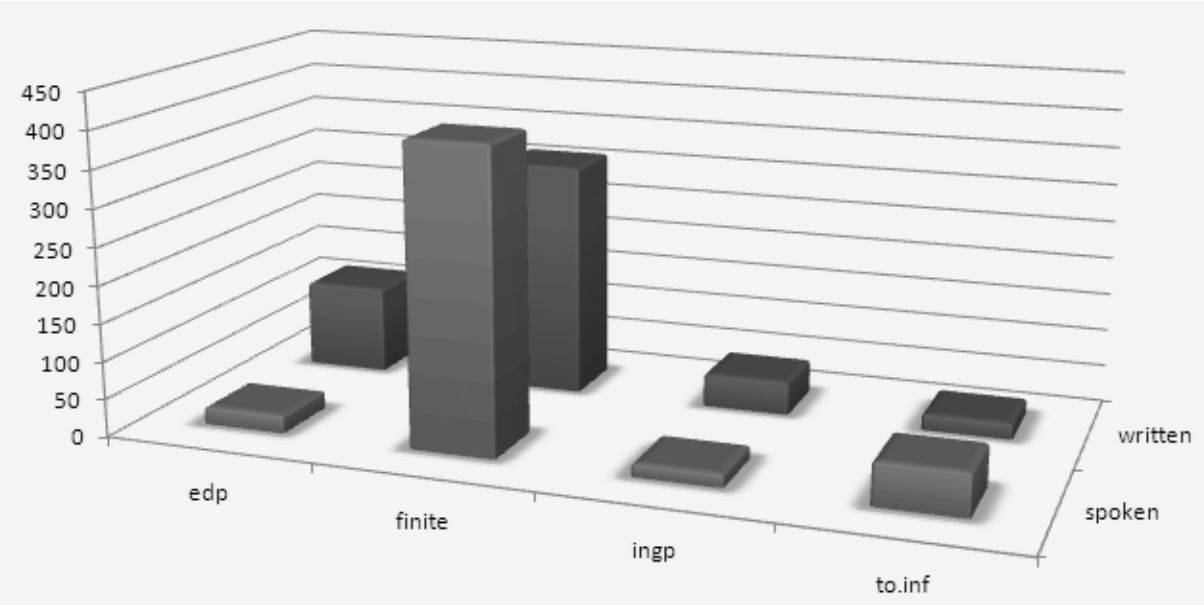

Figure 34: Types of RC (finiteness) across modalities

From this frequency distribution we can observe that 715/1000 RCs are finite, which makes this type the dominant one in both registers. As far as the non-finite types are concerned we

observe that written language is characterized by a stronger reliance on participial forms, whereas we find spoken language to exhibit a greater amount of *to*-infinitival constructions. A chi-square test allows us to disclose the respective associative relationships. Figure 35 presents the obtained results.

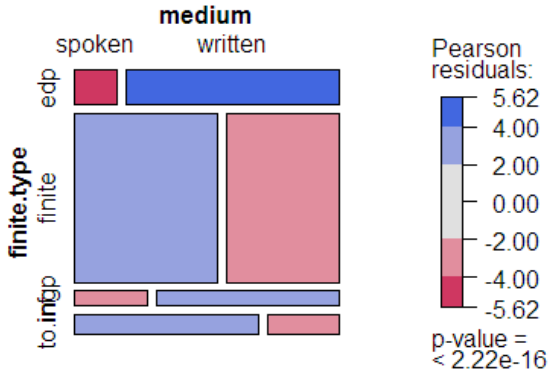

Figure 35: Types of RC (finiteness) across modalities (shaded mosaic)

We observe a positive association between written language and participial RCs in general. This association is particularly pronounced for *–ed* participial RCs. Conversely, we observe that finite and *to*-infinitival forms are associated with spoken discourse. The avoidance of *–ed* participle constructions in spoken discourse is in some sense expected from a processing perspective. The *–ed* participle is formally indistinguishable from past tensed verbs that function as the main verb of the preceding nominal and hence gives rise to a local syntactic ambiguity, namely the main verb/reduced relative clause ambiguity, which we mentioned earlier. The MV/RR-ambiguity has received a huge amount of attention in the sentence processing literature (cf. Ferreira and Clifton 1986, MacDonald et al. 1994, Trueswell et al. 1994, inter alia). It has been suggested that language users tend to avoid ambiguous forms (Temperley 2003) and the low frequency of spoken *–ed* participial RCs is (at least) consistent with that hypothesis. While an ambiguity avoidance principle may seem intuitively plausible, it has not received a lot of empirical confirmation. In fact, recent studies in this domain have either failed to detect the effects predicted by the principle (Ferreira and Dell 2000) or have even found effects in the opposite direction (Arnold et al. 2004). In a similar vein, we may say that the observed preference for finite types in spoken language is consistent with the general idea underlying the ambiguity avoidance principle. Non-finite RCs involve an

implicit argument, which for some sub-types must be recovered on the basis of non-linguistic/pragmatic knowledge. Consider the following examples of *to* infinitival RC for an illustration of this point.

(47) This is $_{NP}$[the book $_{RC}$[to read on a lonely island]] **Implicit role -> object**

(48) This is $_{NP}$ [the man $_{RC}$ [to climb the Mount Everest]]. **Implicit role -> subject**

(49) Intuitively $_{NP}$ [the first mechanism $_{RC}$ [to account for mass movement in the situation of the device structure]] would be electro migration, or diffusion or a combination of both. [W2A-035 #052]

There are no formal cues indicating the grammatical role of the head inside the RC, so the interpreter must use her semantic and pragmatic knowledge about the situation described by the RC predicate and infer the most plausible role for the head in that situation. The examples in (47) and (48) are fictive and were deliberately kept simple to emphasize the contrast of interest. The example in (49), however, is an actual example from the corpus and illustrates the amount of indeterminacy that these structures may introduce. At the time *for* is perceived it still unclear what type of RC we are dealing with. In fact, an interpretation that takes the head of the RC as the complement of the preposition *for* may even be the preferred reading, simply because a) mechanisms are not animate and we have already discussed the tendency of subjects to be animate and also b) mechanisms are things that need to be accounted for, so it is pragmatically plausible for the head to play a non-subject role. So, in light these considerations, we may suspect that the recovery of implicit material introduces additional processing demands. Again, while this may sound plausible, there are nevertheless good reasons to doubt the adequacy of the statement. First, the hypothesis that implicit material adds to the complexity of a patterns is at odds with what has been found in similar domains, e.g. VP ellipsis. Sentences exhibiting VP ellipsis have in fact been shown to be easier to process than their explicit counterparts (cf. Fodor, Bever, and Garrett 1974). Generally speaking, we may assume that material can be omitted if and only if the information it carries

is recoverable from contextual cues. Following the general impetus of this study, we may assume that non-finite clauses have made their way into the system, because they have proved to be effective tools for linguistic communication. That fact that they require an additional amount of inferencing by itself does not constitute evidence for extra processing demand. If anything, we may presume the opposite to be true: human inferencing capabilities are heavily exploited in linguistic communication because "inference is cheap, articulation expensive, and thus the design requirements are for a system that maximizes inference" (Levinson 2000: 29).

So, if it is not processing that can account for the observed distributions, what else can be held responsible? A possible answer to this question can be found in the area of discourse function and genre effect. If certain RC forms are associated with particular levels of formality, we should be able to disclose such association by looking at the text types that host the forms in question. In order to detect potential affiliations of grammatical choices and the text genre, we can refine this picture by again exploiting the text-type annotation of the ICE-GB.

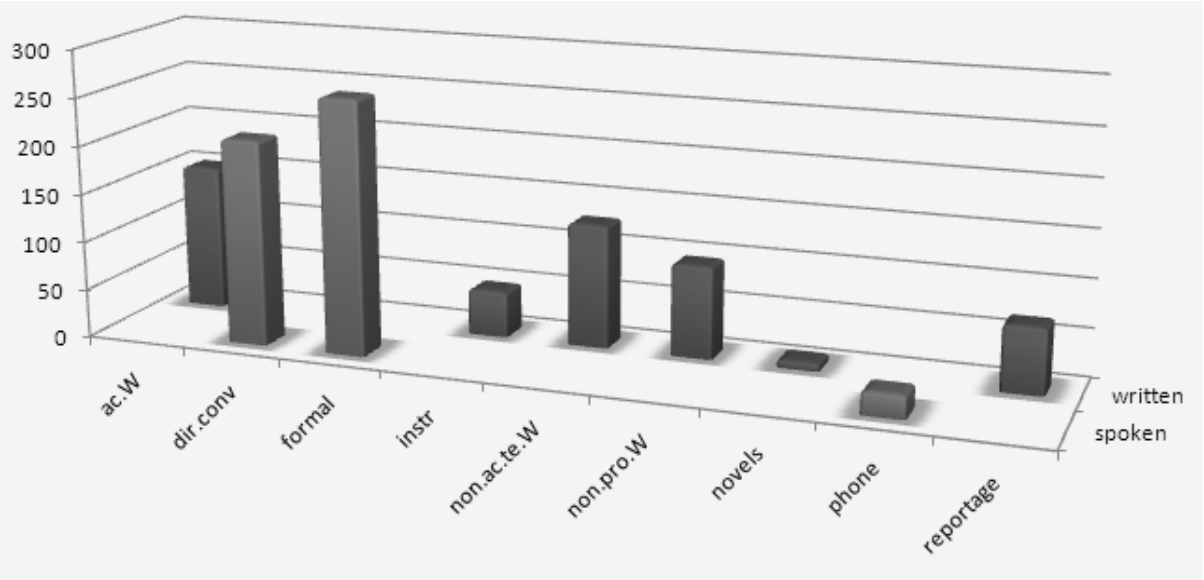

Figure 36: Text type composition for each modality

Figure 36 shows the genre composition of both the spoken and the written part. The spoken part consists of data from more formal contexts (formal), and rather informal contexts,

namely direct (casual) conversations (dir.conv) and telephone conversations (phone). The genres distinguished on the side of written language are academic text (ac.W), non-academic technical writing (non.ac.te.W), non-professional writing (non.pro.W), instructional text (instr), reportage text, and novels. As a reasonable approximation we may treat this order to reflect the degree of formality of the respective type, so that academic texts are most formal and novels are (at least potentially) most informal. Using this more detailed classification, we can better assess potential discourse specific constructional choices, i.e. genre- and formality-dependent affiliations of our RC variants. As the text types are register-specific, we will look at the spoken and written distributions separately. Figure 37 presents an association plot of the results for the written part.

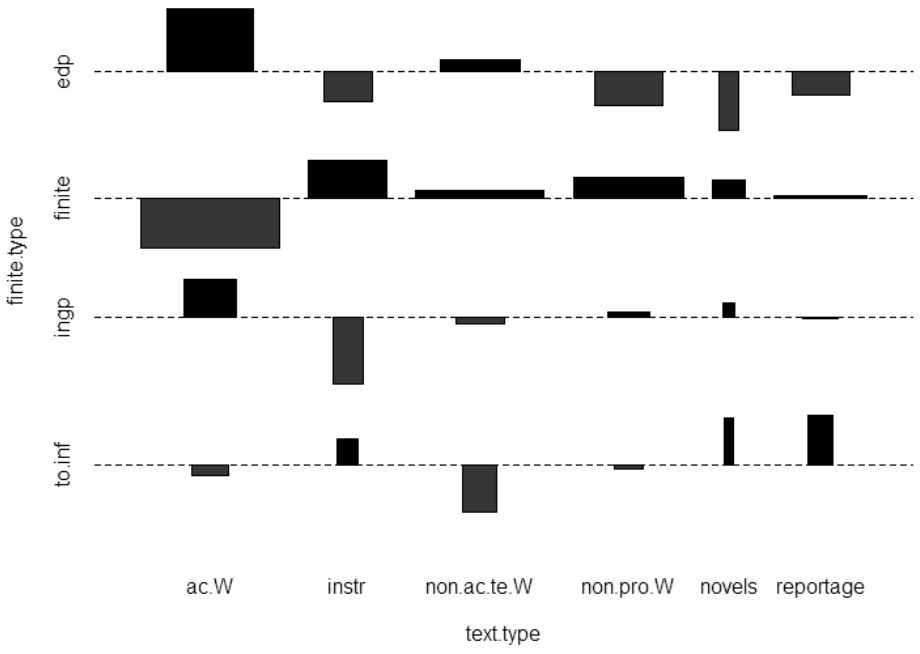

Figure 37: Assocplot RC-type (finiteness) versus genre [written language]

Recall that the dotted lines represent the expected values, bars above the dotted line indicate positive associations, and bars below that designate negative associations (repulsion). We can see that the *–ed* participle indeed prefers formal contexts over informal ones: it is associated most strongly with academic and non-academic technical writing and, correspondingly, occurs with a token frequency below H0 expectations for the remaining more informal text types. It is seems thus safe to say that the *–ed* participial construction is reserved for

informationally dense language. The *–ing* variant exhibits a similar but somewhat less pronounced profile. It too is associated with academic writing, but it is not a preferred choice of non-academic technical writing. Furthermore, we can see that its usage frequency is slightly above chance-level in non professional writing and in novels and slightly below that level in reportage texts. The last non-finite type, *to*-infinitival patterns, apparently is the least formal type, and can be found mainly in reportages, novels, and instructional language. Interestingly, we also learn from Figure 38 that finite RCs are underrepresented in academic writing. Less technical writings, however, exhibit increased amounts of usage of the finite variant.

The characteristics of the spoken patterns complement the observations made for the written part. Figure 38 presents the association plot:

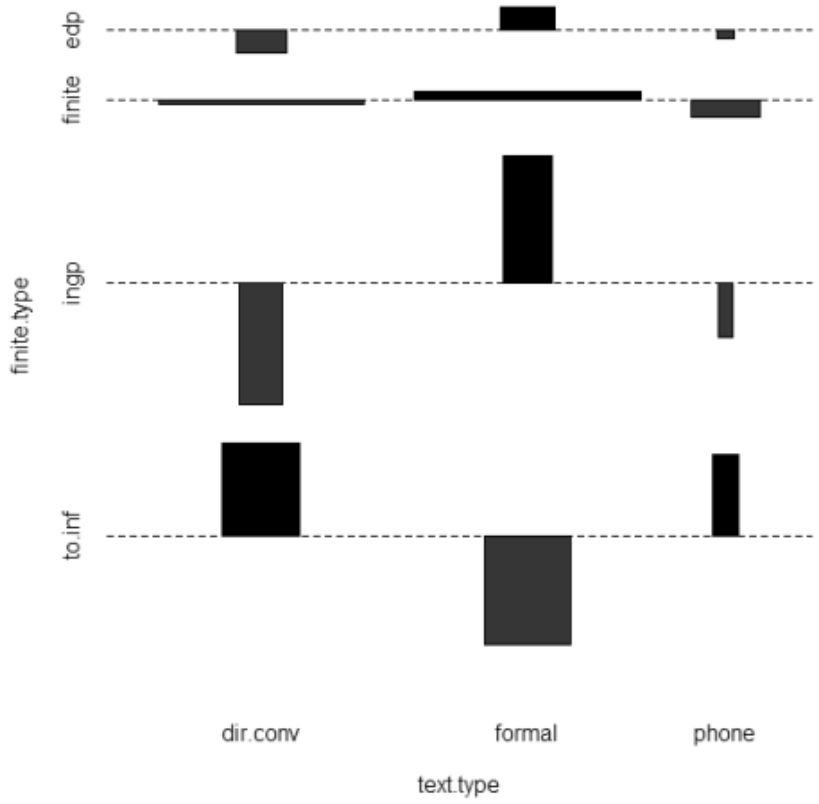

Figure 38 Assocplot RC type (finiteness) versus genre [spoken]

We find that the *–ed* participle construction is used in more formal contexts (and is dissociated from the two more informal genres, i.e. direct and telephone conversations). The

most frequent class of RC constructions, those involving finite RCs, exhibit a similar profile in tendency but are close to being neutral as far as their associated degree of formality is concerned. The clearest difference in genre-contingent usage can be observed for *–ing*-participles and *to*-infinitival constructions: whereas the former are preferred in formal contexts, the latter are typically used in informal settings. Similar to what we observed for the written modality, we find that to infinitival RCs are associated with less formal contexts. In fact, it is very frequent in direct conversations and telephone conversations. Apart from their lower degree of formality, these text types are also characterized by constant monitoring of the ongoing discourse. The communicating parties can evaluate the communicative success of the vehicles they chose to use and can immediately correct potential mishaps. This allows them to maximize inferencing capabilities and use less explicit forms. In more formal contexts a strong reliance on inferencing may be in some sense inadequate just because such utterances leave more room for interpretation, potentially more room than the speaker is willing to admit. The difference in the amount of required inferencing can thus be viewed as bridging the processing and the genre-dependence perspectives.

Another potential explanation for these findings is the discourse function served by the *to*-infinitival pattern. *To*-infinitival RCs are the preferred form to express "ad-hoc concepts", i.e. temporary constructs that arise for specific purposes at particular times (Barsalou 1987, 1992). The construction and expression of such context-sensitive ideas arguably is more important in spoken discourse. Examples are given in (50) and (51).

(50) **The best thing to do** is to turn straight to the index and look through the index […]
[S1A-053 :#013]

(51) And **my first author to come through** was Guy de Maupassant
[S1B-026 #078]

To-infinitival constructions provide an excellent means of expressing nominal concepts for which the language does not provide a lexical form (or—if such a lexeme exists—which the

speaker cannot access at the time of utterance). They work like ordinary definitions in that they first provide the superordinate category and then restrict the set of potential referents by way of a post-modifying element. Clausal post-modifiers, i.e. RCs, offer a much wider range of expressible predicates than alternative forms such as PPs, making RCs the more productive variant (in the sense that a greater number of types can e expressed using this form). The non-finite character of *to*-infinitival RCs is desirable as it results in a shorter overall form of the NP, making the form more cohesive and thus more unit-like (cf. Barsalou 1992 for further discussion). At this point I would like to foreshadow some results to be presented in later sections. We will observe in § 3.3.3 that the majority of spoken finite RCs are object relatives that this preference extends to *to*-infinitival RC as well (cf. § 4.1). For finite RCs this preference has been explained in terms of the function these types serve in the discourse (->grounding/anchoring). It is worthy to note that the preference for object *to*-infinitival RC (like the one in (50)) is somehow connected to this function. While both infinitival RC types can be used to express ad hoc concepts, object RCs involve another implicit argument that typically is a generic subject so that the non-finite clause could be paraphrased by a finite clause with either generic *you* or generic *one*.

(50) The best thing to do is to turn straight to the index and look through the index […] (S1A-053 :#013:1)

(51) The best thing you could do […]

(52) The best thing one could do […]

The underlying semantics of the implicit subject results make these *to*-infinitival RCs similar to finite object RCs, which as we shall see shortly also prefer pronominal subjects. This similarity may lead us to suspect that finite and *to*-infinitival object RC share some of the discourse functional potential of these types. While the discourse functional difference between subject and object *to*-infinitival RCs will not be pursued here, it may be a point of departure for future research.

A tentative conclusion that we may draw on the basis of the distributional analysis of finite and non-finite RC is that it appears to be a combination of processing related and discourse-functional factors that governs the usage frequency of the RC types under investigation. The heavy use of *to*-infinitival patterns in spoken language is licensed by the reliance on on-line inferencing capacities and is motivated by its discourse-functional properties, specifically its function to express ad hoc concepts. The stronger reliance on finite RCs in spoken language may be due to processing factors: finite clauses exhibit less local syntactic ambiguity and can thus be conceived of as being more explicit processing instructions. In contrast, non-finite RCs constitute good formal means to express additional predications as their reduced form may be viewed as an iconic reflection of their secondary predication status. Their drawbacks—local syntactic ambiguity and also less time to process the RC-situation—do not prevent their employment in written language in virtue of their external representation, which allows readers to backtrack and reanalyze more complicated structures. Consistent with this line of thinking we observe that if participial forms are found in spoken discourse at all, they almost always occur in formal contexts, in which processing considerations are likely to be countered by rhetorical considerations. This is to say that in order to achieve certain perlocutionary effects—e.g. impress their audience—speakers are likely to use more demanding forms that in more casual contexts would be avoided or dispreferred for processing reasons.

As the true discourse functional potential of a construction type and also its overall processing demand is probably a function of a quite large set of variable and their interactions, we must of course be careful with our generalizations. This caveat leads us naturally into "multivariate waters" and the configurational view on English relative clause constructions, which we will assume in Chapter 4.

3.3.2 Grammatical features of RC: Transitivity

The second variable in this group concerns the transitivity of the RC. As indicated earlier, the value for transitivity is taken here as indirect expression of the complexity of the described situation. Following the treatment in the ICE-GB corpus, five levels of transitivity were distinguished. Difference in transitivity values reflects differences in valency (arity) and

semantic contentfulness of the respective predicates. We may assume the following relative ordering to represent degrees of associated processing demand: copular << intransitive << monotransitive << complex transitive << ditransitive. The primary criterion for the ranking was the number of required arguments. In case two factor levels require the same number of arguments, the ranking was compiled on the basis of the semantic content of the predicate. For instance, copular and intransitive constructions both require only a single argument, i.e. a (logical) subject. However, the predicates expressed by intransitive verbs are usually richer in their semantics than those of copular verbs, which are very close to being semantically empty. Similarly, monotransitive and complex transitive patterns both require two arguments, but complex transitives are more demanding (on average) as they ascribe an additional property to the second argument. Finally ditransitive constructions express ternary predicates and were hence considered to describe the most complex situations. Figure 39 presents the distributions of these types for the relative clause.

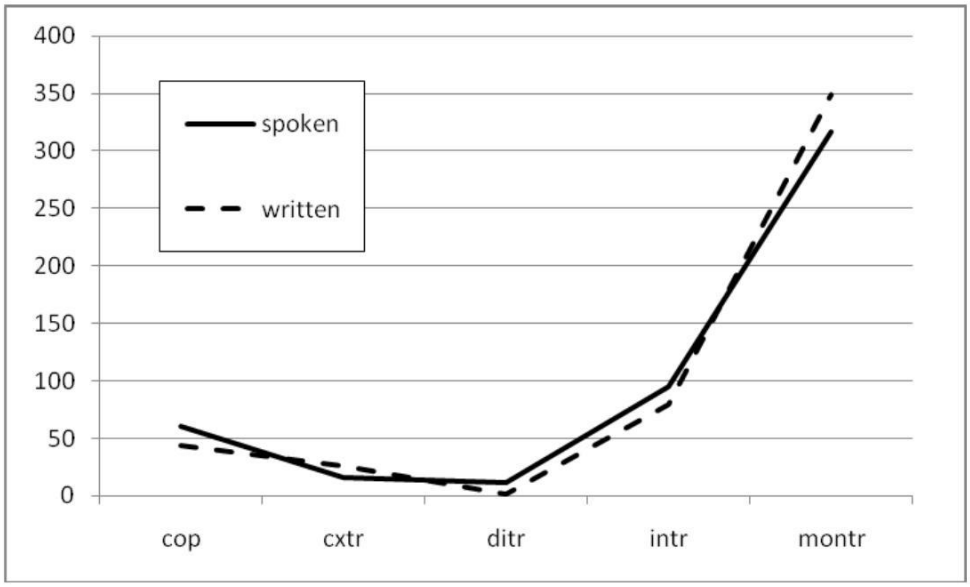

Figure 39: Transitivity of relative clause (n=1000)

Using our by now familiar plot triple we can easily evaluate the modality specific differences (cf. Figure 40):

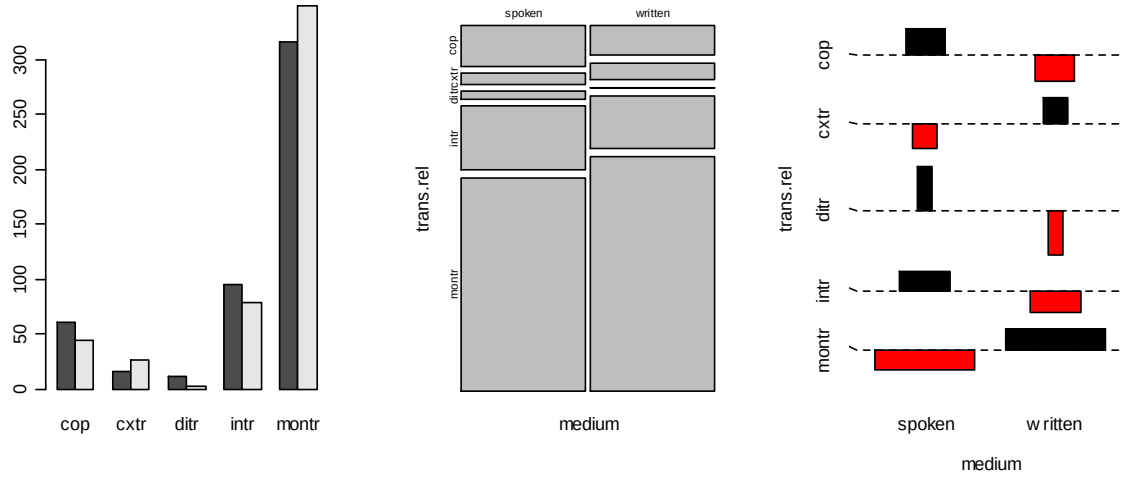

Figure 40: Transitivity of relative clause across modalities

We observe that the transitivity distributions are quite similar across modalities. None of the differences is statistically significant. Monotransitive RCs (montr) are by far the most frequent pattern in both registers, followed by intransitives (intr) and copular (cop) constructions. Complex transitive (cxtr) and ditransitive (ditr) play a subsidiary role.

Since we could not detect any modality-contingent distributional differences, there is nothing in urgent need of explanation. We may note, however, that the typical level of complexity of the situation described by the RC is of intermediate and specifies a relation between two participants. The prevalence of monotransitive patterns is not surprising if one assumes a cognitive linguistic point of view. Like all linguistic categories, the concept of a clause is seen as grounded in basic human experience. Langacker calls the concept underlying what is arguably the most typical kind of event the *canonical event model* (Langacker 2008: 354ff.). An instance of this model is identified as a bounded, forceful event in which an agent acts on a patient to induce a state and monotransitive clauses are the linguistic means to of coding such events. We will return to the issue in § 3.4.1 and 3.5.1 when we have a look at the transitivity of the main clause and the typical constellations of RC and MC transitivity. At this point however, we may proceed to our next variable of interest, which concern the type of relative clause.

3.3.3 Grammatical features of RC: Relativized role

We turn next to the factor whose values provide the labels for relative clauses in the linguistic literature, i.e. the internal syntax of the RC. Let us first have a look at the distribution of grammatical roles that underwent relativization and again see if there are notable differences across modalities. Figure 41 presents the frequencies of the RC types across modalities.

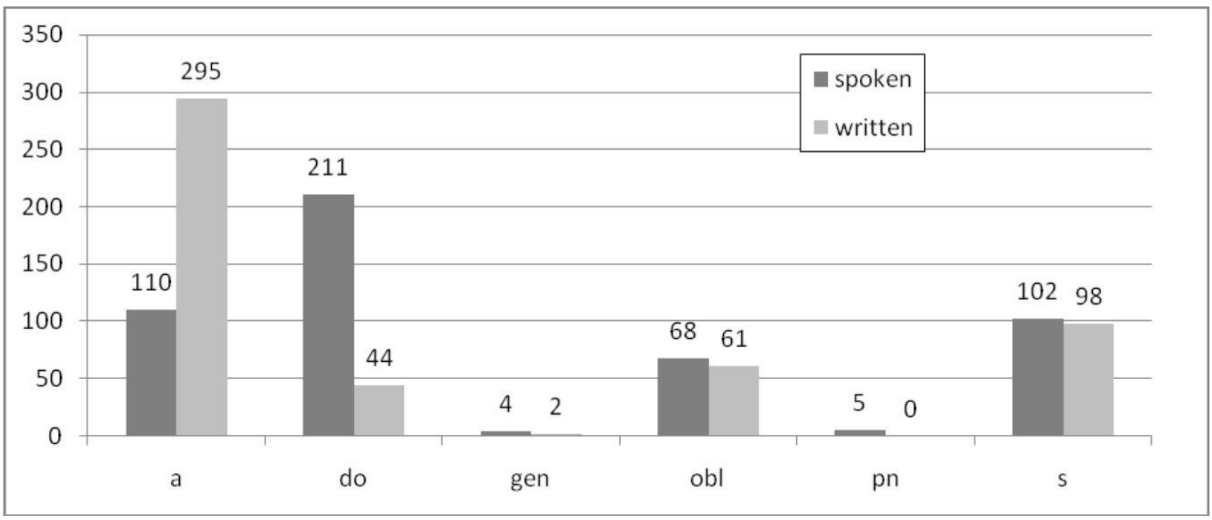

Figure 41: Relativized roles across modalities

We observe that there are in fact striking differences across modalities. While written language is dominated by subject relatives, we observe a fairly even split between subject and object relatives in spoken language. The category "obl" occupies third place and is evenly distributed across modalities. In the coding used here, it includes both obliques proper but also various types of adjuncts. In collapsing these subtypes we are focusing here on their commonalities (the respective nominals are in most cases sisters of a preposition). The object-like character of an oblique was considered to be of only secondary importance here. The remaining types, predicate nominal (pn) and genitive (gen) play only marginal roles in both registers. Submitting these data to statistical analysis, we learn that there are indeed quite strong associations between a) subject relatives and written language and b) object relatives and spoken language ($\chi 2 = 168.35$, df = 3, $p < 2.22e\text{-}16$, Cramer's V = 0.41).[16]

[16] The value 2.22e-16 is a constant expressing the smallest positive Lisp float that can be added to 1.0 to

Figure 42 presents the results graphically:

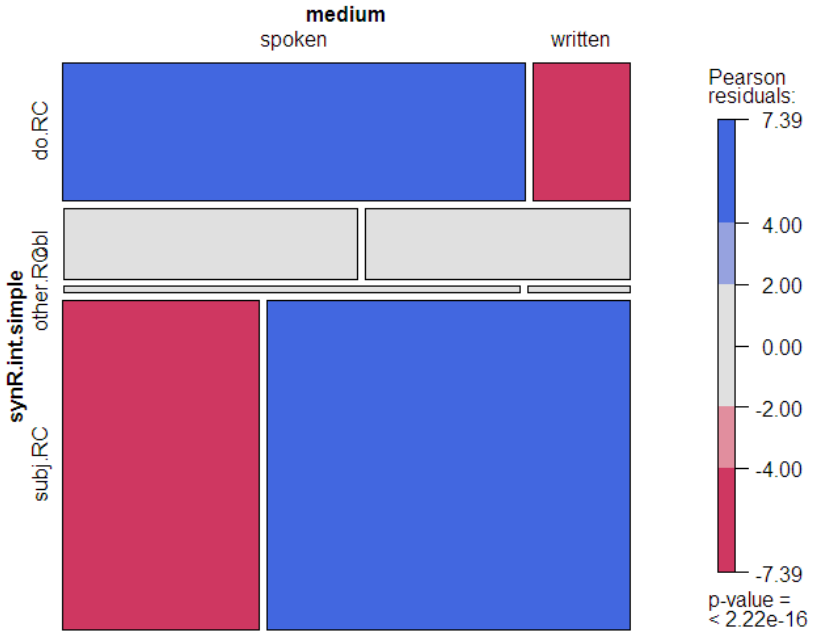

Figure 42: Relativized role (simple) versus medium

Of course, this picture may be misleading as it is possible for the variable finiteness to confound the results: we have seen in § 3.3.1 that there is a greater number of *–ed* participial RCs in written language and these type imply that the head plays a subject role within the RC. However, if we reduce the data set and focus on only those RCs that do not require subject roles—that is if we focus on finite and *to*-infinitival RCs—we still find very similar results ($\chi 2 = 108.66$, df = 3, $p < 2.2e\text{-}16$, Cramer's V = 0.358).

In light of these distributions and associations, the alleged increased processing demand of object relatives (as compared to subject relatives), which is maybe one of the most pertinacious beliefs in recent psycholinguistics (Ford 1983, Frauenfelder et al. 1980, Holmes and O'Regan 1981, King and Just 1991, King and Kutas 1995, inter alia present evidence for this processing asymmetry), is not predicted by a exemplar-based view on sentences processing as it is assumed here. However, this belief is beginning to dissolve under the

produce a distinct value for IEEE machines. Adding a smaller number to 1.0 will yield 1.0 again due to round-off. Lower p-values are thus not meaningful on contemporary home computers.

weight of more recent findings, which have shown other factors such as animacy (Mak et al. 2002, 2006) or NP type configuration (Bever 1974, Gibson 1998, Gordon et al. 2001, 2004) to modulate the processing difficulty of a RC. We will turn to these issues in § 3.5.4. The fact that the subject-object relative clause processing asymmetry is not predicted from the results found here is therefore favorable for the general account presented here.

3.3.4 Grammatical features of RC: Corpus comparison

This final section in our discussion of grammatical features of RCs is somewhat different in its approach. More precisely, we will return to some variables we have already discussed, e.g. finiteness, and combine these with other features we have not yet discussed, e.g. voice. This more synthetic approach to the description of RCs is in some sense a first precursor of the configurational view we will assume in later chapters, albeit on a smaller scale. The reason why this step was taken here is mainly to allow for a more direct comparison of the present data and results with recent corpus work on relative clauses.

Douglas Roland and colleagues have provided a comprehensive analysis of the distributional frequencies of a number of structures that have played a prominent role in psycholinguistic modeling of language comprehension (Roland et al. 2007). Among these structures are a variety of relative clause patterns, which are presented in Table 7:

Table 7: relative clause type investigated in Roland et al. 2007

Structure	Example (from Wall Street Journal corpus)
Subject relative	The researchers who studied the workers
Object relatives (full)	The 25 countries that she placed under varying degress of scrutiny
Object relatives (reduced)	The foreign stocks they hold
Passive relatives (full)	Delmont D. Davis, who was named president and chief operating officer in August
Passive relatives (reduced)	A competing psyllium-fortified cereal called Heartwise
Infinitive subject relative	A computer system to map its waterworks
Inifinitive object relative	A blind girl to cure
Infinitive passive relative	The last woman to be executed in France

The classification is different from the one adhered to in the present study and may thus require some familiarization. To get a quick hold of the classification, it is helpful to think of the contrasted sub-types in Table 7 as factor level combinations of the variables internal role

(subject versus object), finiteness (finite versus infinitive versus participial), voice (active versus passive), and relativizer (present versus absent). The table then presents the configurations that are permissible in English. While some types have received more attention than they have in the present study, e.g. *infinite passive relatives*, other types were neglected altogether in the Roland study (e.g. -*ing* participle RC and relatives that are neither subject nor object relatives). It is only fair to say at this point that the exclusion of *–ing* participial constructions is not uncommon in the grammatical treatment of relative clauses. In fact, Quirk and colleagues treat these types as postmodifying non-finite clauses but distinguish them from relative clauses proper (Quirk et al 1985: Chapter 17). The reason for the exclusion of these types is the indeterminacy of their function. Consider the examples in (53) and (54) for an illustration:

(53) The man [wearing such dark glasses] obviously could not see clearly.

(54) [Wearing such dark glasses] the man obviously could not see clearly.

The bracketed clause in (53) meets all the requirements of a (non restrictive) non-finite *–ing* participial RC. However, as shown in (54), it is possible to extrapose the clause without changing its meaning in any obvious way. But RCs in English always follow their heads. The sentence initial position, however, is quite usual for certain types of adverbial clauses. So *–ing* participial RC—if we wish to postulate this category—behave somewhat differently from archetypical (finite) RCs and this may motivate a different grammatical treatment. However, by the same logic we should also exclude *–ed* participial RCs. Quirk and colleagues do so and are thus at least consistent in their treatment. Consider the examples in (55) and (56).

(55) The substance [discovered almost by accident] has revolutionized medicine.

(56) [Discovered almost by accident] the substance has revolutionized medicine.

The last remaining non-finite RC type, the *to*-infinitival RC sub-type, must also be considered less typical for the category relative clause as it is often hard to distinguish from it from adverbial clauses expressing a purposive relationship. Schmidtke (to appear) presents a thorough discussion of this issue incorporating perspectives from linguistic typology and grammaticalization theory. The examples in (57) and (58) were taken from this work and illustrate the fuzzy boundary between relative and purpose clauses in English:

(57) Two other books [to read on holiday] were lent to me by Tina.
[RC-like postmodifier]

(58) Tina lent me them [to read on holiday].
[purposive adjunct]

Whatever the reason for the exclusion of *–ing* participial RC (but inclusion of the other non-finite types) might have been, the majority of sub-types are considered in both studies so that we may compare Roland's results with those of the present study. Before we do so, however, a few words on the investigated corpora are in order.

Roland and colleagues investigated the distributions of the types in Table 7 across a respectable set of corpora: the Brown corpus, the British National Corpus (BNC), the spoken part of that corpus (BNC spoken), the Switchboard corpus, and the Wall Street Journal Corpus. The Brown corpus is the result of Henry Kucera and W. Nelson Francis' pioneering work at Brown University in the nineteen-sixties. It comprises roughly a million words that distribute across 500 samples from 15 genres (Kucera and Francis 1967). Compiled in the mid-nineties, the BNC comprises roughly 100 million words, of which 90% are from the written register (Burnard 2007). While both Brown and BNC can be considered fairly balanced, i.e. general, corpora, the remaining two corpora are more specialized: the Switchboard corpus consists of 2,400 telephone conversation between unacquainted adults recorded in the nineties, which in its transcribed form amounts to roughly 3 million words (Godfrey et al 1992). Finally, the Wall Street corpus is a parsed corpus consisting of roughly

1 million words of written texts, which mainly revolve around business issues. The distributions of the various RC types reported in Roland et al. (2007) are presented here as Figure 43.

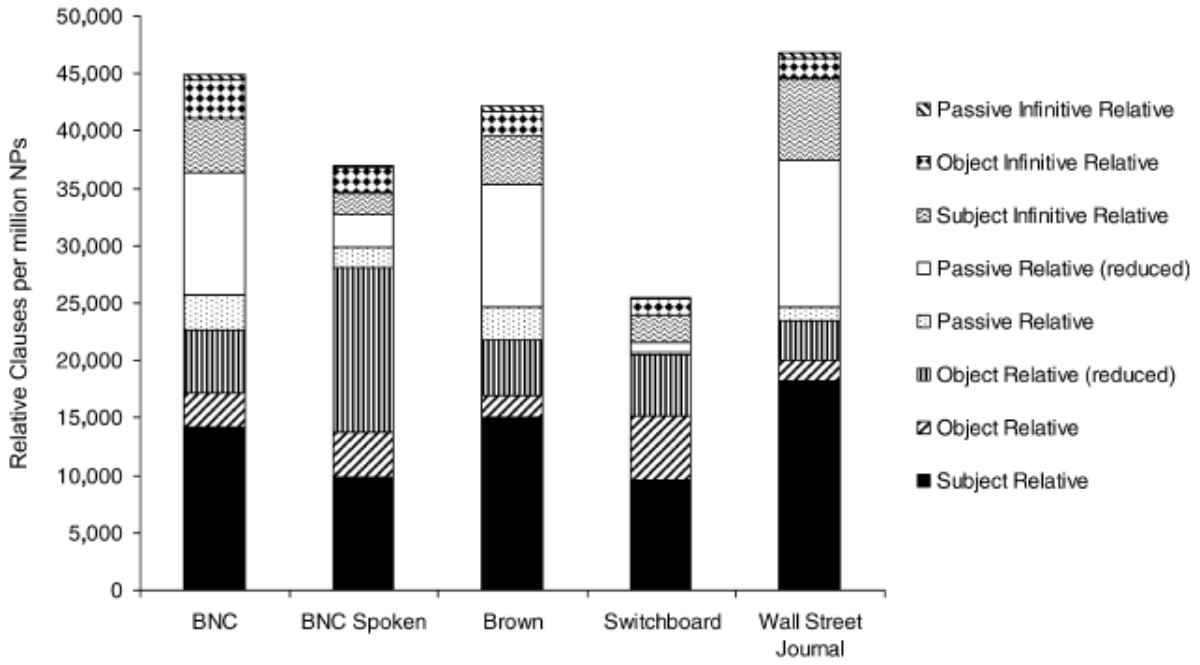

Figure 43: Distribution of relative clauses across corpora per million NPs (taken from Roland et al. 2007)

Figure 43 shows the frequencies of each RC type per 1 million noun phrases and entails that

1. subject relatives are more common in written corpora than in spoken corpora
2. passive relatives and passive infinitives are more common in written corpora
3. object relatives are more common in spoken corpora

Roland and colleagues also considered the proportions of full and reduced object relative clauses across corpora and report somewhat inconclusive results. The percentages of reduced object relatives are presented as Table 8:

Table 8: Proportion of reduced object relatives across corpora (Roland et al. 2007)

Corpus	Reduction
BNC spoken	79%
Brown	71%
BNC	65%
Wall Street Journal	65%
Switchboard	49%

Table 8 leads us to assume that register has no role to play in the explanation of object relative reduction. The spoken corpora occupy both first and last position in the ranking and the overall differences are not exactly very pronounced (with the possible exception of the low score obtained for the Switchboard corpus). However, Roland and colleagues note that these results have to be taken with a certain amount of caution as they have noticed a systematic error in the BNC tagging. In addition to this source of error, the BNC spoken data consist to a considerable degree of speech that is not exactly representative of that register such as speeches, lectures, and news broadcast. In contrast, the grammatical analyses of present ICE data were checked manually and it was ensured that the spoken texts are in fact representative for the registers at hand (cf. § 3.1).

In order to directly compare these results with the results obtained in the highly controlled data set used in the present study, the relevant data were synthesized so as to give rise to the same categories used in the Roland study. Figure 44 shows a stacked bar plot like the one presented as Figure 43 above and presents the raw frequency distributions of RC types of the complete data set (n = 1000).

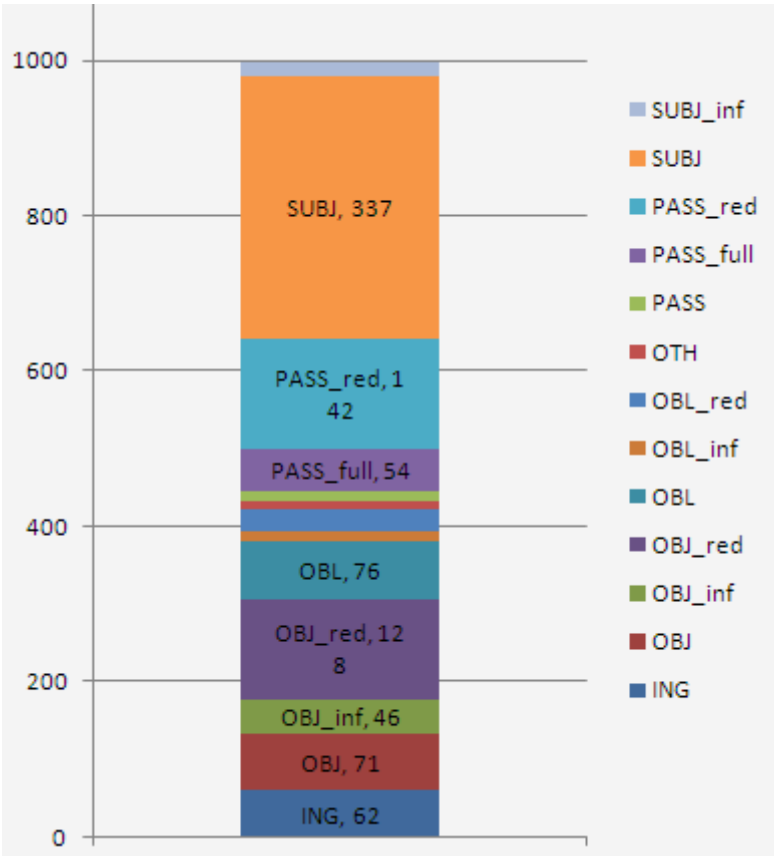

Figure 43: Type distribution (Roland classification)

Abstracting away from the modality specific contrasts, we observe that subject relatives constitute the largest group (~ 34 %), followed by reduced passives (~ 14%), and reduced object relatives (~ 13 %). Full (non-reduced) object relatives and oblique relative contribute roughly 7 % of the cases each, *-ing* participles make up 6% of the data and full passive are the last category to pass the 5% threshold. If we "zoom in" again and look at the modality specific distributions, we arrive at the distibutions presented as Figure 45:

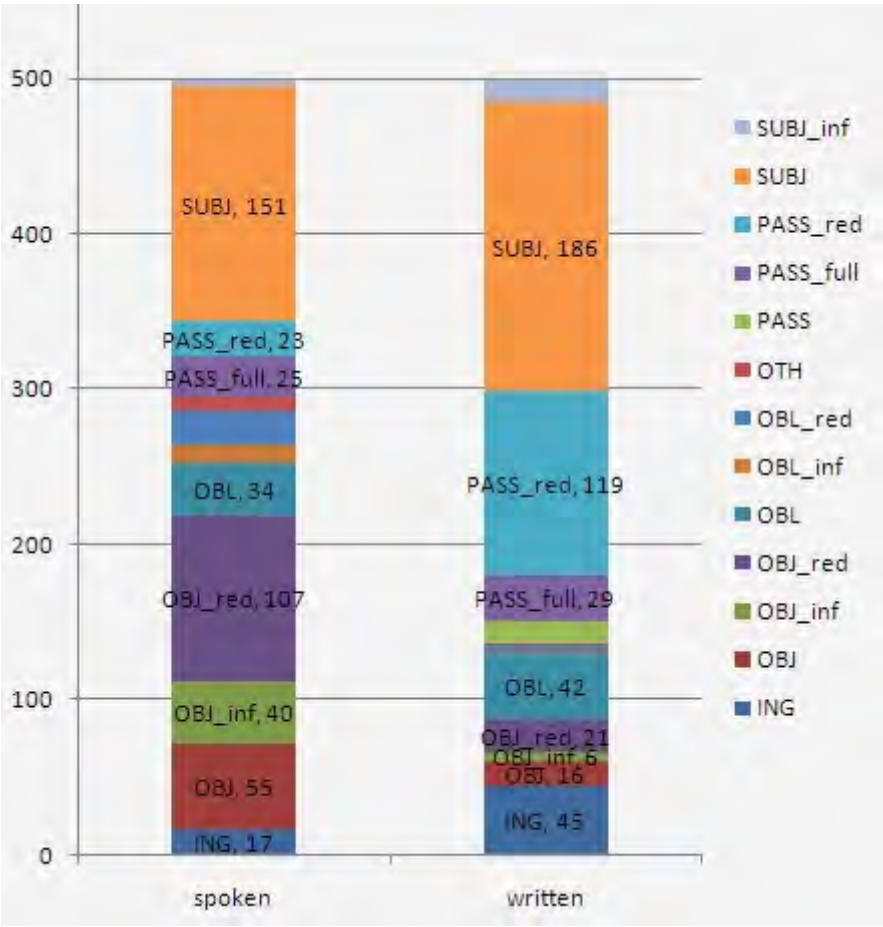

Figure 45: Distribution of RC types (Roland) across modalities

The distributions in Figure 45 are fully compatible with the main findings of the Roland study. Subject relative and (reduced) passives are more common in written language are more frequent in written discourse, while there are object relatives in the spoken register. We can help ourselves to a clearer overview, when we eliminate the marginal categories to focus on the quantitatively most prominent types. Figure 46 presents the corresponding overview, which presents the data from the seven most frequent types (n=884):

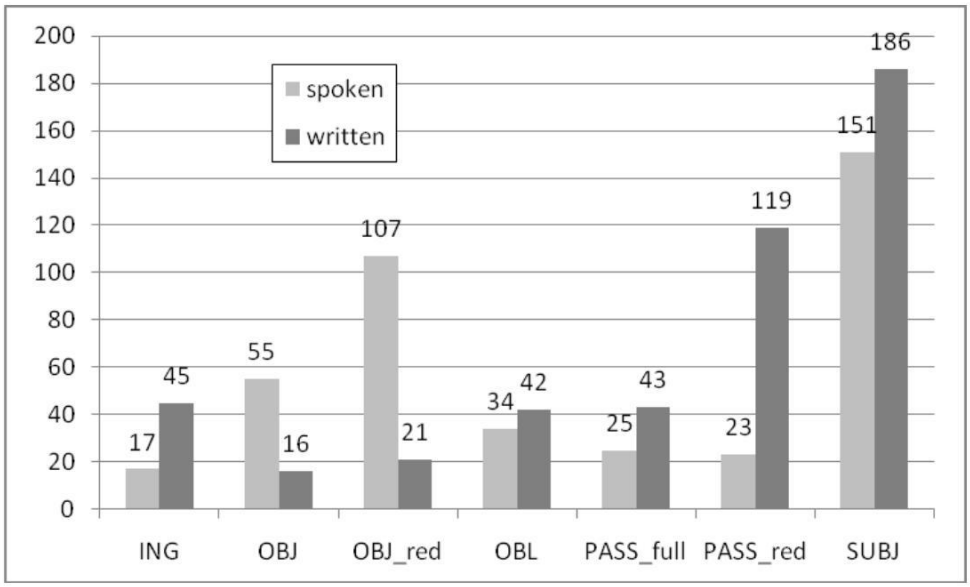

Figure 46: Distribution top seven RC types across modalities

Notice that while the clear tendencies of the Roland study are reflected in the present data, the manually checked data used here exhibit a more pronounced distributional difference between the modalities within the category of reduced object relative (= *that*-less relatives). According to the present data, *that* omission is much more common in spoken discourse (with a ratio of roughly 5:1). The statistical meaningfulness of these distributional differences is confirmed by the associative statistics that underlie the association plot in Figure 47.

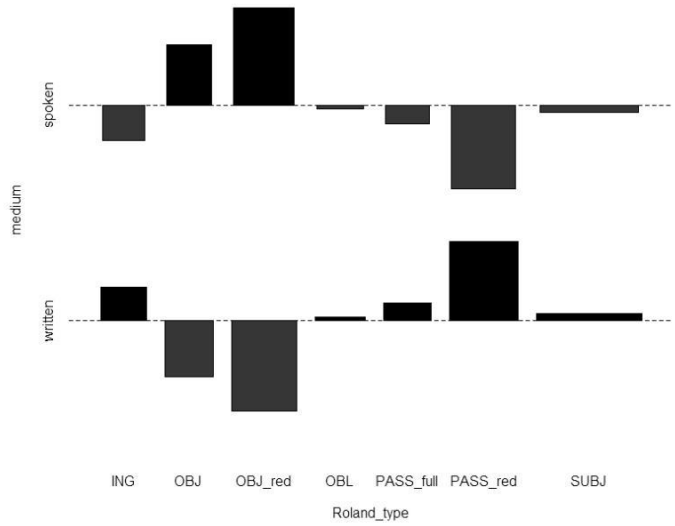

Figure 47: Association plot of RC types (Roland)

As we will discuss in more detail in § 4.3.3, this finding is comforting, if we wish to assume that the omission of non-obligatory relativizers are largely (though not exclusively) governed by processing factors.

The comparison of the present data with the results obtained in the large-scale study conducted by Roland and colleagues provided us with some additional confidence in asserting that the results presented here are not just idiosyncratic properties of the data used but rather that the results are quite stable across different corpora. This stability of the distributional results is especially comforting because a) the corpora differ greatly in sample size (1 million to 100 million words), b) the corpora were compiled from language from at least two generations of speakers (1960ies – 1990ies), and c) the corpora represent different varieties of English (British and American English).

This concludes our overview of the grammatical properties of the RC proper. We will noun turn to factors that influence the processing of RCC and that are encoded on the dominating main clause.

3.4 Features of the main clause

As dependent clause relative clauses are rarely found in isolation and require a dominating main clause. In this section we will have a look at some grammatical properties of the MCs in an RCC. We will start this discussion with an assessment of typical transitivity values of these structures and compare these with the preferred RC transitivity discussed in § 3.3.2. It will be argued that we can expect to find differences in typical patterning based on the different functions these clauses fulfill in the (English) language. We will then turn to issues concerning the positioning of the RC in the dominating clause. While the most salient contrast in this domain certainly is that between center and right embedded RCs, we will also look at the external syntactic properties at a more detailed level and ask what syntactic role in the MC typically receives RC modification.

3.4.1 Grammatical features of MC: Transitivity

As we have already motivated the reason for the inclusion of transitivity as a quantity of interest in § 3.3.2, which focused on the transitivity of RC, we may at this point dive into empirical waters right away. As all MCs are annotated with transitivity information in the ICE-GB, we can investigate the complete data set (n=1000). An overview of the frequencies of the types distinguished in the ICE-annotation scheme is given in Figure 48:

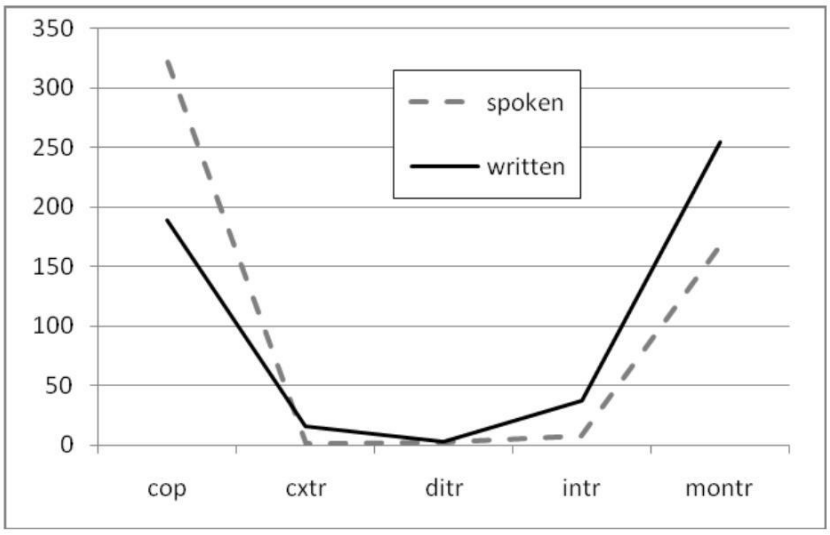

Figure 48: Transitivity of MC (overview)

We observe that for main clauses, copular and monotransitive constructions together account for the lion's share of the data. These two constructions type together basically account for most token in both registers differences. The remaining levels do only play marginal roles. A comparison of the two registers reveals that while the dominant type in spoken language is the copular construction, written language is dominated by monotransitive patterns. Submitting these data to statistical analysis, we learn that globally the distribution significally departs from what would be expected under the assumption of statistical independence of the factors transitivity and medium ($\chi^2 = 89.020$, df = 4, $p < 2.22e\text{-}16$, Cramer's V = 0.292). The local departure from independence is shown in the association plot in Figure 49.

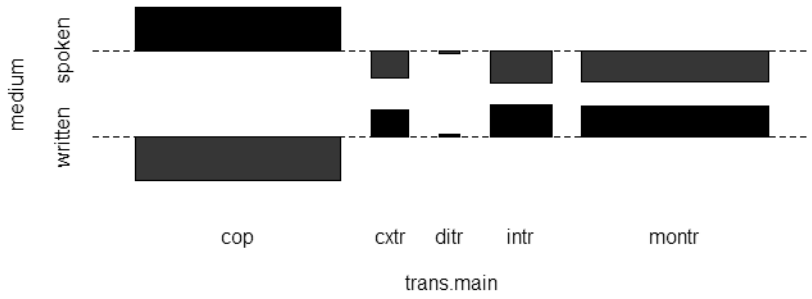

Figure 49: Transitivity of the main clause across modalities (assocplot)

As indicated by the width of the bars, we observe that the residuals from the categories cop (=copular constructions) and montr (=monotransitive constructions). This is not particularly surprising as 933 out 100 cases are instantiations of either one of the two dominant patterns.

The heavier use of monotransitives in written language is consistent with the hypothesis that more demanding patterns are to be found with the written modality. But while this hypothesis is consistent with the observed data, it does not seem to be an adequate explanation of the distributional facts. Especially as we noted earlier that monotransitive clauses constitute the linguistic means for expressing the most basic types of situation (cf. the discussion of the *canonical event model* in § 3.3.2). Simply put, monotransitive are certainly not complex enough so as to be avoided in spoken discourse. A more plausible explanation for the frequency asymmetries may be located at the level of discourse function. We will discuss the discourse-functions of various RCC types in later sections, when our descriptions have reached higher degrees of specificity (-> Chapter 4). At this point suffice it to say that a constructionist view on grammar would assign different functions to different forms. It is trivial to note that forms of RCC with monotransitive MC differ from those with copular MC. However, this general statement is sufficient to postulate different functions and hence motivates the idea that there are different communicative needs to be satisfied across the two modalities.

Another noteworthy contrast concerns the transitivity of the MC and the RC. If we directly compared the clause transitivity of the MC with that of the RC (cf. § 3.3.2), we

observe an interesting difference. Figure 50 presents a direct comparison:

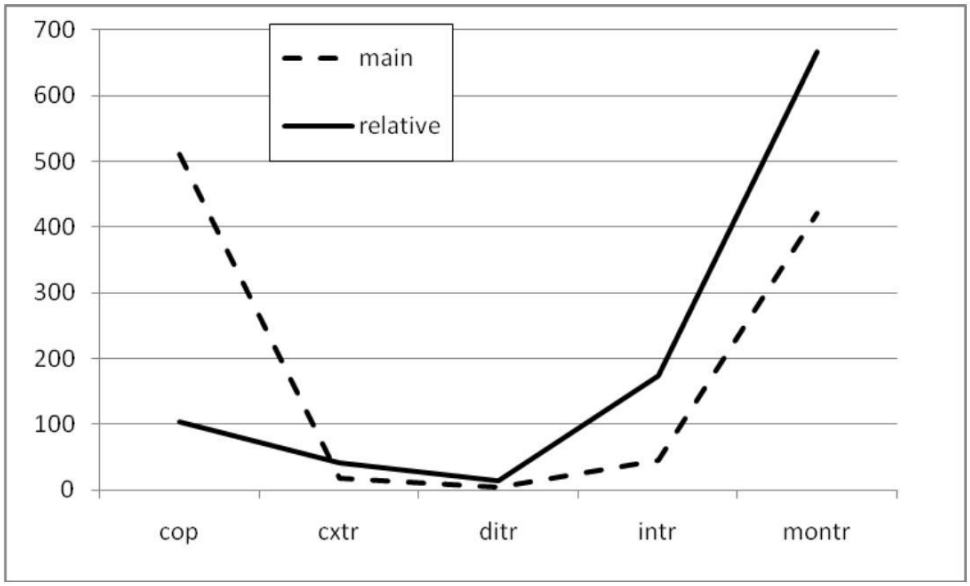

Figure 50: Transitivity of MC and RC

Figure 50 reveals that the transitivity profiles are quite similar across clause types with one exception: the ratio of copular clauses. While more than 50% of the MCs in RCCs are copular, this is only the case for a mere 10% of RCs. If we treat clause transitivity as an indirect measure of clause complexity (cf. § 3.3.2), we would be forced to assume that the RC constituent of an RCC on average tends to be more complex than the MC constituent. This *relative propositional weight* of the clausal constituents has been described in a manner fitting such a statement by various researchers (Lambrecht 1988, Bates and Devescovi 1989, Diessel 2004 inter alia).

While the relative low propositional content of the MC predication relative to the MC predication in RCC certainly is an important finding, it appears that the low proportion of copular RC is predictable from the functional role of RCs in the linguistic system of English. Let me elaborate a little on this point. Languages like English employ numerous ways to refer to some entity in the real world. A speaker can choose from a set of types of referring expressions, with different semantic properties. One way of referring to an entity is using a definite description, which applies to a unique individual. Such definite descriptions can assume various forms in English, two of which are exemplified in (58) and (59).

(58) [Det AdjP N] as in *The oldest building in Jena…*

(59) [Det N RC] as in *The building which is the oldest one in Jena…*

Both expressions can be used refer to some building for which it is true that it is the oldest one in Jena. The first example uses a pre-modifying AdjP for the ascription of this property, while the second one employs a post-positioned clausal modifier. So, in principle speakers have a choice as to what form they wish to employ to express the thought they desire to communicate. These forms, however, obviously have different intrinsic properties; (58) is for instance longer than (59) and hence requires more effort than its alternative. At this point, we should connect our discussion with the contents of § 1.2.: If we assume that speakers tend to use forms that are maximally efficient, we would therefore predict that they prefer (58) over (59). Both the AdjP and the RC can be used to express monovalent (unary) predicates. But it is hard to express more complex predicates of greater valency (or arity) by way of a simple AdjP. In contrast, RCs—qua being clauses—present no such limits. They can of course employ verbs expressing predicates/relations of arbitrary valency. Now, given the available linguistic means in English and their properties, using a RC to express a unary predicate simply is a waste of energy. The low frequency of copular RCs can thus be explained by a general 'Minimize Form' principle. A recent formulation of such a principle can be found in Hawkins (2004). Hawkins formulates it as follows:

Minimize Forms

"*The human processor prefers to minimize the formal complexity of each linguistic form (its phoneme, morpheme, word, phrasal units) and the number of forms with unique conventionalized property assignments, thereby assigning more properties to fewer forms. These minimizations apply in proportion to the ease with which a given property P can be assigned in processing to a given F.*" (Hawkins 2004:38)

Hawkins' principle is of course not new to linguistic theorizing and he acknowledges the obvious connections to work on economy (e.g. Haiman 1983, 1985) and inferential pragmatics (e.g. Grice 1975, Levinson 2000). What underlies all these strands of research is a commitment to the idea that language users behave as if they were trying to produce the *minimal information sufficient to achieve their communicative ends*. In fact, this principle goes far beyond linguistic behavior and has counterparts in decision theory and game theory, as well as in areas of statistics and philosophy (cf. Casti 1996 for an approachable introduction into the mathematics of these issues and the underlying principle, the *minimax theorem*).

We will return to the issue of transitivity for a third (and last) time in § 3.5.1, when we look at the distributions of RCC transitivity configurations. In this section we will ask which sequences of transitivity values are common and which sequences tend to be avoided.

3.4.2 Grammatical features of MC: External role and type of embedding

The head of the RC not only plays a role in the RC proper but is also the central element in the dominating NP that has a grammatical function in the main clause. This is often referred to as the external syntax of an RC. The grammatical function of the element that gets modified by way of a relative clause stands in an intricate relationship to the overall structure of the clause. Usually the first constituent in an English declarative sentence is the subject of that clause so that modification of the subject results gives rise to *center embedding* (CE) or *nesting* of the RC. Modifying VP internal arguments in English results in a right branching structure, i.e. *right embedding* (RE). We may use the following examples as the basis for our discussion:

(60) The juice [(that|which|∅) the child spilled] stained the rug. [CE]

(61) The child spilled the juice [(that|which|∅) stained the rug]. [RE]

The impact of the type of embedding on the overall processing difficulty of a linguistic

construction has intrigued researchers from the earliest days of psycholinguistic inquiry. While some studies have reported that CE structures with just a single embedding to be on par with (or even easier than) their RE counterparts (Holmes 1973, Baird and Koslick 1974, Gibson et al. 2005), the vast majority of the experimental work on this phenomenon reports that center embedded structures are harder to process than right embedded ones (e.g. Miller & Isard 1964, Blumenthal 1967, Blumenthal and Boakes 1967, Fodor & Garrett 1967, Marks 1968, Schlesinger 1968, Foss & Cairns 1970, Blaubergs & Braine 1974, Larkin & Burns 1977, King and Just 1991). And there certainly is no shortage of theoretical accounts of these results. The most influential explanations include the following:

i. Center embedding results in deviance from canonical word order
ii. Center embedded structures make it more difficult to match the MC subject with its predicate.
iii. Center embedded structures involve a greater distance between subjects and verbs (=longer dependency domains).
iv. Language users exhibit a tendency to treat nesting as coordination
v. Center embedding gives rise to local syntactic ambiguity with some verbs (see or like) but not others (hit or slap)

Despite the problems posed by CE structures, their comprehensibility may improve in the presence of semantic cues (Bever 1970, King and Just 1991, Schlesinger 1968, Stolz 1967). However, center embedded structures become incomprehensible even when such semantic cues are available, if the number of embeddings is greater than three (Gibson & Thomas 1999).

3.4.2.1 Interlude: some words on working memory, connectionism, constraint-satisfaction, and analogy

At this point in our discussion, I would like to pause for a minute and have a closer look at what I consider to be the "backbone of the argument" underlying all the accounts of CE/RE processing differences, namely the role of memory limitations in the processing of linguistic structures. We will see that the role of memory limitations depends a lot on assumptions about the nature of the mechanism and the operations that we assume are at work in language processing and the types of model we make use of in order to understand these mechanisms. In light of this dependency, it appears sensible to provide a discussion of the general workings of the mechanisms that fit the theoretical treatment advocated here. This section will thus outline the relationship of working memory and the class of models of sentence processing that the present account is most sympathetic to.

As indicated by the very formulation of the proposed reasons of the increased processing demand associated with center embedding, most accounts assume that the processing difficulties observed in the experimental settings originate from elevated demands on (verbal) working memory. That is to say that most accounts assume that there is a limited resource and that cognitive processes, e.g. comprehending linguistic structures, consume certain amounts of this limited resource. This idea is so intuitively plausible that it is actually hard to even conceive of an alternative view. It certainly was the dominant view in the second half of the 20th century, which is of course closely connected to the theory of computing and the advent of the digital computer. Ever since the late fifties the computer has provided the dominant metaphor for the conceptualization of the human mind at least in western culture. This conception of the mind, and by implication language processing, is beginning to change though. There are many angles from which the computational view of the mind can—and in fact has been—criticized. This however is not the place to even sketch this complex and difficult debate (cf., e.g., Van Gelder 1995 for a discussion of the computational view and its alternatives). Many key ideas of such a criticism can be traced back to Aristotle and his ideas about mental associations, which have ascended and descended again and again in the history of western thought (cf. Sutton 1998). However, the arrival of parallel distributed processing

(PDP) models and connectionism must be considered an important step in the development of alternative conceptions of the mind (cf. Rumelhart and McClelland 1986, Elman et al. 1996, inter alia, Selfridge et al. 1988 presents an annotated bibliography including many influential papers on the issue). Much thinking has been devoted to what exactly connectionism could believably be, i.e. how exactly it contrasts with symbolic computational modeling (cf. Fodor and Pylyshyn 1988 present a critical analysis). For our purposes we shall confine ourselves with a rough indication of the central conceptions in these accounts. Following Rumelhart and McClelland (1986), we shall treat the term *connectionism* as a referring to a style of modeling that is based upon networks of interconnected simple processing devices. So, a connectionist model (of some empirical domain) is a model that is based upon networks of interconnected simple processing units. An important difference to more traditional approaches to computation is the parallel nature of processing. Instead of going through a sequence of serial processing steps, a connectionist model processes (potentially) huge amounts of information at the same time. In order to understand this difference a little better, we need to understand the central process of most connectionist models, namely the process of spreading activation. Each processing unit in the network can in principle assume two states: it can either be active or inactive. A unit becomes active when it receives a sufficient amount of stimulation from the units that it is connected with. The central idea of the process of spreading activation is that the activation of a set of nodes may results in the propagation of activation through the net thereby activating other nodes, which in turn may stimulate yet other nodes and so forth. So, we can think of spreading activation as a kind of chain reaction. Let us illustrate the general idea behind the process on the basis of a very influential model of word recognition namely McClelland and Rumelhart's *Interactive Activation model* (IAM; McClelland and Rumelhart 1981). Consider Figure 51.

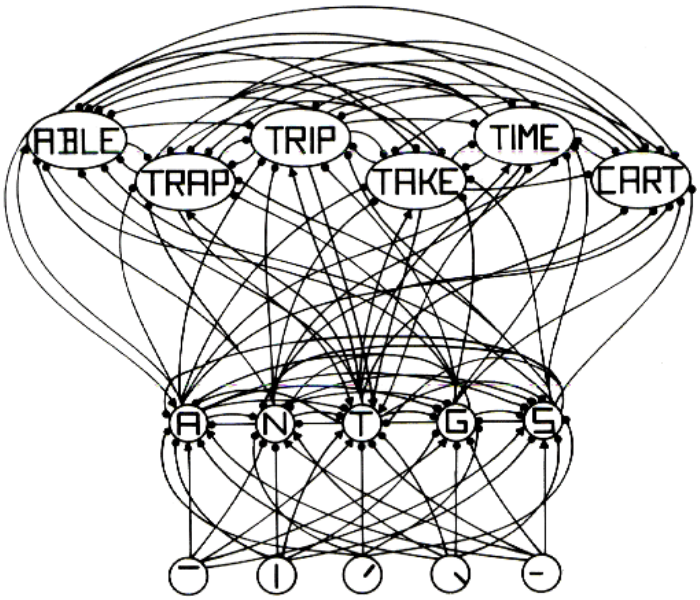

Figure 51: McClelland and Rumelhart's Interactive Activation model

Each circle in Figure 51 represents a node in the network and the lines that connect individual nodes represent links between the respective nodes in the network. The model consists of three hierarchical levels: the top-most level is the most complex level, in which each node represents individual words. The next lower level represents constituents of these words, say letters and the bottom level represents the most elemental level of constituents of letter representations, i.e. different sorts of vertical, horizontal, and diagonal lines. The nodes at the most fundamental level can receive their activation by the perception of the corresponding feature. This is to say that in this model it is assumed that the human perceptual apparatus effectively comprises *feature detectors* that respond only to certain stimuli. Ever since the groundbreaking experiments by Hubel and Wiesel, who would later receive the Nobel prize for their work on sensory processing (cf. Goldstein 2001 for an overview of that work), feature detectors are routinely assumed in neurophysiological treatments of perception. The presentation of a particular stimulus will cause some low-level nodes to become activated and these will in turn pass on activation to those nodes they are connected with. When a node at the next higher level has received a sufficient amount of activation, it too will become active and will send activation to connected units at the next higher level. Word recognition then corresponds to the activation of a node at the word level. The architecture of the IAM for

instance allows activation to spread in both directions (and is thereby capable of accounting for top-down effects) and also assumes both excitatory links and inhibitory connections among nodes. While the former can raise the activation level of the recipient node, the latter can lower the activation level of the receiving node. Such model types are powerful statistical pattern associators, which relate input patterns to output patterns and in the course of this association physically change their structures (most importantly the connection weights between units (cf. McClelland and Rumelhart 1989: Ch. 2). A quarter of a century has passed since McClelland and Rumelhart's formulations of the PDP-approach and half a century has passed since the developments of its earliest predecessor, the Perceptron (Rosenblatt 1958). It should not be too surprising that model architectures continue to develop further and can now address issues that could not be handled by these early models, e.g. the handling of temporal extend in parallel architectures.

The here presented description of the IAM in particular and connectionist models in general is of course a crude simplification of the matter. The goal at this point is only to convey the most fundamental ideas (for introductions cf. Feldman and Ballard 1982, Anderson 1995, Elman et al. 1996, McLeod et al. 1998, Marcus 2001, inter alia). The details of particular model architectures need not bother us here. What is important though is that connectionist models do not in any obvious way make recourse to working memory at all. Given the importance of this point, let me quote an authoritative statement at length:

> "*Within the connectionist framework, the processing of an input is achieved not through the action of rules or productions operating on declarative knowledge in a computational workspace [viz. working memory, DW] but rather through the passing of activation through a multilayer network. In this framework, the network's capacity to process information varies as a function of the input […], the properties of the network […], and the interaction of these properties—how much the network has experienced similar input before. […] [W]here is working memory? To the extent that it is useful to talk about working memory within these systems, it is the network itself* […]"

(MacDonald and Christiansen 2002:38)

This brings us (back) to a crucial property of connectionist models which is directly relevant for the present study and the overall force of the argumentation, namely the role of frequency. Even though particular connectionist models may very well vary in the details of their learning algorithm, they all share that connections between units are 'weighted' and this weight (or strength) is flexible and contingent on the co-activation frequency of the respective units. So once we assume that connectionist models present the right type of model of the human processing system, it appears problematic to utilize the concept of memory limitations in our explanations of processing difficulty (or any notion that presupposes the existence of a working memory component for that matter).

Once we fully adopt the frequency dependency of processing difficulty, many common sense considerations fall into place: people generally do not struggle so much with problems which they encounter and solve frequently. The task at hand may be inherently complex, but once it has been mastered this complexity produces no behavioral reflexes any longer. This is not only true for the activity of tying a shoe, but apparently carries over to linguistic behavior, too. The view on processing in general and comprehension difficulty in particular assumed here can be summed up by the following statement.

> "*Comprehension difficulty [...] emerges from several competing structures ultimately derived from distributional patterns of language use*" (Gennari and MacDonald 2006)

While certain intrinsic properties of linguistic structures such as complexity may be one cause of the observable frequency distributions, we cannot deduce the complexity of a structure from its usage frequency. And any *resource limitation* account is doomed to make the wrong predictions, if it fails to recognize the impact of usage frequency on resource consumption. As we have seen a number of times already in this study, it certainly appears as

if discourse-functional need is a better predictor of frequency and, a fortiori, processing difficulty.

The here advocated alternative view on language comprehension came to be called the *constraint-based* or *constraint-satisfaction* view on language comprehension (cf. MacDonald and Seidenberg 2006 for a general discussion). In these accounts comprehension difficulty is understood as a classification task that requires the satisfaction of a number of probabilistic constraints. In this account, language comprehension is characterized as continuous and homogenous and a great many of informational sources and processes are used at the same time. From a computational perspective, it is very natural to employ connectionist models as a means to implement these constraint-satisfaction processes.

At this point, we may return to the issue of center embedding and the difficulty of patterns that exhibit this structural property. It is interesting to note that many studies that have reported the difficulty of center embedded structures have used test sentences like the following (taken from Davis 1995).

(62) The child [the dog bit] developed rabies. [single CE]

(63) The child [the dog [the man shot] bit] developed rabies. [double CE]

The observation is that sentences like (62), which exhibit a single embedding, are relatively easy, while sentences like (63), which exhibit double embedding, are nearly incomprehensible. One possible way to account for these differences is to focus on their structural properties and there is a strong tradition in the psycholinguistic literature to adopt such a perspective. Starting with Kimball's *Principle of two sentences* (Kimball 1973), which states that the constituents of no more than two sentences can be parsed at any one time, we can trace a huge amount of literature working from the assumption that it is the intrinsic complexity of a pattern and the architecture and way of operation of the human parser that governs these phenomena (cf. Koster 1978, Gibson 1991, Stabler 1994, Lewis 1995, inter alia). Alternatively, we could explain the observed difficulties on the basis of the very low

frequency of double center embeddings. Pointing out the relative infrequency of the pattern would be the first step in the explanation. This first step targets a mechanistic level: as a rough approximation we may posit that high frequencies of use facilitate future processing, because each processing event has a direct impact on the connection weights among the involved units. A high frequency value of an expression type E counts as a sufficient condition of E being relatively easy to process. So, high frequency entails low difficulty. The frequency hypothesis predicts that low frequency items are harder to process. But not all constructions that are equally infrequent are equally difficult to process. People have less trouble with multiple right embeddings, which are also rather infrequent. If we assume—for the sake of argument—that triple right embeddings are as infrequent as triple center embeddings, the frequency account would clearly make wrong prediction (it would predict the same amount of difficulty). While this certainly looks like a strong argument against the frequency account, there is still room for vindication. The frequency account may receive unanticipated back-up from analogy. Triple right embeddings are very similar to simpler (and highly frequent) structures. In fact, a sentence like (64) can be processed as a sequence of simple transitive sentence:

(64) Wayne likes Cathy$_i$ [who$_i$ likes John$_j$ [who$_j$ likes Mary$_k$ [who$_k$ likes Peter]]].

If we assume that the occurrences of *who* in (64) are pronominal referring expressions, then (64) can be interpreted in the very same way we can interpret four conjoined simple transitives with identical anaphoric relationships. Consider (65).

(65) Wayne likes Cathy$_i$ and she$_i$ likes John$_j$ and he$_j$ likes Mary$_k$ and she$_k$ likes Peter.

Evidence for the processor's sensitivity of the close relationship between right embedded subject RC and conjoined simple transitive clauses comes from first language acquisition. The structural similarity between these types has been suggested to account for the relative

ease with which children acquire subject relatives in English, which has been reported in numerous studies (cf. Bever 1970, Tavakolian 1981, Diessel and Tomasello 2005, inter alia). Diessel and colleagues have argued repeatedly for the idea that analogical processes guide the acquisition process (Diessel and Tomasello 2005, Brandt et al. 2008, Diessel 2009). If the processing of an expression E is influenced by the frequency of a similar structure E', the frequency-based explanation would receive further reinforcement. Simple transitive structures without doubt are highly frequent and if they play a causal role in the processing of similar structures, we would expect multiple right embeddings to be easier. The analogy hypothesis has recently received further confirmation from computational modeling: Fitz and Chang (2008) report the learning behavior observed for children could be re-produced by a connectionist model (Type: Simple Recurrent Network, cf. Elman 1990), which was trained on the basis of 10,000 simple sentences of various syntactic types. Figure 52 presents their results

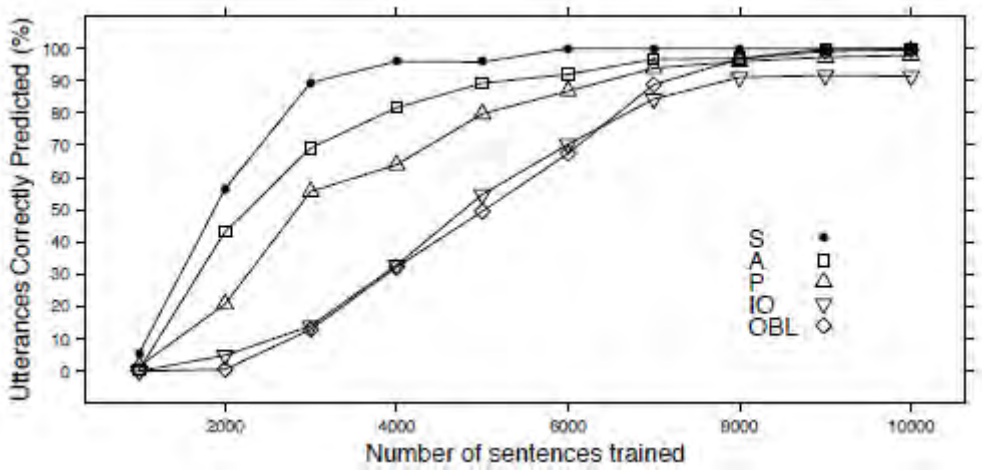

Figure 52: Order of RC acquisition in Fitz and Chang's model (taken from Fitz and Chang 2008)

Figure 52 plots the number of correctly predicted utterances in percent against the magnitude of the training set. We observe that subject RCs (S=intransitive subject RC; A=transitive subject RC) required the lowest numbers of training sentences to reach adult like performance. Note that Fitz and Chang's results not only corroborate the work from Diessel and colleagues, but also suggest that some type of connectionist model can in fact mimic human behavior. While these findings certainly are very encouraging for the view presented

here, we should note that human-like behavior constitutes only a necessary and not a sufficient condition for the adequacy of connectionist models to emulate human symbolic processing.

We have seen that a frequency-based approach to sentence processing can in fact account for the empirical phenomena, when it is complemented by an account of analogical processes. There can be no doubt in the idea that analogy is a central concept in contemporary cognitive science and for many it is actually the very core of human cognition (Gentner et al. 2001 presents an overview of research from various fields including developmental and comparative psychology, neuroscience, artificial intelligence, linguistics; and philosophy). Now, once this first *why*-question (Q: Why is the pattern difficult?) has been answered (A: It is extremely rare), we must of course be prepared to account for the relative infrequency of the pattern in question. While intrinsic properties may or may not play a role here, there certainly is more than that. As Limber (1976) notes, even simple center embedding may very well be infrequent for pragmatic reasons. Center embedded RCs tend to modify subjects and subjects are topics in English. From such an information structure/discourse-pragmatic perspective it is not exactly surprising that center embedded RCs are rather infrequent in natural discourse and hence harder to process. Of course, it follows that from this view multiple center embeddings are even harder as it appears rather difficult to even conceive of a communicative situation in which a need for such a structure would arise.

I should be perfectly clear about the fact that it is not my contention to deny that structures do have intrinsic properties and that some of these properties do have an impact on the processing demand associated with that structure. It certainly is plausible that CE may be harder for structural reasons. In RE structures both clausal constituents can be processed in a serial fashion. Hearers can first construct a situation model for the MC and then construct one for the RC. In the CE case however, the processing of the MC is interrupted, which might very well influence the ease of building the situation model (cf. Slobin 1973 for a discussion of this non-interruption hypothesis). Rather what I would like to argue for is that we should not overrate the impact of these intrinsic features. Formulating explanations in terms of

consumed resources (working memory loads) may be misleading as it goes too far beyond what is observable. The discussion of the connectionist models was supposed to show that we can do without such a component altogether. In fact, should it turn out that connectionist models are indeed the right type of model and these model do not assume a WM-component then all our WM-dependent explanation "go right out of the window". My point is that we should try to avoid unnecessary stipulations and resist the urge to formulate explanation with theories that make such unnecessary stipulation.

3.4.2.2 Corpus analysis: Center versus right embedding

Without further ado, we may now proceed by looking at some data. Figure 53 presents the general distribution of the embedding types in the present data set (n=1000).

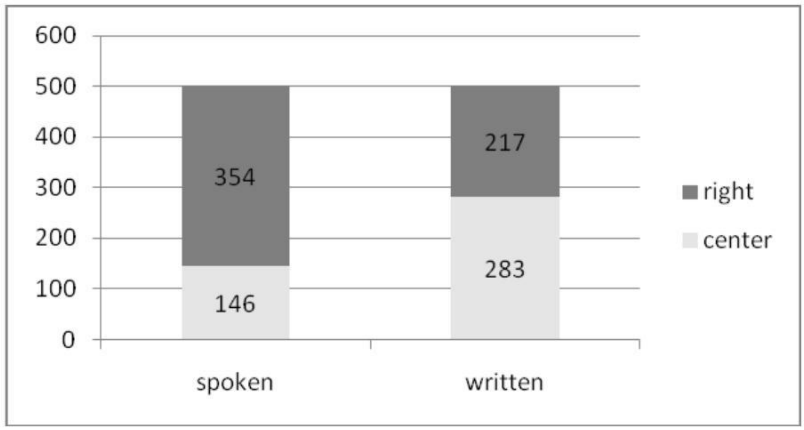

Figure 53: Type of embedding across modalities

The number of right embedded RCs is greater than that of the—allegedly more difficult—center embedded ones, but the difference is not as pronounced as one might have a suspected: 429/1000 RCs are center embedded. But as Figure 53 shows, most of these are contributed by the written sample, in which they actually outnumber their right embedded counterparts. In spoken language we indeed observe a strong bias towards right embedding. A statistical evaluation of the distribution gives rise to Figure 54.

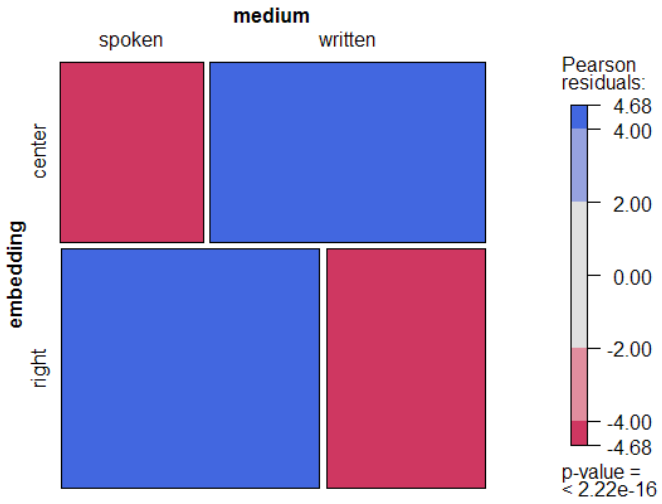

Figure 54: Type of embedding across modalities

We observe that the positive association of center embedding and the written register is quite pronounced (globally: Chisq = 76.62, df = 1, p-value < 2.22-16, Cramer's V: 0.277). There is a highly significant positive association between right embedding and spoken discourse. Again it appears as if language users do not shy away from more complex constructions in writing but prefer simpler ones under on-line constraints. Alternatively, we may presume that there is a strong need for the discourse function carried by right embedded RC in spoken language, whereas there is a strong need to express the function carried by center embedded RC in written language.

To restrict our perspective to a single contrast, viz. type of embedding, certainly has the potential to cloud other interesting relationships. In order to get a clearer idea of what may underlie the distribution in Figure 55, it is helpful to consider some additional variables. Figure 56 provides an overview of the frequencies of certain patterns we can distinguish by letting the variable FORMALITY, INTERNAL ROLE, and FINITENESS into the picture (all variables were coded as binary factors to minimize the complexity of the resulting hierarchical structure).

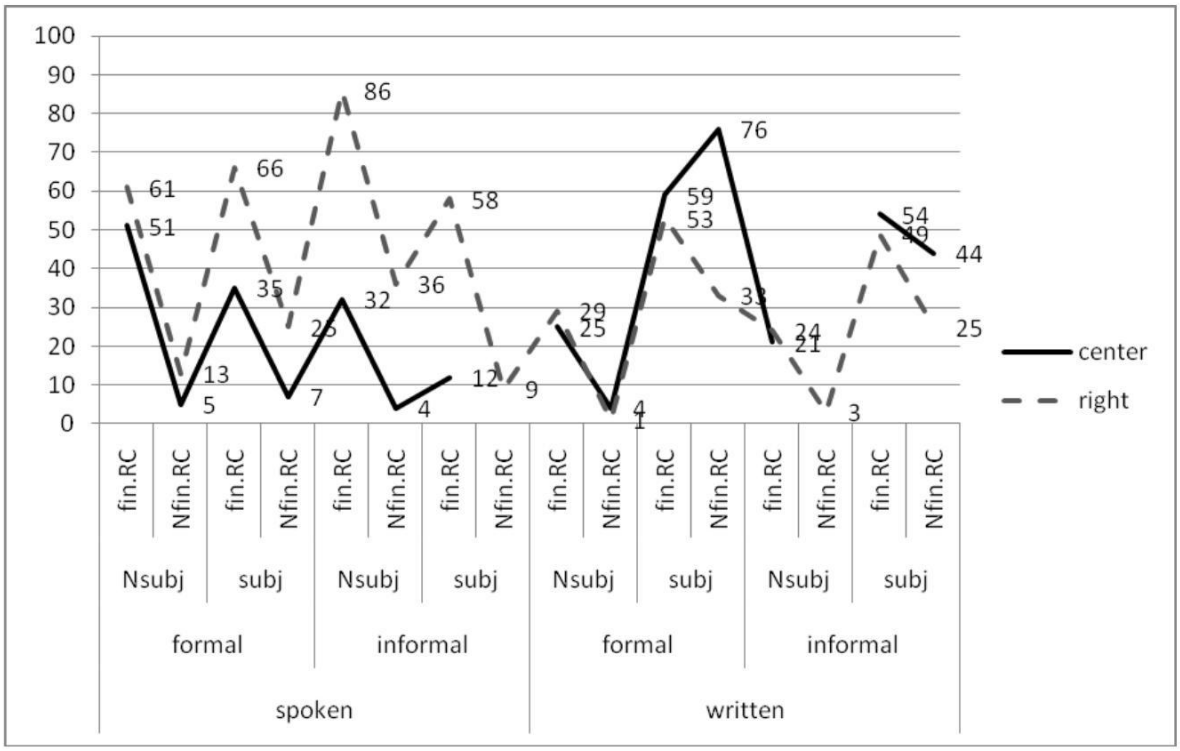

Figure 55: Embedding across sub-types

We need not discuss each of the sixteen subtypes in Figure 55. What we should note however is that the most pronounced bias towards right embedding is attested for finite non-subject RC in informal spoken language. The subset of RCs that instantiate the feature list {- FORMAL; - SUBJECT, + FINITE } exhibits a ratio 86:32 (in favor of RE) in spoken language. The exact opposite pattern, i.e. the feature list {+ FORMAL; + SUBJECT, - FINITE} exhibits a ratio of 33:76 in written language, thereby constituting the strongest CE bias of the contrasted subtypes. The pronounced formal differences of the two patterns strongly suggest different functions, when viewed from the perspective of construction grammars, which usually subscribe to the principle that "a difference in syntactic form always spells a difference in meaning" (Bolinger 1968: 127). The differences in register and especially formality are very reconcilable with the discourse functional explanation. We will postpone a statistical analysis of these distributional differences until later sections as the analysis would strongly benefit from a multivariate treatment. However, we may at this point acknowledge the promises of a configurational view, which we will embrace in Chapter 4.

3.4.2.3 Corpus analysis: Role of modified MC element

Having discussed the distribution of the types of embedding, we may now have a more detailed look at the external syntax of the RC and look at what grammatical roles of the MC are modified by way of a RC. Figure 56 presents an overview.

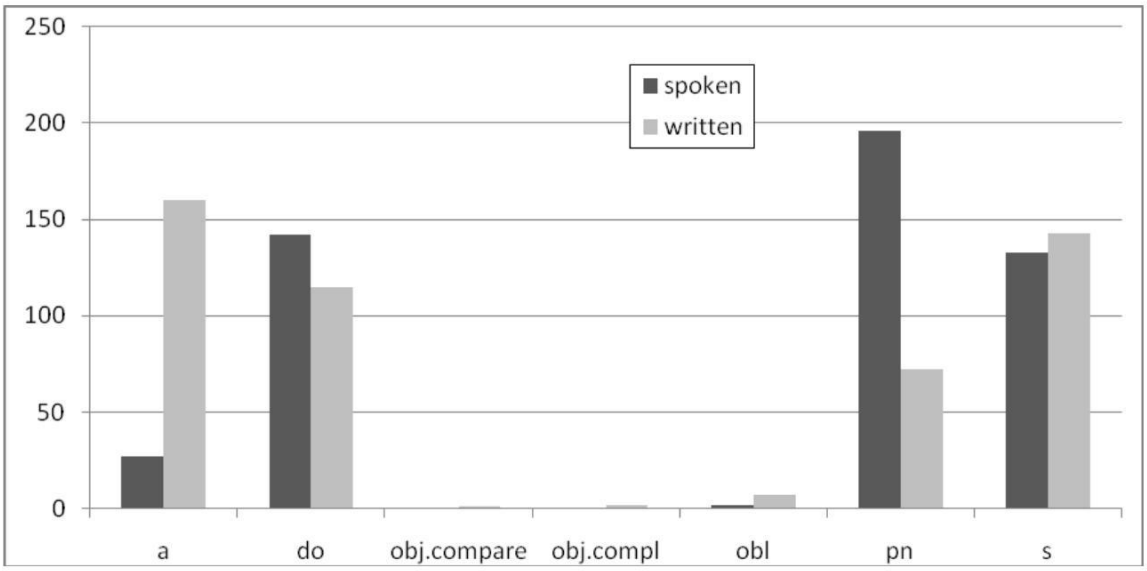

Figure 56: External role of head across modalities

We observe that spoken RCs tend to modify either predicate nominals (pn) or direct objects (do) or subjects (a/s). The category subject was subdivided into subjects of a transitive clause (a) and subjects of an intransitive clause (s). The latter is about five times more frequent than the former. The remaining roles are negligible and barely exceed the 1% margin (their joint frequency is 12/1000), which is why they have been erased from the data that were submitted to statistical analysis (n=988). A chi-squared test of the interaction of EXTERNAL.ROLE and MEDIUM discloses significant associations (χ^2 = 155.01, df = 3, p-value < 2.2e-16, Cramer's V=0.396). These associations are illustrated in Figure 57.

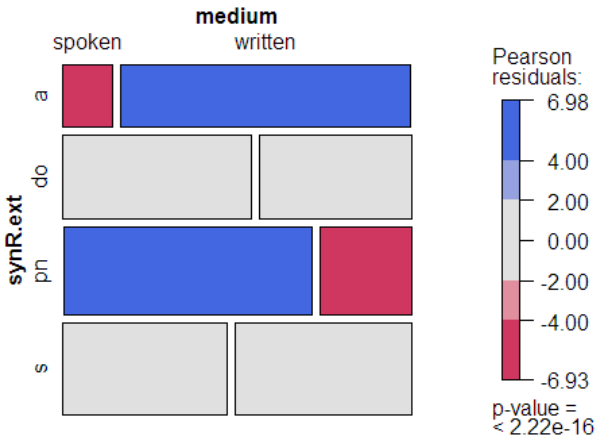

Figure 57: External syntax across modalities

There is a strong positive association between written language and modification of transitive subjects. Spoken language shows a strong preference towards modification of a predicate nominal. The distributions of intransitive subject and direct object modification are very much with the limits of statistical expectation.

Again there are good reasons to believe that these preferred patterns are the result of the discourse functions that the respective structures serve and the typical communicative needs that are typically present in the respective registers. We can help ourselves to some first tentative empirical evidence for this view, when we add the variable formality into our overview. The idea behind this addition is that the level of formality can be viewed as a crucial factor in typical discourse function. Figure 58 presents the results.

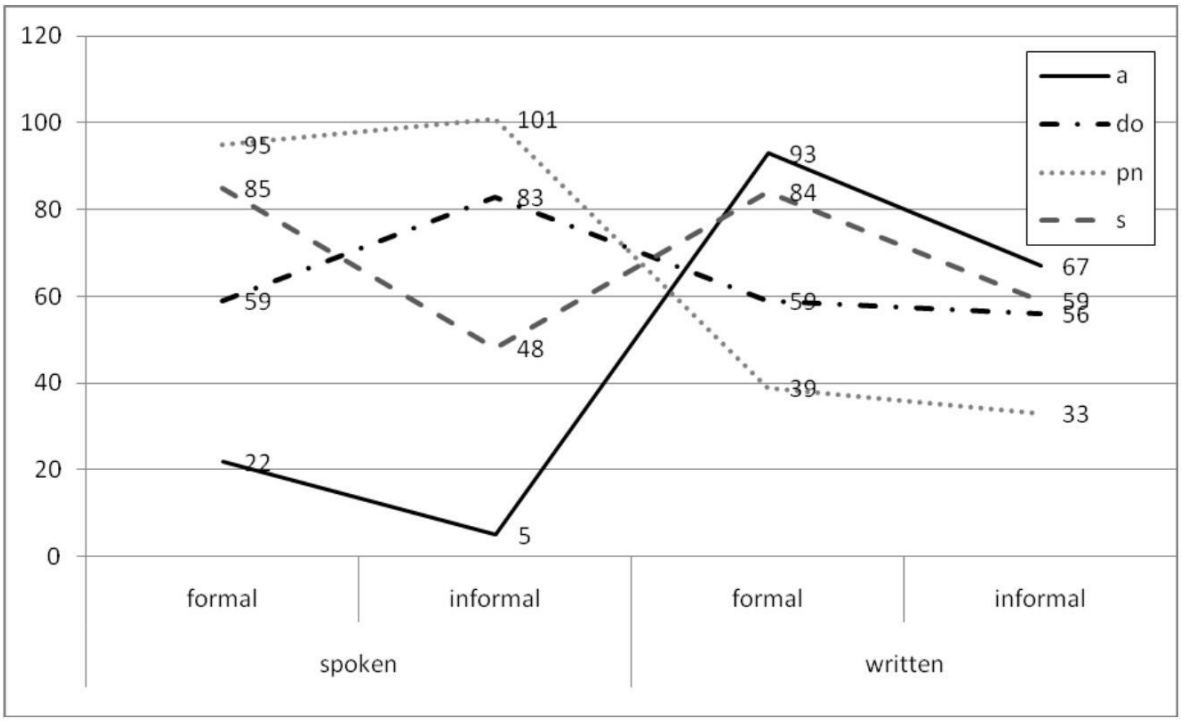

Figure 138: Modified MC role across registers and degree of formality

The four graphs denote the contrasted syntactic roles of the MC that have received RC modification. As the marginal roles have been dropped, we are looking at a total of 988 usage events that are classified with respect to both modality and formality. The numerical contributions of the respective factor levels of the variables register and formality are fairly even ($N_{written}/N_{spoken}$= 490/498; $N_{formal}/N_{informal}$=536/452), which I hope excuses the lack of normalization here. Taking levels of formality into consideration, we are able to observe further hints for the discourse-functional motivation for the frequency of the four types (S, A, PN, DO). The preference for A-modification attested for written language is stronger in formal contexts than in informal contexts. The converse relationship can be observed for the spoken modality, which exhibits a strong preference for PN-modification. This bias is stronger in less formal situations (even though the difference is not very pronounced). Interestingly we can also observe a difference in the usage of DO-modification in spoken language, but not in written language. And finally, there is a strong difference in S-modification usage across formality levels in both registers (in opposite directions). All these differences are very much in line with the idea that it is discourse functional pressure that

accounts for usage frequency. However, we will again postpone our discussion of these discourse functions for the simple reason that the formal specification is too vague to allow for a decent mapping from form to function. Once we have reaches more detailed levels of formal description (in Chapter 4), we will make an attempt to pin down the discourse functions that the different patterns may serve.

We may however discuss a more general point that concerns the relationship between the intrinsic complexity of an expression, the frequency of that expression and its and discourse-function. The point I would like to raise may be presented as follows: Even if we assume for the sake of argument that we can indeed predict the processing difficulty of a pattern E from the frequency of E (and—via the argument from analogy—the frequency of similar patterns) and even if we can explain in turn these frequencies on the basis of their discourse functions, does this rule out an account that explains processing difficulty on the basis of E's intrinsic properties, say its complexity? The answer I would like to propose is that this question presupposes a false dichotomy. Instead of utilizing either the discourse-function of E or the complexity of E, we may think of the complexity of E as an integral part of the discourse-function of a linguistic expression. While this may require a widening of the notion discourse-function, it certainly yields some quite intuitive results: a language user may for example choose a highly complex syntactic pattern in order to attain certain communicative goals, say impress his audience. Similarly, a language user may use very simple language when talking to an uneducated audience to signal group affiliation. This is to say that the complexity of an expression has certain effects beyond processing demand. We may spell out a theoretical treatment of such choices either in terms of perlocutionary effects as assumed in speech act theory (e.g. Searle 1969) or in terms of adaptive behavior if we frame the phenomenon in a more sociolinguistic context (e.g. Hymes 1974) or in some other framework that acknowledges the effects that linguistic choices have on the negotiation of social relationships and power.

3.5 Cross-clausal features

We may now turn to our last set of features of interest. As the name suggest, these factors are characterized by the fact that they target a particular relationship between the clausal

constituents of an RCC. We will start our discussion with a look at the transitivity configurations of RCCs. We have already had a look at the transitivity of each clause and are now enriching this information by looking at the interdependencies of the values, which again will bring us one step further in our attempt to understand of the typical patterning of RCCs. We shall then turn to issues of parallelism. As the head of the RC is standardly portrayed as playing a role both the structure of the RC and the MC, we can in principle have two scenarios: the roles can be identical or they can be different. In the former case, we may speak of *role parallelism*. Section 3.5.2 will focus on syntactic parallelism, i.e. the grammatical role played by the head inside and outside the RC, whereas § 3.5.3 will investigate parallelism on a semantic level, i.e. the thematic role played by the head inside and outside the RC. Finally, the last sub-section will investigate the characteristic of both the head and the subject of non-subject RC. A number of grammatical and conceptual features have been proposed to influence the processing demand of a RCC type.

3.5.1 Cross clausal features: Transitivity configurations

We have seen in § 3.3.2 and § 3.4.1 respectively that the typical transitivity value of an RC in biclausal RCCs is monotransitive in both registers, while the MC tends to be either copular or monotransitive. While the former is typical for spoken language the latter was the dominant type in written language. While we expect that the relative frequencies of these sets result in large intersections, i.e. a large number of $MC_{(COP|MONOTR)}$ [RC_{MONOTR}] structures, it is still worthwhile to test whether this or other combinations is statistically special or not. Figure 59 presents an overview of the transitivity patterns across modalities.

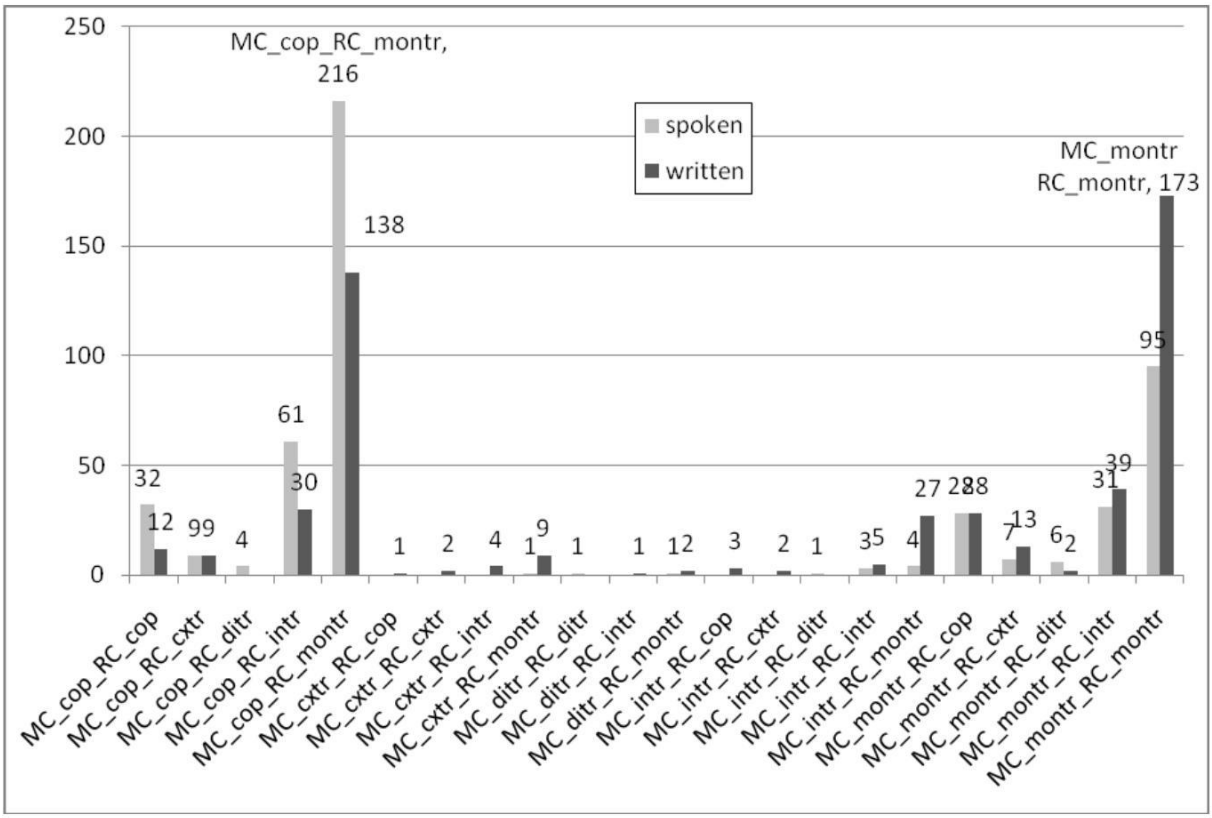

Figure 59: Transitivity configurations across modalities (overview)

If we focus on those combinations that occur at least ten times, we reduce the number of patterns from 22 to 10 and retain a data set of 962 data points.

Table 9: Top ten transitivity configurations

	medium		
TrMC_TrRC	spoken	written	TOTAL
MC_cop_RC_cop	32	12	44
MC_cop_RC_cxtr	9	9	18
MC_cop_RC_intr	61	30	91
MC_cop_RC_montr	216	138	354
MC_cxtr_RC_montr	1	9	10
MC_intr_RC_montr	4	27	31
MC_montr_RC_cop	28	28	56
MC_montr_RC_cxtr	7	13	20
MC_montr_RC_intr	31	39	70
MC_montr_RC_montr	95	173	268
TOTAL	484	478	962

The dominant construction type for spoken language is the combination of a copular main clause and a monotransitive relative clause. This suggests that for spoken RCCs, the

subordinate clause does actually carry more information than the corresponding main clause. Due to its relatively low informational status, the main clause is not likely to serve as the focal point of attention in the RCC structure. Rather it is likely to (merely) provide a syntactic framing of the heavier RC predication, so that the schema underlying such RCC types may serve specific discourse functions, say putting the RC predication in focus. The (numerically) dominant pattern in written language is the combination of two monotransitive clausal constituents. If we use transitivity to approximate informational richness, we are led to assume that both clausal are equally loaded in this modality. A global chi square test of independence discloses a significant divergence form independence (χ^2= 85.684, df = 9, p-value = 1.191e-14). Figure 60 helps us grasp the most distinguishing configuration types (for expository convenience, only the top five have been plotted here).

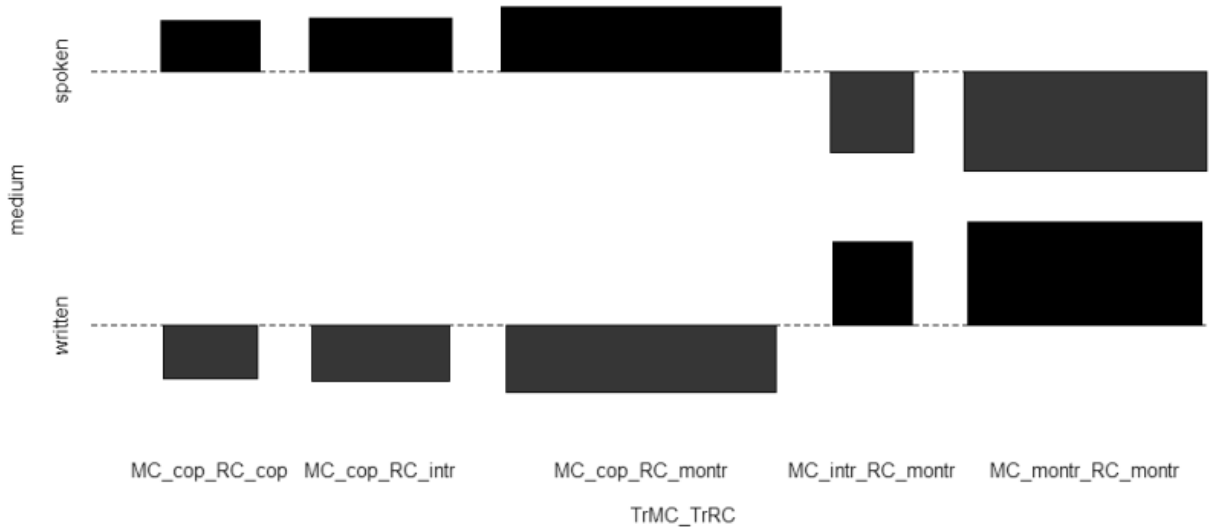

Figure 140: Association plot transitivity configurations X medium (Top five)

While Figure 60 confirms some of the characteristics we have already pointed out, namely the preference for $MC_{copular}[RC_{monotrans}]$ patterns in spoken language and the preference for $MC_{monotrans}[RC_{monotrans}]$ in written language, it also presents additional information that is relevant for our purposes. The utilization of the abovementioned top configurations is consistent with the simple and straightforward idea that written language is simply more complex than spoken language. The dominant RCC pattern in written language involves more arguments and can thus be considered to express heavier propositions. However, by that

logic it certainly is surprising to see significantly more $MC_{intrans}[RC_{monotrans}]$ patterns in written discourse as these involve the same number of arguments. Together with the fact that all three patterns that are characteristic of spoken language are specified for a copular MC, this finding provides additional evidence for the hypothesis that it is discourse-function and not complexity that best accounts for the observed distributions.

3.5.2 Cross-clausal features: Syntactic parallelism

It has been suggested in the psycholinguistic literature—predominantly in the context of language acquisition—that RCC are easier to process when the head plays the same role in both clausal constituents. This proposal has been termed the 'parallel function hypothesis' (Sheldon 1974). Sheldon tested 3 to 5-year-old children in their comprehension of four types of RCCs: $subject_{MC}$-$subject_{RC}$ (SS), $subject_{MC}$-$object_{RC}$ (SO), and $object_{MC}$-$subject_{RC}$ (OS), and $object_{MC}$-$object_{RC}$ (OO) constructions. She observed that SS-RCCs and OO-RCCs are significantly easier to comprehend than the mixed patterns. From the perspective taken here, we would expect there to be a reflex in the frequency signatures of the patterns such that those patterns which show identical should be frequent or at last more frequent than expected on the basis of chance. Table 10 presents the results of the corpus analysis.

Table 10: Syntactic parallelism (complete set)

		RC				
		do	obl	other	s	
MC	do	**70**	38	3	146	257
	other	1	4	0	7	12
	pn	80	44	6	138	268
	s	104	43	2	**314**	463
		255	129	11	605	1000

We observe that SS patterns are indeed the most frequent type (n=314), which is very much in line with the idea that these patterns are easy. It should be easier for children to pick up patterns that are highly frequent in the ambient language. The experimentally observed ease of SS patterns can thus be attributed to frequency effects. However, the ease of OO patterns does not seem to follow from the frequency hypothesis. Their co-occurrence frequency is

neither particularly high nor is it statistically significantly greater than expected. Figure 61 presents the results of a statistical analysis of the data in Table 10.

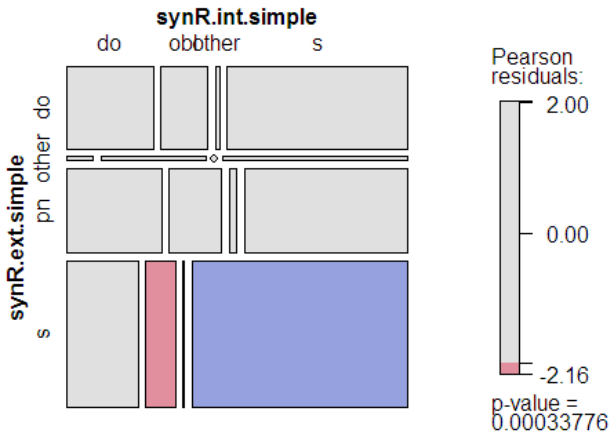

Figure 61: Syntactic parallelism (complete set)

As indicated by the (light) blue coloring there is indeed a statistically significant association between subject modification in the MC and subject extraction. The type S-Obl, i.e. the type where an oblique relative modifies the main clause subject, is somewhat avoided. It occurs significantly less frequent than expected. This is likely to be due to semantic reasons: while subjects are associated with high semantic roles, obliques (as they are defined here) encode low roles expressing the goal or manner of an action or specify spatio-temporal parameters. Such a discrepancy may very well be hard to conceptualize or simply unpractical for most discourse purposes. It certainly prevents the referent of the NP in question to be human as human referents make no good obliques or adjuncts. But the S-Obl pattern does occur in the data (n = 43). The example in (66) is typical for this type.

(66) Some of the **novels** by women, **in which** the narrator concentrates on male politics […], are also centered on a heroine […]. (W2B-009 #025:1)

So, why do the corpus results only partially mirror the experimental findings? A possible reason is the fact that we have not yet looked at register differences. We have already observed for many variables that virtually all processing predictions are usually met only by

spoken language. Hence, we should get a clearer idea of the degree to which the corpus results reflect Sheldon's experimental findings and corroborate the frequency hypothesis, when we look at the registers in isolation. A Cochran-Mantel-Haenszel test reveals that there is indeed a significant difference in the proportions across modalities (Cochran-Mantel-Haenszel $M^2 = 17.27$, df = 9, p-value < 0.044). However, the direction of the difference is not as predicted by a processing account. The distributions for spoken register, which should be more sensitive to processing factors, look rather disencouraging. Consider Table 11.

Table 11: Syntactic parallelism (spoken discourse)

		RC				
		do	obl	other	s	
MC	do	55	19	3	65	142
	other	0	0	0	2	2
	pn	75	34	5	82	196
	s	81	15	1	63	160
		211	68	9	212	500

The distribution shown in Table 11 is not statistically significant ($\chi^2 = 13.32$, df = 9, p-value = 0.1486). Furthermore, both the SS pattern and the OO patterns are even a little less frequent than expected on the basis of chance (SS: $F_{exp} = 67.85$; OO: $F_{exp} = 59.92$). So, the statistically significant association of the subject roles reported earlier has to be due to what happens in the written part. Table 12 presents the corresponding data.

Table 12: Syntactic parallelism (written discourse)

		RC				
		do	obl	other	s	
MC	do	15	19	0	81	115
	other	1	4	0	5	10
	pn	5	10	1	56	72
	s	23	28	1	251	303
		44	61	2	393	500

But maybe the findings reported here are incompatible with the parallel function hypothesis because this hypothesis was derived from only a subset of RCC types. The RCs that Sheldon

employed in her experimental setting were exclusively finite RCs and were restricted to the class of subject and object RC. This minimal contrast between just two roles of course results in a greater likelihood of parallelism as two out of four possible states instantiate parallel roles. To allow for a better comparison with Sheldon's results the analysis was restricted so as to include only these subtypes (n=425). Figure 62 presents the results.

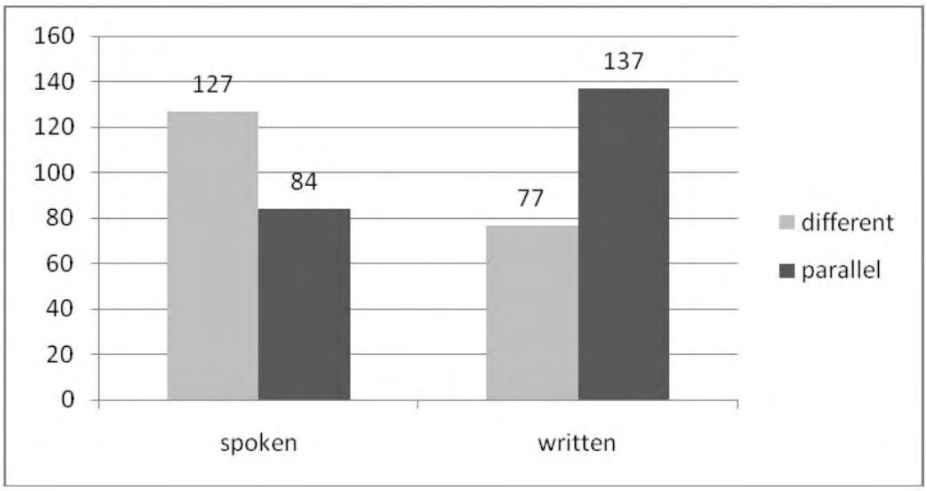

Figure 62: Parallelism across modalities (finite SRC -ORC)

We can compute a chi-square goodness of fit test for each modality to evaluate the distributions. The prior probability of parallelism was set to 0.5. For the spoken distribution we observe that there actually is an above chance bias towards non-parallelism ($\chi 2 = 87.73$, df = 1, $p < 0.0031$). For the written modality, we do get the predicted bias ($\chi 2 = 16.82$, df = 1, p-value = 4.105e-05). When cross the variables SYNTACTIC PARALLELISM and MEDIUM, we observe a statistically significant association deviation from independence ($\chi 2 = 24.945$, df = 1, p = 5.89e-07, Cramer's V = 0.242). Figure 63 presents the corresponding mosaic plot.

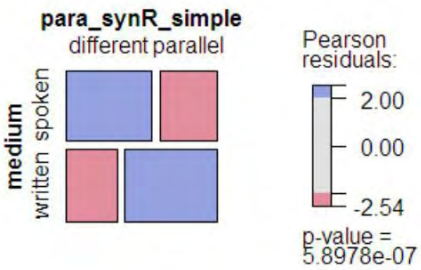

Figure 63: Mosaic-plot PARALLELISM X MEDIUM

The fact that written but not spoken language shows the bias towards syntactic parallelism surely is at odds with the parallel function hypothesis, if it is applied to adult on-line processing. However, it is possible that the hypothesis still describes a psychologically real phenomenon of language acquisition. It certainly is conceivable that a contrast that is relevant in early stages of the human linguistic competence but loses its relevance once an adult like state has been reached. As the present corpus data are samples of adult language, no claims can be made about child language and principles at work in processes of language acquisition.

We may close our discussion of syntactic parallelism with an overview of what configurations underlie the parallel/non-parallel distinction. Figure 64 presents the frequencies of the investigated RCC subtypes across modalities (finite RC; only subject or object roles; n=425).

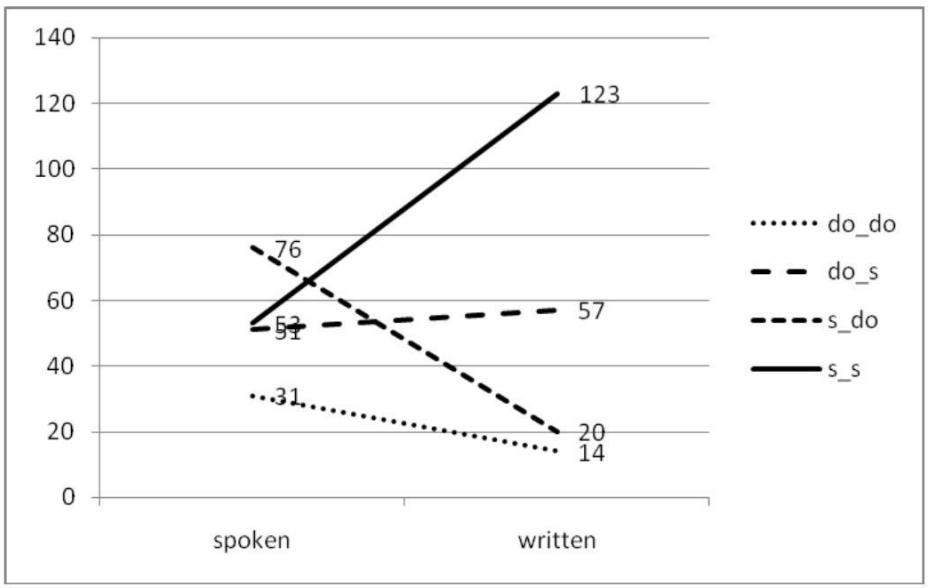

Figure 64: Role configurations across modalities (finite SRC - ORC)

Figure 64 can be viewed as a more coarse grained view on the distribution presented in Tables 12 and 13 and presents a refinement of the +/- parallel role overview in Figure 64. The first part of the description of the contrasted subtypes indicates the role the head plays in the MC, whereas the second element specifies the relativized role. We observe that the distributions of relatives modifying the direct object of the MC (do_s & do_do) are quite

similar in their frequencies across modalities. But while spoken language comprises the by far larger amount of object relative modifying the MC subject (s_do), finite written RCCs are dominated by subject relatives modifying the MC subject (s_s). A more subtle observation we can make at this point is that at least (76 + 52 =) 128 out of 146 (~ 88%) center embedded RC in spoken discourse are finite RC (cf. § 3.4.2.2). In contrast, written center embedded RCs are more likely to be non-finite. The finite types listed here account for (123 + 20 =) 143 out of 283 (~ 50%) of the center embeddings in that register. But again our multivariate descriptions in Chapter 4 will allow for more precise characterizations.

So, while SS relatives are quite frequent (at least in written language), which would to some account explain the ease of processing, the corpus-data do not appear to corroborate a more general parallel function hypothesis (at least not if that parallelism is conceived of as manifesting itself on a syntactic level).

3.5.3 Cross-clausal features: Thematic parallelism

It is possible though that the relevant parallelism is actually situated at a semantic level, rather than on a syntactic one. Maybe what the parallel function hypothesis is really about concerns the roles that the entities talked about play in the situations described. There is of course an intricate relationship between syntactic and semantic roles, but the mapping is quite complicated. Even though agentive roles are likely to occupy subject positions, this need not be the case. A principled exception to this tendency can be observed for passive constructions (syntactic subjects encode low semantic roles). Given the role that voice plays for the relationship between syntactic and semantic roles, a look at the distribution of active and passive constructions appears worthwhile. Figure 65 presents an overview.

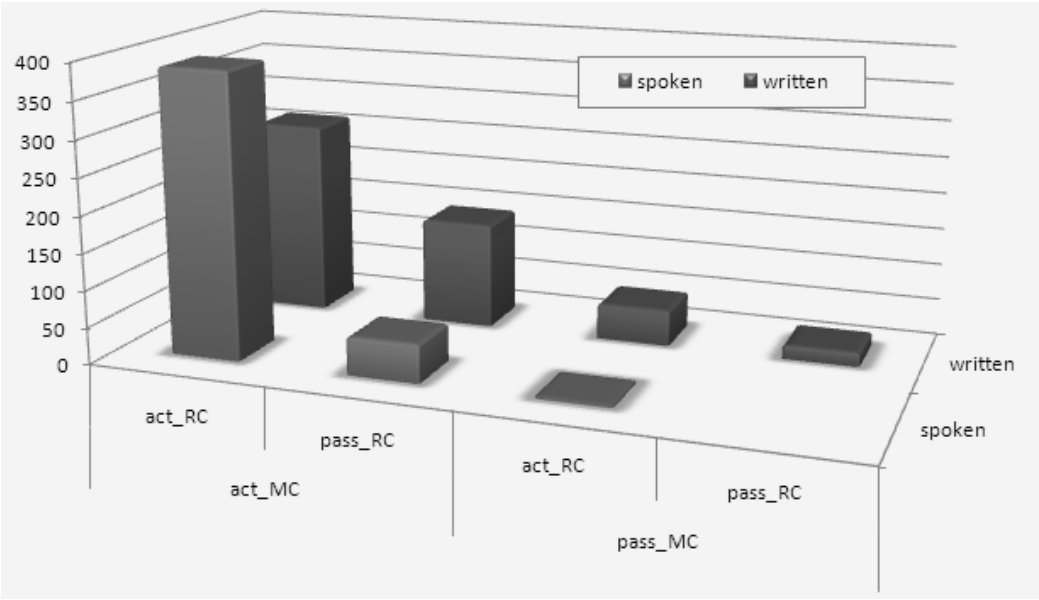

Figure 65: Voice of the clausal constituents across modalities

Figure 65 shows that RCC types with an active relative clause embedded in an active main clause (ACT_{MC}_ACT_{RC}) are dominant in both modalities, although passives are more common in written language.[17] It is interesting to see that RCCs with a passive main clause are so rare—particularly in spoken language where there are just three occurrences attested, all of which show an active relative clause. Querying the complete corpus for ((,CL(¬exclam, ¬inter, ¬imp, ¬subjun))), i.e. for declarative clauses, for each voice level reveals a ratio of (57417/11056=) 5.2 in favor of the active construction. If we consider spoken language only, this ratio is even more pronounced. Here active patterns are (38476/4446=) 8.6 times more frequent than passives. So, given the general preference for active constructions over passive ones, the RCC type ranking is in line with what we would expect. However, for RCCs the voice difference is a lot more pronounced than it is for declarative in general: only 3 out of 443 RCC have a passive main clause, which means that for every passive type there are 147.66 active ones. We do not need to bother about statistical hypothesis testing to

[17] The relative frequencies of RCC are very similar across registers and the difference across modalities is not significant (Mantel-Haenszel $X^2 = 1.2976$, df = 1, p-value = 0.2547, common odds ratio = 0.69).

acknowledge that the difference of 5.2 (or 8.6 for that matter) and 147.7 is rather striking. But at this point we can only speculate why this is the case. Part of the reason may be that passives tend to have fewer overt arguments than their active counterparts and hence offer fewer attachment sites. This overview is of course still quite coarse-grained and confounded with many variables that may distort the picture. However, there may be more to it and given the straight cline in the frequency of the respective types, i.e. $ACT_{MC}\_ACT_{RC} > ACT_{MC}\_PASS_{RC} > PASS_{MC}\_ACT_{RC} > PASS_{MC}\_PASS_{RC}$, a processing explanation should not be excluded.

But let us return to the semantic issue and the analysis of the distribution of semantic roles. We may start with a rather fine-grained overview. All roles were assigned according to the criteria described in Quirk et al. (1985). Figure 66 presents the roles played by the head in the main clause.

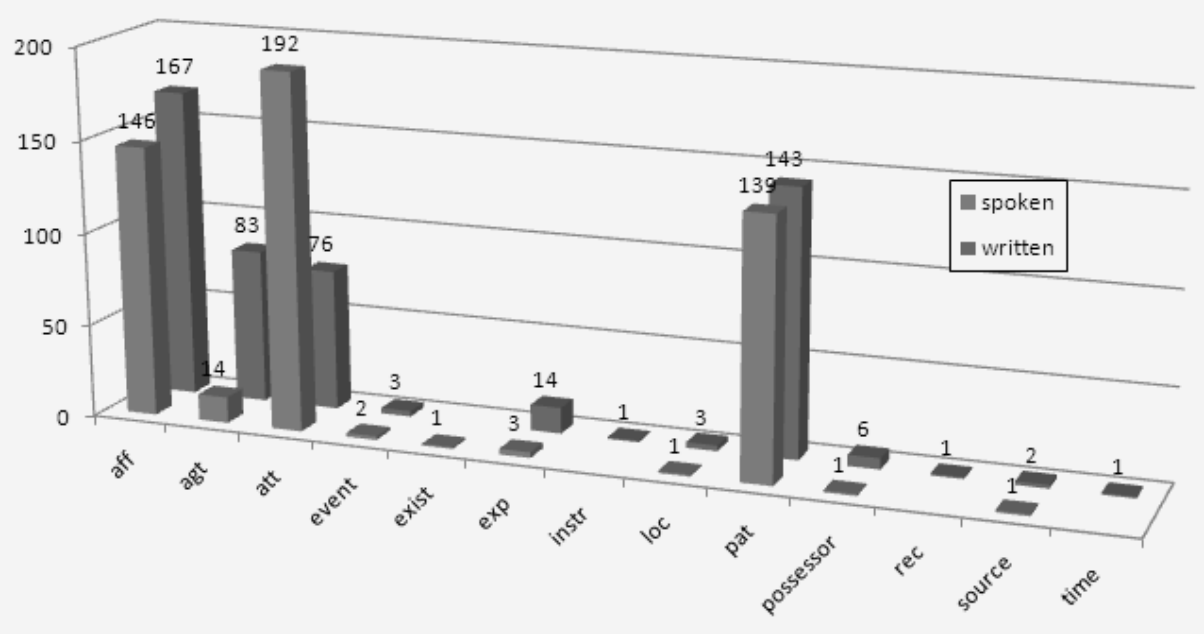

Figure 66: Thematic roles (main clause)

Figure 66 shows that many of the roles used in the semantic annotation are negligible. The lion's share of the examples (96%) assumes one of four values: AFFECTED, AGENT, ATTRIBUTE, PATIENT. The role AFFECTED is typically assigned to the subject of a copula construction. Correspondingly, ATTRIBUTE is the label for the role of the predicate nominal of such constructions. Given that the other roles, AGENT and PATIENT, are associated with

transitive constructions, the semantic role overview allows some direct inferences about the syntax of the constructions as well. The set of head roles within the relative clause is even more condensed. Here the PATIENT role is clearly dominant with AGENT and AFFECTED accounting for the majority (~80%) of the cases (cf. Figure 67).

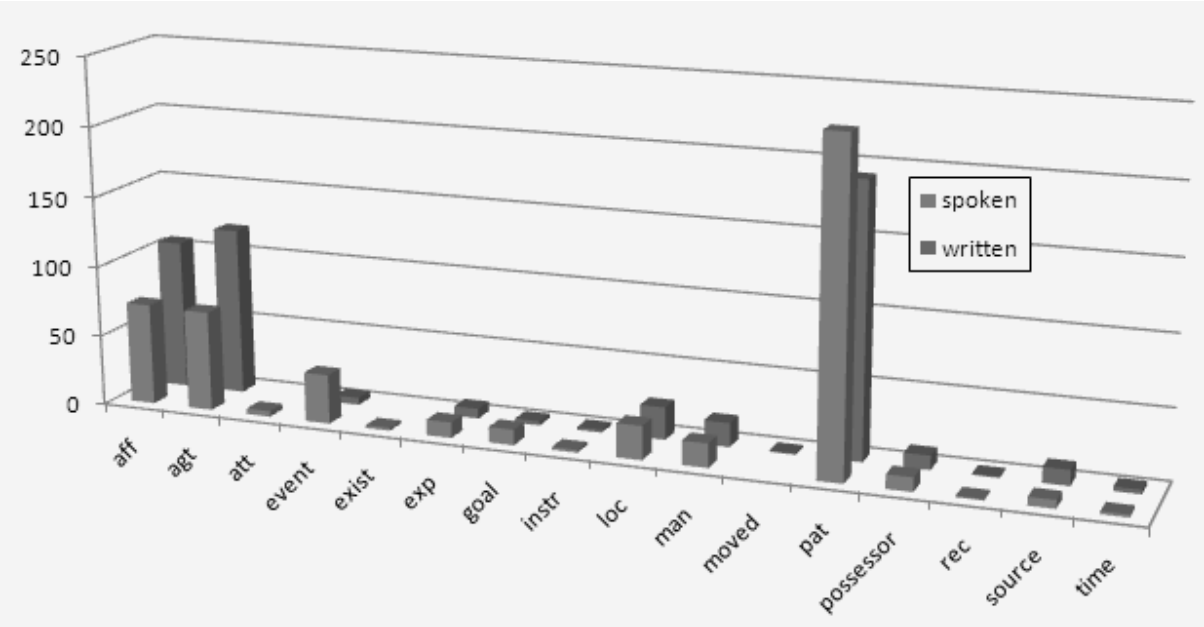

Figure 67: Thematic roles (relative clause)

The next step from here is the simplification of the table. Instead of just dropping those cases that do not belong to either of the dominant categories, the complete data set was re-coded so as to allow only three levels: PRIMARY, SECONDARY, and TERTIARY. Primary roles are those that are associated with the exudation of energy or force and to PERCEIVERS/COGNIZERS in events of perception or cognition. Secondary roles were assigned to all remaining argument roles and tertiary role are reserved for all adjunct roles.

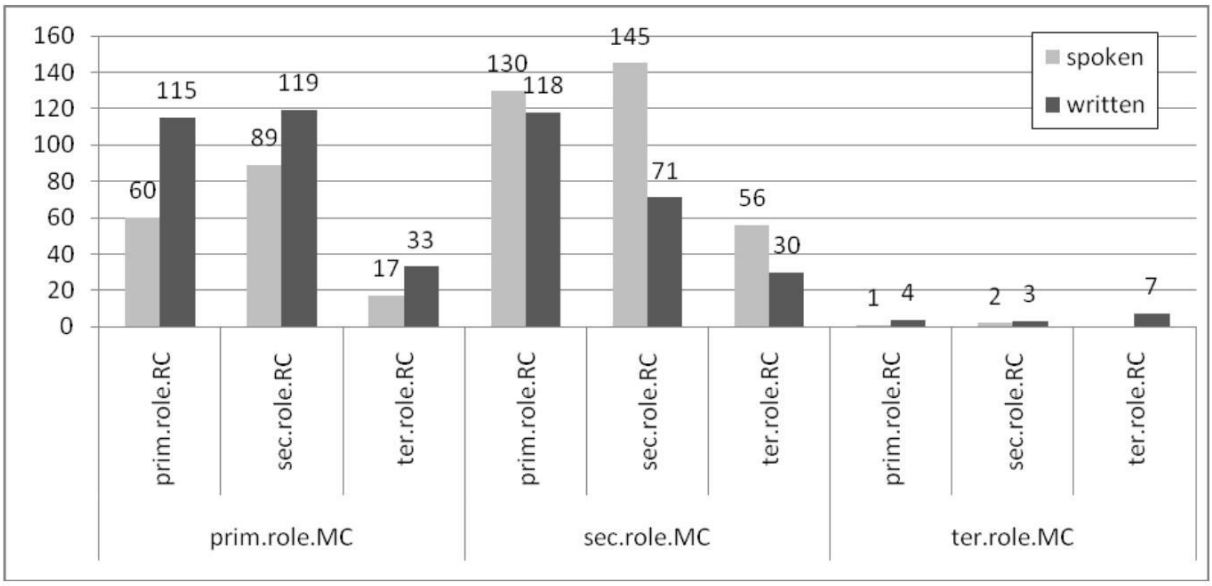

Figure 68: Semantic roles of the head in MC and RC across modalities

Table 68 shows that one can hardly speak of a tendency towards parallelism even if only three levels are distinguished. For the case that the role of the head in the main clause is primary, we observe for both modalities that a tertiary role within the RC is rather rare. Overall the most likely scenario is one where the RC role is secondary, i.e. not parallel. This tendency is more pronounced in spoken discourse. The strongest evidence for semantic parallelism can be observed for cases where the MC role is secondary. Here, parallel role assignment is indeed dominant, at least for the spoken modality. The low token frequencies observed for tertiary roles in the main clause prevent any qualified statements on this category. A Chi-square test of independence reveals that there are also no significant associations between levels of parallelism and medium ($\chi 2 = 0.94$, df = 1, p-value > 0.33, Cramer's V=0.031).

This concludes our "search for parallelism". Neither the reduction of to be distinguished syntactic roles nor the switch to a semantic level, which controlled for voice effects, led to results that could plausibly be interpreted as evidence for the parallel function hypothesis. If the effect is nevertheless real, its reality certainly cannot be explained with reference to induction. That is to say that the ambient language, which in the usage-based view provides sufficient information about grammar in the form of statistical regularities, cannot be the source of an alleged parallelism effect. At least not if we grant that the data

sample used here does not diverge dramatically from the general population it is supposed to represent. And given the close fit with the corpus data reported in Roland et al (2007) that could be observed for a number of RC properties (in § 3.3.4), we have no reason to doubt the representativity of the data used in this study.

3.5.4 Head versus RC- Subject: Interference and discourse-function

We have now reached the final section of this chapter, in which we will investigate the role of the two adjacent NPs in finite non-subject RCC (head and RC subject). As indicated earlier, this section will finally present a discussion of why exactly such NP properties should be relevant for the processing of an RCC. We may thus start with a sketch of the research in this area.

Research into sentence comprehension is traditionally divided into two different phenomenological areas both of which cause processing difficulties. Monitoring the way people process such problematic cases allows researchers to draw inferences about the underlying processing architecture. One of these areas involves (local) syntactic ambiguities, which raise the level of uncertainty about the structure currently being processed, which in turn may add to the difficulty of the pattern. The other focuses on the role of the intrinsic structural complexity of unambiguous sentences. A great variety of model types has been employed in the attempt to assess which knowledge sources are used when (Crocker et al. 2000 presents an excellent overview spelling out all major dimensions of contrasts along which different models may differ).

Recent years have witnessed are growing interest in the role the characteristics of the NP constituents have on the processing of such structures (Bever 1974, Gibson 1998, Gordon et al. 2001, 2004, Mak et al. 2002, Traxler et al. 2002, Warren and Gibson 2002). This research has disclosed that processing ease varies systematically with certain properties of the referring expression, i.e. the types of NPs used in the structures. Bever (1974) observed that sentences like (67) are easier than sentences in (68) even though their syntactic structures are arguably identical.

(67) The reporter **the politician the commentator** met trusts said the president won't resign.

(68) The reporter **everyone I** met trusts said that the president won't resign

Both sentences exhibit double center embedding and so the observed differences in processing difficulties have been attributed to the linguistic realization of the referring expressions in bold print. The relevant properties are both formal (e.g. lexical vs. pronominal NP) and semantic/conceptual in nature (e.g. animate vs. inanimate referent) and so by studying the processing of such structures, researchers have hoped to learn something about the types of representations that are present in working memory during syntactic processing.

While the empirical finding that the NP types modulate processing difficulty is quite robust, the theoretical treatment is somewhat controversial. In terms of psychological theorizing, most accounts are committed to the belief that there is a resource limited working memory, which as we have seen is not a necessary component of a language comprehension model. On the linguistic side, these accounts assume that the structures in (67) and (68) differ only in their lexical choices. While the relationship between psychological theories of language and linguistic theories surely is an intricate one, it is also by necessity asymmetric. Linguistic theories may or may not aim at psychological plausibility (cf. Gazdar et al. 1985), but psychological theories have to make some assumptions about linguistic categories, especially if it makes reference to a capacity limited storage device. If this storage device is conceived of as being capable of storing up to *k* units, we need to know what exactly counts as a unit in order to test the predictions of a given model. As we have discussed in § 2.1, the traditional bi-partite distinction between syntax and lexis has been questioned in recent linguistic theorizing and has in fact been abandoned in constructionist treatments of grammar. These accounts—as we have seen—assume all of language to be symbolic and postulate a continuum of symbolic units of varying degrees of complexity and schematicity. So, assumptions on the linguistic side have straightforward repercussions on psychological models particularly when it comes to the number and type of units that are assumed to be instantiated in a given structure and thus have to be accessed in the processing of that

structure. The examples in (67) and (68) are well suited to illustrate a crucial the difference that follows from the linguistic commitments. From a constructionist perspective it is possible to account for the relative ease of (68) by treating the string *everyone I* VP as an instance of a highly salient unit (or construction), which—qua being a unit—can be accessed as a whole and which thus does not require any serious amount of structural processing. We can easily find evidence for a difference in representational status of the stings in question (unit vs. composite structure) by submitting them to a quick-and-dirty Google search. The string *the politician the commentator* yields exactly two hits (both of which refer to psycholinguistic studies in which the sentence was used). Now, the frequency of the string *everyone I* is capped at 19,000,000 occurrences. So, the string *everyone I* is at least 9.5 million times more frequent in the data underlying the Google search. While these numbers certainly need to be taken with a considerable amount of caution, they certainly suggest that there are enormous differences in the way the strings are treated in the mental grammar (cf. § 2.2.1).

The basic point is that a more traditional view on grammar would assume that the number of units in (67) exceeds that of (68) by exactly two (the occurrences of the definite article), while from a construction grammar perspective the number of relevant processing units may differ from that assessment depending on the number of prefabricated chunks that are present in the structure. In this view, it is imperative to identify these intermediate units and assess of their degrees of entrenchment.

This issue can be extended to other types of RCCs and their processing, specifically to those that have an embedded non-subject relative clause so that the head is (immediately) followed by the subject of the RC.

With these theoretical prerequisites in place, let us have a closer look at the scenario in which the processing system faces two consecutive NPs that need to be integrated into the current discourse model, i.e. the head NP and the NP that constitutes the RC subject. Consider the following examples.

(69) This is [someone|the teacher] that [you|the student] likes so much.

(70) This is [something|a product] that [it|the company] sells.

The RCCs in (69) and (70) exemplify some of the formal variation that we observe with respect to the morphosyntactic (pronominal|lexical) and semantic (animate|inanimate) properties of the two NP in brackets, i.e. the head and the subject of the relative clause. It has been suggested in the psycholinguistic/experimental literature (e.g. Gordon et al. 2001) that language users have problems with RCC that exhibit two similar NPs in these positions. This similarity can be morphosyntactic, i.e. both NPs are either pronominal or lexical (Gibson 1998, or semantic both NPs have either animate or inanimate referents (e.g. Traxler et al. 2002, Make et al. 2002, Warren and Gibson 2002), or both (Gordon et al 2004).

One typical finding with respect to animacy is that object relatives are easier when the head is inanimate. This result is expected if we assume that humans exploit statistical regularities in the ambient language to build up structural expectations. Inanimate objects are more likely to be patients and patients are more likely to occupy an object position in English. So let us have a look at finite non-subject RCs and specifically the morphosyntactic forms and animacy values of the RC subject NPs. The analysis is based on a total of 329 observations of such constructions, which is to say that about a third of all RCC under investigation meets the constraints {[+finite] & [-subject relative]}. We should note right away that these 329 examples are not evenly distributed across the two registers. The construction is far more frequent in spoken discourse (n_{spoken}=230, $n_{written}$=99). Figure 69 presents an overview.

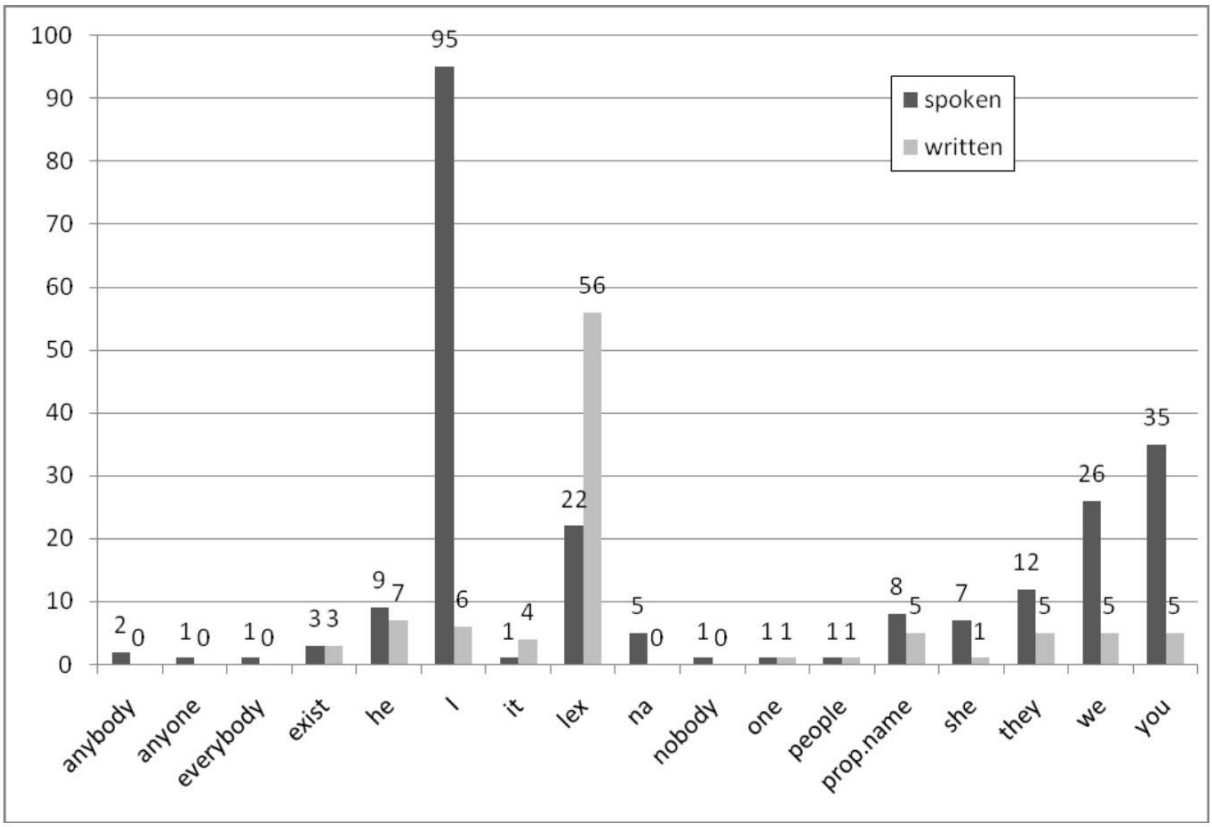

Figure 69: Lexical realization of subject of finite non-subject RC (n=329)

Despite the fact that the number of spoken constructions outnumbers that of the written modality by more than 2:1, we can already read off certain modality specific tendencies. The frequency difference of *I* as the subject of the RC, for instance, certainly cannot be attributed to the greater number of spoken cases. Conversely, the large number of lexical NP in written language certainly suggests the presence of a statistically meaningful effect. With this first orientation in place, we may turn to a more detailed characterization of the subject NP. The characterization will focus on the factors ANIMACY and MORPHOSYNTACTIC REALIZATION. Even though certain additional (potentially relevant) variables most notably DEFINITENESS of the NP or CONCRETENESS of the respective referents have been investigated, they were excluded from the discussion as they did not add to the general patterning of the findings. Figure 70 presents the possible scenarios at this level of description.

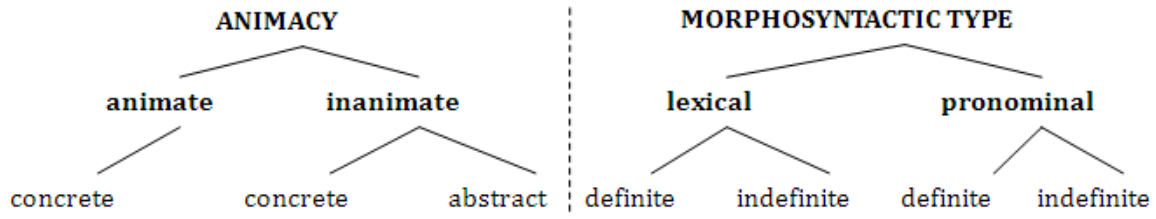

Figure 70: Granularity of description

The morphosyntactic contrast (lexical|pronominal) incorporates differences at the level of definiteness. Similarly, the semantic difference captured by the variable concreteness is incorporated in the next higher lever that distinguishes animate from inanimate objects. Further distinctions at lower levels turned out to be unnecessary.

The first factor of interest concerns the animacy of the .referent of the subject NP. Figure 71 presents the findings across modalities.

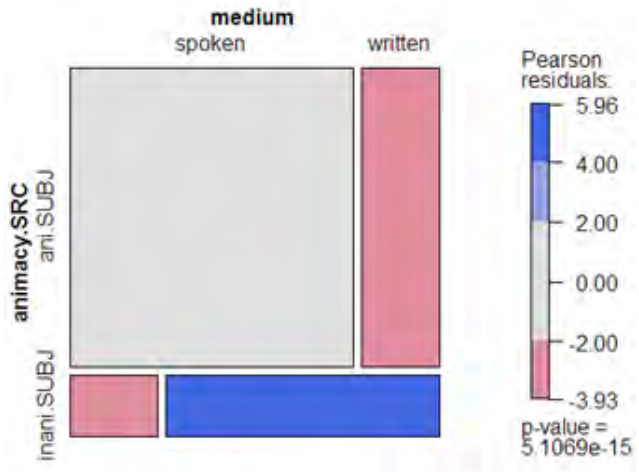

Figure 71: Animacy of RC subjects across modalities

The overall distribution is far from what is expected under the assumption of statistical independence ($\chi 2 = 61.219$, $p_{Pearson} < 5.1$ e-15; df = 1, Cramer's Phi = 0.44). We observe a strong bias towards inanimate subjects in written discourse (as indicated by the deep blue coloring). For spoken language we observe that animate subjects far outnumber inanimate ones, but the co-occurrence frequency is not beyond our statistical expectations.

The distributional differences for the morphosyntactic realization are even more pronounced. There is a strong preference for pronominal subjects in spoken language, while

written language shows a strong preference for lexical subject NPs ($\chi 2 = 86.121$, $p < 2.22$ e-16, df = 1, Cramer's Phi = 0.52). Figure 72 presents the corresponding mosaic plot.

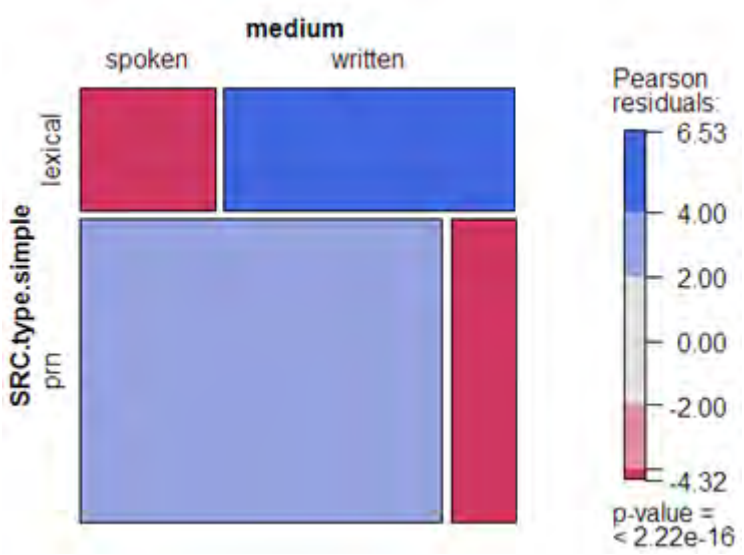

Figure 72: Morphosyntactic realization of RC subjects across modalities

Again, a plausible explanation for both of these findings is to be found at the level of information structure. Fox and Thompson (1990) have taken a discourse functional stance to the grammar of relative clauses and have tried to explain typical RC patterning with reference to the interlocutors' state of knowledge. They argued that pronominal subject NPs with animate referents are often used in (spoken) relative clauses as *anchors*, which allow a new entity—the referent of the head NP—to be linked to an entity that is highly salient in the context of utterance, usually the speaker or the addressee.

Now that we have characterized typical RC subjects, we may include the head into the picture so that we can assess preferred construction types. Recall that we have introduced the idea that the processing demand of an RCC increases when the two consecutive NPs in these constructions are similar in their syntax and/or semantics. Gordon and colleagues propose that this effect is due to similarity-based interferences. They hypothesize that the human processing system has problems with the organization of a set of units that are very similar to each other. If we reason from processing difficulty to expected frequency, so that easier patterns would be used more extensively, we would expect to find a greater number of dissimilar NPs than similar ones. If, however, we reason form frequency to processing

difficulty, we would expect discourse-functionally useful tools to be frequent and these frequent types would be predicted to be easy. Figures 73 and 74 present the attested factor level combinations of the variables morphosyntactic type and animacy respectively.

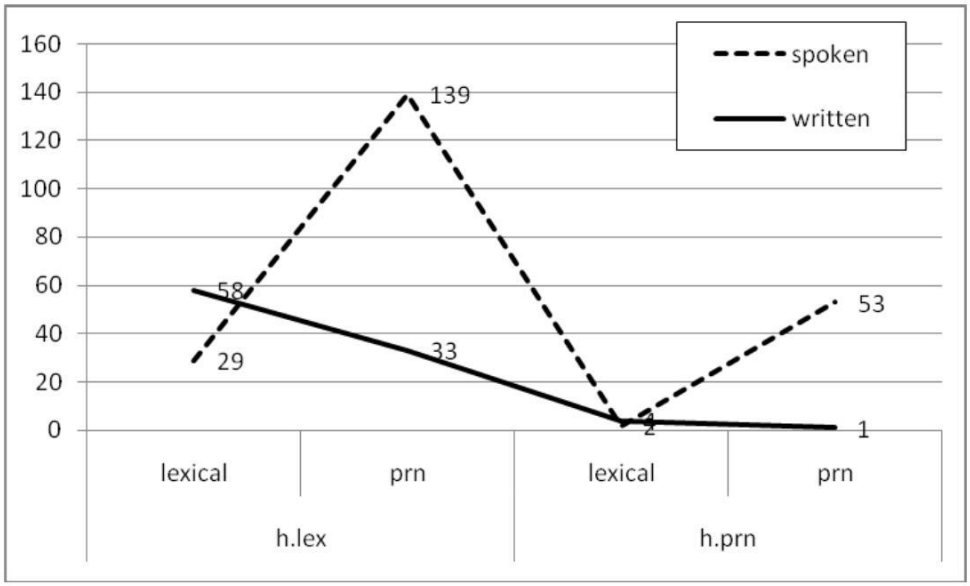

Figure 73: Register comparison :RC subject and head (morphosyntactic types)

For the morphosyntactic factor and spoken language, we observe that lexical heads combined with pronominal subjects are very frequent. This finding is consistent with both hypotheses that relate processing ease to and frequency (processing ease of E ⇒ frequency of E; frequency of E ⇒ processing ease of E). Notice that the interference hypothesis (if combined with frequency assumptions) not only predicts a high number of lexical heads followed by pronominal subjects (as these are dissimilar), but also a high number of pronominal heads and lexical subjects (as these are just as dissimilar). This latter tendency, however, is not borne out by the data. Pronominal subject are preferred even if the head is also pronominal. In contrast, the competing hypothesis, which mediates frequency and processing ease via discourse function, is fully compatible with the observed distribution. Non-subject RCs with a pronominal subject are useful tools to anchor new referents in the discourse. The low frequency of lexical-lexical configurations in this view is due to the fact that such patterns are not particularly useful. Lexical NPs indicate lower degrees of givenness/accessibility and are thus not well suited to help integrate new entities into the discourse (model).

For the written modality, we observe a straight downturn in frequency from "lexical (head)-lexical (subject)", to "lexical-pronominal", to "pronominal-lexical", to "pronominal-pronominal". As written language is generally less dependent on processing factors, we must be prepared to observe mismatches with our processing hypotheses. But the fact that "lexical-lexical" is the preferred type in that modality certainly is not easily accounted for by the interference hypothesis.

For the semantic factor, animacy of the NP referent, we observe similar results. Spoken language is clearly dominated by inanimate heads followed by animate RC subjects. This finding is predicted by both hypotheses. The overall distribution, however is not easily explained with reference to interference only, because the other dissimilar combination, i.e. animate-head and inanimate subject, is the least frequent type. This general preference for animate subject is predicted by the discourse-functional approach, because animate things are better anchors than inanimate entities.

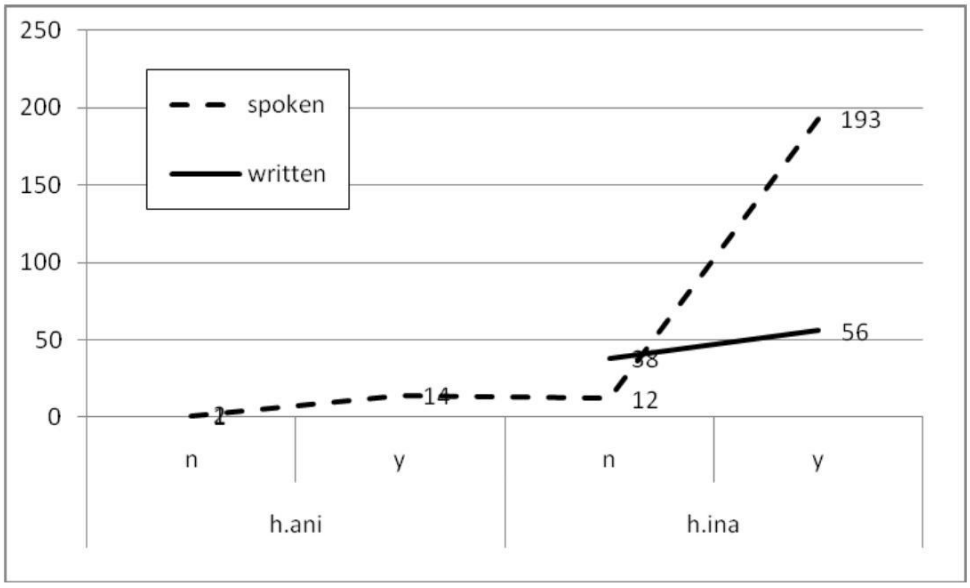

Figure 74: Register comparison: RC subject and head (animacy)

As the properties animacy and morphosyntactic realization are logically independent, we may help ourselves to a better overview of what is typical or atypical across registers by crossing the intersections of these variables. Figure 75 presents an overview of the frequencies of the resulting patterns across modalities.

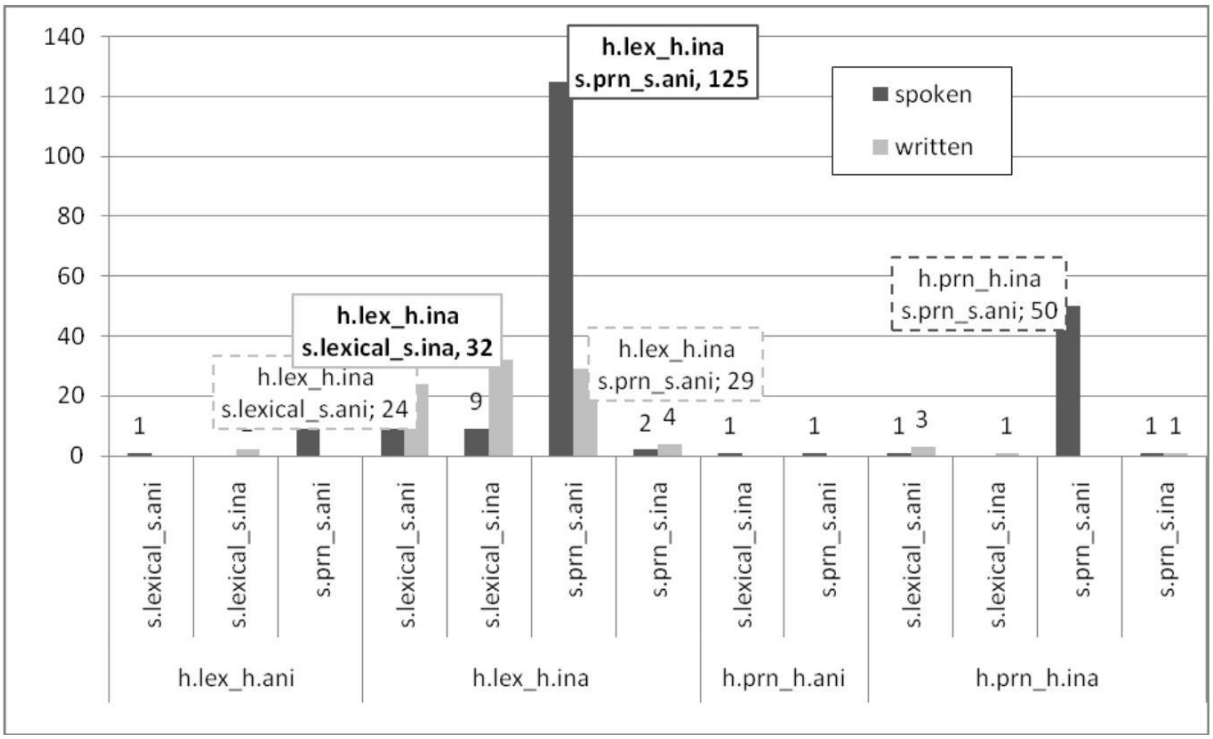

Figure 75: Heads and RC subjects (overview)

Looking at this overview we observe that the by far most frequent type in spoken language is "h.lex_h.ina ↷ s.prn_s.ani", which corresponds to a lexical head referring to an inanimate entity is followed by a pronominal RC subject referring to an animate entity. There is another frequent pattern, "h.prn_h.ina ↷ s.prn_s.ani", in which the same type of RC subject follows a pronominal head, which too refers to an inanimate entity. All remaining patterns exhibit considerably lower frequencies. From a discourse-functional perspective this patterning is very much expected. The fact that in both dominant patterns the RC further describes an inanimate entity can be explained by the fact that inanimate entities are more likely to be new (i.e. not given or less accessible) in the discourse because inanimate entities cannot possibly participate in a discourse and the most accessible participants are always the interlocutors themselves (i.e. speaker and hearer). So, in the present classification the most frequent head type is the one in strongest need of being anchored. This anchoring is best achieved by relating the new entity to a highly accessible (given) entity and pronouns signal such high accessibility values.

In written language, the situation is a little different (as usual). We observe that the

attested cases distribute more even across different types. The types that incorporate most data points all exhibit a lexical head that refers to an inanimate entity. The strong dominance of this head type is shown in Tables 13 and 14, which present the distributions in terms of proportions separately for each register.

Table 13: Percentages of "head ↷ RC subject"-types (spoken language)

SPOKEN		RC subject				
		s.lexical & s.ani	s.lexical & s.ina	s.prn & s.ani	s.prn & s.ina	TOTAL
head	h.lex & h.ani	0.45	0.00	5.45	0.00	5.91
	h.lex & h.ina	**7.73**	**4.09**	**56.82**	**0.91**	**69.55**
	h.prn & h.ani	0.00	0.45	0.45	0.00	0.91
	h.prn & h.ina	0.45	0.00	22.73	0.45	23.64
	TOTAL	8.64	4.55	85.45	1.36	100.00

Table 14: Percentages of "head ↷ RC subject"-types (written language)

WRITTEN		RC subject				
		s.lexical & s.ani	s.lexical & s.ina	s.prn & s.ani	s.prn & s.ina	TOTAL
head	h.lex & h.ani	0.00	2.08	0.00	0.00	2.08
	h.lex & h.ina	**25.00**	**33.33**	**30.21**	**4.17**	**92.71**
	h.prn & h.ina	3.13	1.04	0.00	1.04	5.21
	TOTAL	28.13	36.46	30.21	5.21	100.00

While a lexical inanimate head is present of roughly 70% of the spoken cases, we find it in over 90% of the cases in the written modality. When we translate these proportions back to the underlying frequencies, we can submit the highlighted lines to a contrastive analysis. A chi-square test of independence, which evaluates the relationship between the modality of the lexical inanimate head and the type of RC subject, yields highly significant results ($\chi 2$ = 62.021, df = 3, p< 2.17e-13, Cramer's V = 0.506). Figure 76 presents the corresponding association plot, which allows us to allocate the sources of the departure from independence.

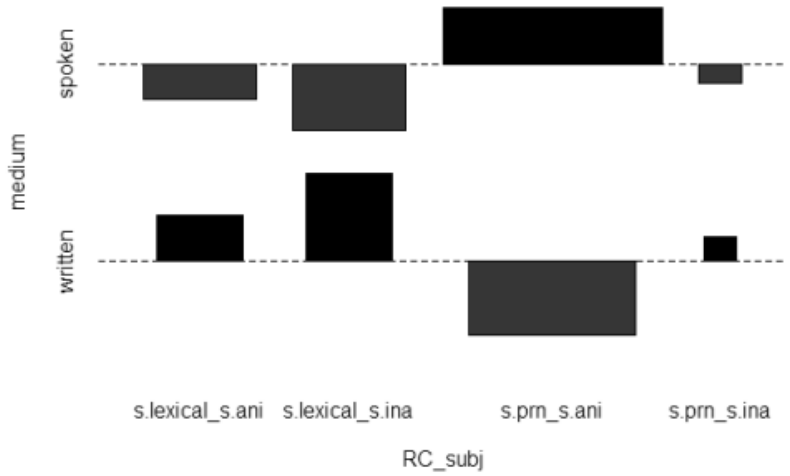

Figure 76: Lexical inanimate heads and their RC subjects across modalities

In summation: we observe that written language and spoken language differ significantly in their patterning. Written language preferably utilizes lexical RC subjects. The high proportion of inanimate NP in that function indicates that the function of the RC in these constructions is not likely to be that of grounding or anchoring a new referent into the discourse because inanimate lexical NPs are not good anchors. In spoken language, however, the anchoring/grounding function of RC presents a plausible explanation for the frequency distributions.

4 Expanding horizons: RCC in ambient configuration space

Having provided a general overview of how some crucial parameters relevant for the characterization of English RCCs distribute across the two modalities, this chapter will present a widening in scope and lead our discussion to a multivariate point of view. With this change of perspective we will also approach a central technical goal of the present study, namely to detect highly entrenched RCC schemas, i.e. complex patterns whose elaboration sites are instantiated by constituents that (co-)occur frequently enough to suggest that the overall pattern approaches unit status. In the attempt to detect such routinized RCC schemas we will make use of two statistical techniques both of which aim at finding (unanticipated) patterns implicit in complex data sets. The first procedure to be described, a association rule mining (ARM) techniques, is borrowed from *data mining* and *knowledge discovery*, i.e. the area of inquiry concerned with sorting through large amounts of data in the attempt to pick out relevant information. Association rule mining techniques are generally geared to detect regularities in the data set, which are then disclosed in the form of probabilistic (inexact/fuzzy) rules. The second technique to be used, configural frequency analysis (CFA), typifies a different approach to a similar class of problems and aims at identifying those factor level combinations (configurations) of a complex contingency table that are statistically special.

Our first area of application for these techniques of pattern detection is the class of non-finite RCCs. As indicated earlier the focus of this study clearly lies on finite RCCs and so the analysis of the non-finite types is best conceived of as a means to introduce the methodology on the basis of a data set of manageable complexity. This discussion can be viewed as a sanity test of the proposed methodology. So in order to gain some confidence in the proposed methodology, we will start with a discussion of non-finite RCCs in the transparent (bivariate) fashion employed so far expand our approach from there. So having investigated the bivariate relationships, we will first introduce, apply and discuss the results of the association rule mining technique, and then go through the same steps using the second method, i.e. configural frequency analysis. We will focus on three variables central to their

grammatical description and their (potential) distributional differences across modalities.

4.1 Non-finite RCCs (bivariate prelude)

The first things to note are a) that non-finite RCCs make up roughly 30% of the total data set (285/1000) and b) that they are more common in written language than in spoken language (written: 186/500; spoken 99/500). Our overview of their properties starts with an investigation of the following variables:

Type of non-finite RC:	*-ing* participle, *-ed* participle, *to*-infinitive
Internal syntactic role:	subject versus non-subject
Type of embedding:	right versus center
Medium:	spoken versus written

Table 15 presents a general summary of their distribution.

Table 15: Summary of non-finite RCC (n=285)

medium		**sub-type**		**internal role**		**embedding**	
spoken	99	*-ed* prt	143	non-subject	66	center	140
written	186	*-ing* prt	63	subject	219	right	145
		to-inf	79				

When we cross modality (MEDIUM) with the other three variables, we can observe first, a pronounced preferences for *-ed* participial constructions in written language, and an even stronger preference for *to*-infinitival constructions in spoken language, which necessitates a corresponding bias in terms of the RC-internal syntax: *to*-infinitival RCs tend to have an object gap as in "*the book to read _* " whereas the head always plays the role of subject in participial RCs as in "*the horse _ raced past the barn*". Figure 77 presents an overview of the frequency distribution.

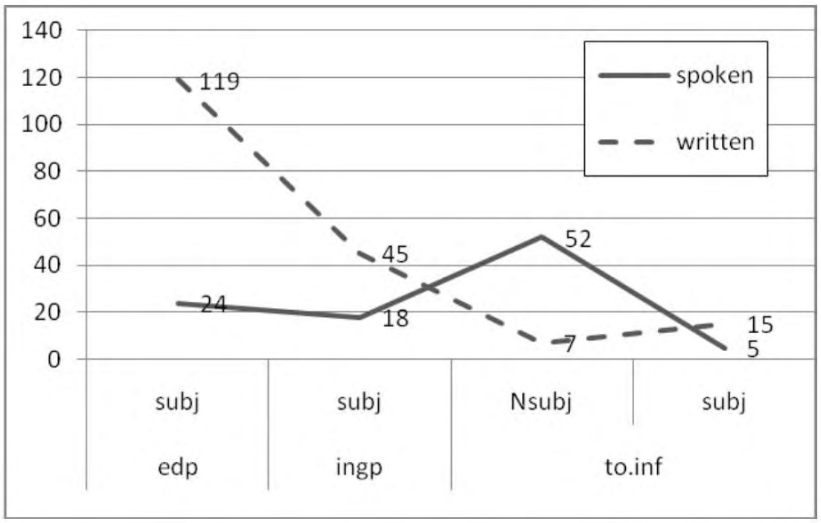

Figure 77: Internal roles of non-finite RC across modalities

At the most general level, we observe that participial RCs are more frequent than their infinitival cousins in written language, whereas the opposite is true of spoken language. The variants of participial constructions distribute quite evenly in spoken language ($n_{ed\ PTC}$ = 24; n_{ingPTC} = 18), while *–ed* participles are clearly dominant in written language ($n_{ed\ PTC}$ = 119; n_{ingPTC} = 45). A statistical analysis of the relationship between the factors FINITE.TYPE and MEDIUM reveals a statistically significant association between written language and *–ed* participial RCs and spoken language and *to*-infinitival ones ($\chi 2$ = 108.17, df=1, p < 5.79e.15, Cramer's V = 0.613). Figure 78 presents the corresponding mosaic plot (cf. also § 3.3.1).

Figure 78: Non-finite RCC: type of non-finite RC across modalities

Interestingly, there appears to be an interaction between the subtypes of *to*-infinitival RC and

modality to the effect that spoken discourse shows a greater proportion of non-subject RC. A statistical analysis reveals a significant effect of considerable size (subset: *to*-infinitival RC INTERNAL ROLE X MEDIUM: $\chi 2$ = 29.632, df = 1, p = 5.22e-08. Cramer's V = 0.612). If, however, we look at the residuals, we learn that the more surprising fact about the table—statistically speaking—is the positive correlation between written *to* infinitives and subject roles. The high number of non-subject *to*-infinitival RCs in spoken language is within the range of expectation. If we compare the distributions of subject and non-subject RCs in the complete data set with the ones observed for the subset of *to*-infinitival constructions, we discover quite striking differences between the modalities (cf. also § 3.3.3).

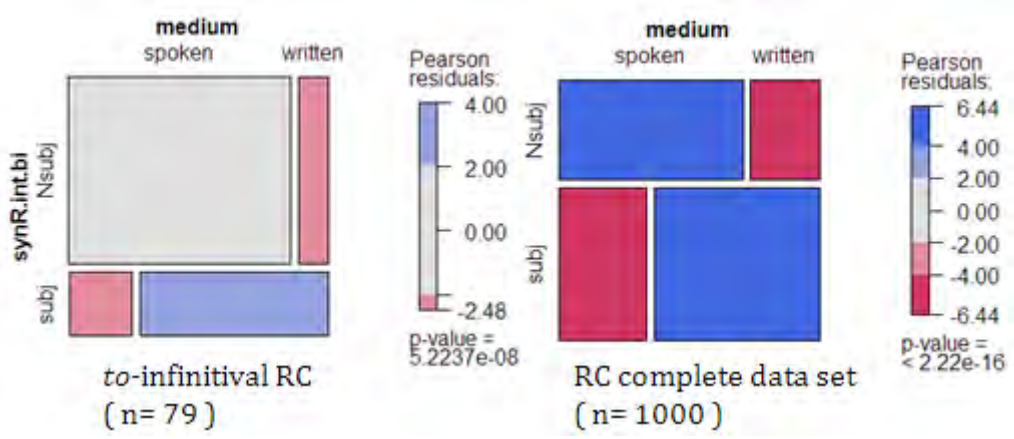

Figure 79: Internal role of *to* infinitival RC across modalities

As p-values are dependent on sample sizes, we need not worry too much about the much lower p-value of the right hand table and, correspondingly, the higher degree of saturation of the coloring. The effect size of the association is clearly more pronounced in the *to*-infinitival subset (Association$_{\text{complete data}}$: $\chi 2$ = 137.09, df = 1, p= 2.22e-16, Cramer's Phi = 0.37; Association$_{\text{to-infinitival data}}$: $\chi 2$ = 27.988, df=1, p= 5.223e-08, Cramer's Phi = 0.61).

So let us turn the third (and final variable) employed in our characterization of non-finite RC, i.e. type of embedding. We have already seen in § 3.4.2.2 that there is a strong tendency for spoken RC to occur in VP internal positions, whereas written RCs are predominantly used to modify the subject of the MC. Figure 80 presents an overview.

.

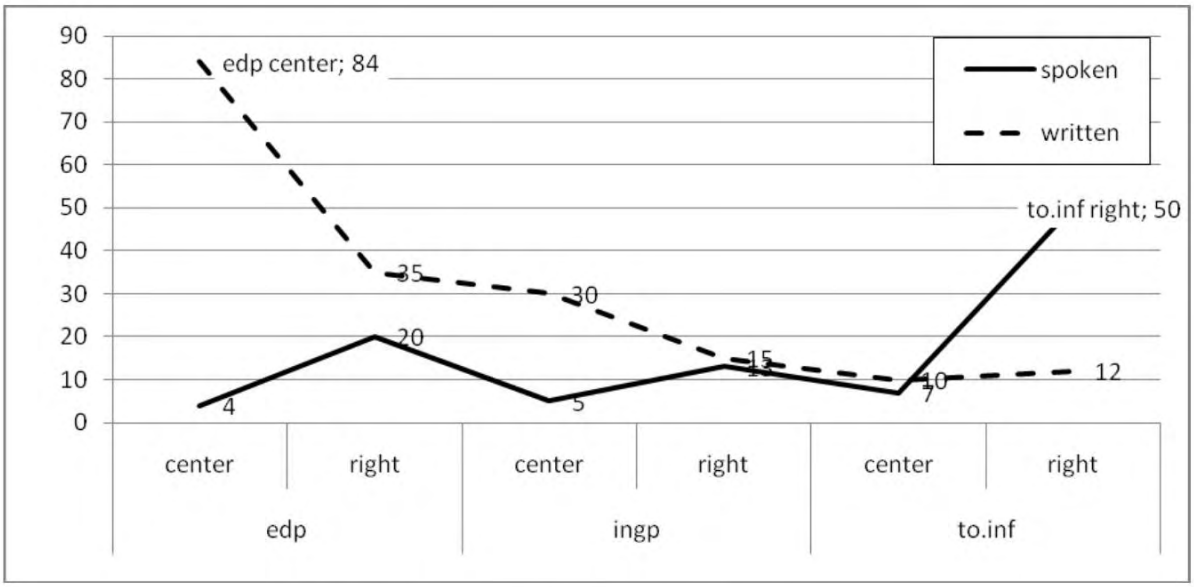

Figure 80: Embedding in non-finite RC across modalities

When we cross EMBEDDING and MEDIUM for the complete data set and the subset of non-finite RC, the picture in Figure 81 emerges.

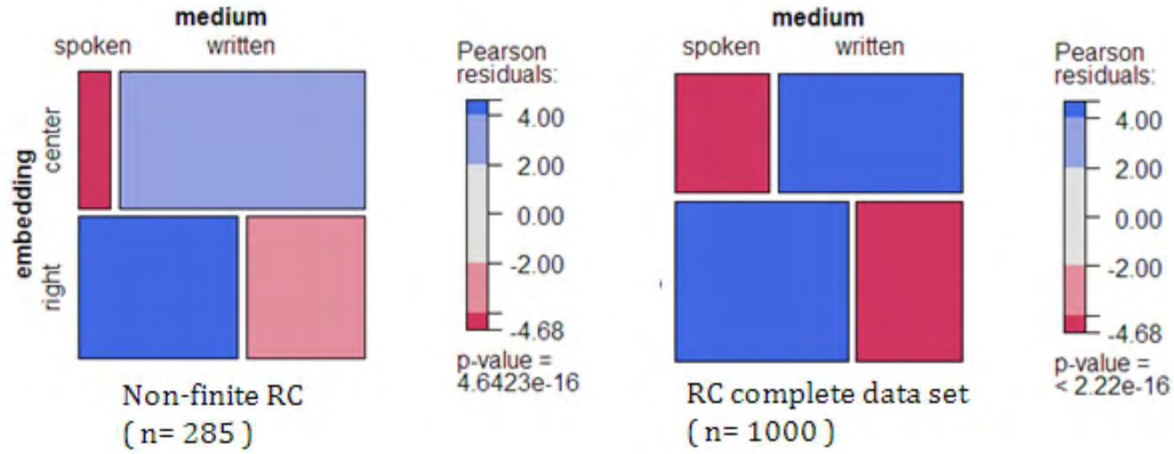

Figure 81: Type of embedding versus medium across data sets

We observe that the embedding tendencies are even more pronounced with non-finite RC subset than they are for the more general class. We obtain a much larger effect size for the non-finite types (Association$_{\text{complete data}}$: $\chi 2$ = 76.621, df = 1, p= 2.22e-16, Cramer's Phi = 0.277; Association$_{\text{non-finite data}}$: $\chi 2$ = 65.943, df = 1, p= 4.44e-16, Cramer's Phi = 0.481).

Our step-by-step bivariate analysis of the non-finite RCC revealed that it characteristic of spoken language to exhibit a tendency towards right embedded *to*-infinitval non-subject RCs, whereas written language heavily employs center embedded *–ed* participial RC, which by necessity are subject RCs. But how can we account for these observations? The results for embedding are very much in line with certain processing expectations: spoken language prefers the simpler variant, i.e. right embedding, whereas the more complex type, center embedding, is primarily used in the written register. Even though there are many possible ways to characterize the processing demand associated with the different positions that RCs can assume within the MC-structure, the literature on (syntactic) complexity and processing difficulty agrees on the idea that if overall sentence complexity is modulated by the type of embedding at all, then it is the center embedded variant that should impose greater difficulty. We can characterize the increase in complexity with respect to structure building processes (e.g. Gernsbacher 1990). The comprehension of right embedded RC should be easier because the comprehender does not have to delay the construction of the situation model corresponding to the scene described by the MC until the situation described by the RC has been completed. Approaches that focus on parsing (e.g. Hawkins 1994, 2004) make the same predictions. Again as modifications of the subject nominal (leading to center embedding) and modifications of a VP internal nominal (resulting in right embedding) are available as alternative ways to express a particular thought only in very specific circumstances (if at all), the difference in preferred embedding cannot be accounted for solely on processing grounds but again could be due to differences in communicative tools utilized in spoken and written discourse respectively (cf. § 3.4.2.2).

We may also presume that the preference of *–ed* patterns in written language and its infrequency in spoken discourse is compatible with the processing view, as passive constructions are often viewed as being harder than active ones. However, the relationship of voice and complexity is likely to be more intricate than that and there are presumably numerous interacting factors, e.g. semantic type of argument, presence of negation etc.. Also, the fact that *–ed* participle construction are more frequent in written language is consistent with the idea that complex patterns are preffered in this but not the other modality. Passive participial construction of this type may introduced local MV/RR ambiguities (cf. § 3.1.2.2),

which may be argued to impose further indeterminacy and hence processing difficulty.

We should also note that the data for SUB-TYPE OF RC and INTERNAL ROLE are somewhat problematic, when processing demand is operationalized in terms of length of the dependency domain, i.e. the distance from the head nominal to its canonical position within the RC proper. If deeper roles caused more processing difficulty, then we would expect spoken language to show a preference for participial RC and subject roles, because those result in minimal dependency domains. Processing accounts that work with memory demand that is contingent on the length of the dependency domain (e.g. Gibson 1998) thus predict the exact opposite patterning. But of course, *to*-infinitival RC and *ed*-participial RCs do not constitute a typical choice phenomenon. They can be alternative ways to express a particular meaning if and only if their respective verbs express converse predicates. The closest thing to a choice between using either a *–ed* participial RC construction or a *to*-infinitival one, requires certain adjustments within the VP of the RC. For an illustration consider the examples given in (71) and (72).

(71) The book [to buy __ (by NP_i from NP_j)] VP_{MC}

(72) The book [__ sold (by NP_j to NP_i)] VP_{MC}

But even at the level of semantic content, i.e. a truth-functional level, these two sentences can hardly be considered to mutually entail each other, if only because of their different temporal properties. As this "predicate reversal"-constraint makes it dubious to assume that speakers do in fact have a constructional choice between participial and *to*-infinitival RCs, a simple processing explanation like the one just sketched appears to be oversimplified. In cases like these, the observed distributional difference might very well be better explained from a discourse-functional level, particularly by an account focusing on typical communicative contexts and associated speech acts that are likely to differ across modalities. After all speakers do of course try to verbalize the thought they wish to express and do not produce utterances just because they are easy to produce.

This leaves us with the observation that *to*-infinitival RCs pattern differently across modalities with a preference for non-subject variants in spoken discourse. Again, this is surprising if we wish to assume that shorter dependency domains lighten the processing demand of a pattern and that frequencies mirror processing difficulties. If there is a difference in patterning, we would expect there to be a greater proportion of infinitival subject RC in spoken discourse, but the opposite is attested.

Stronger support for a processing account of the observed preferred patternings can be expected from an analysis that looks at these variables not in isolation but focuses on their co-occurrence. It is only when we look at the "bigger picture", i.e. the overall form of the RCC that we can hope to identify the discourse-functional motivation that may ultimately explain the frequency of a given pattern in a specific modality (or genre for that matter). The sections to follow will explicate how this higher order entrenched pattern recognition can be achieved.

4.2 A configural perspective on non-finite RCCs

The preceding section has shown that spoken non-finite RCCs tend to be right embedded, are predominantly *to*-infinitival, and show relativization on the subject role. In contrast, we have observed written discourse to exhibit a tendency towards center embedding and *ed*-participial RC constructions. However, as these variables are clearly (logically) independent from each other nothing in the analysis presented so far suggests that those properties occur together (center embedded & *-ed* participle). The observed preferences for both modalities may result from different cases, i.e. the overall preference of a given modality for certain values for embedding and type of RC need not be realized together but could be distributed over different examples. We shall address this issue in turn as we present the first results from our pattern-oriented perspective.

Without further ado, we may now turn to a discussion of our pattern recognition techniques, starting with the measure from data mining.

4.2.1 Association rule mining: *k*-optimal patterns analysis

Association rule mining is a popular and well-researched way of disclosing interesting

relations among a set of variables in a large dataset. Such interesting relationships are—as the name suggests—expressed as rules that relate certain properties on the basis of their degree of association. The rules assume the following general form

$$\{\text{PROPERTY}_A, \text{PROPERTY}_B \ldots, \text{PROPERTY}_K\} \Rightarrow \{\text{PROPERTY}_L, \text{PROPERTY}_M \ldots, \text{PROPERTY}_N\}$$

The notation of such rules resembles the one used in propositional calculus, i.e. there is an antecedent (or left hand side of the rule, LHS) specifying a set of conditions and a consequent (or right hand side of the rule, RHS), which associates the antecedent with another set of properties. However, the operator '$\Rightarrow$' does not denote material implication, but rather designates a probabilistic implicational relationship such that the set of properties listed on the left hand side of the rule makes likely the set of properties specified on the right hand side. In other words, the method aims at identifying rules that specify some set of properties that are associatively connected to another set of properties, hence the label *association rule mining*.

Methods like these are particularly relevant in the context of market basket analysis, or more generally in business activities for which it is interesting to know which products are frequently purchased together. Such information may then be exploited for commercial purposes. Formally, we may define the problem tackled by association rule mining in these contexts as follows: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of binary attributes called items and $D = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions called a database. Each transaction in D has a unique translation ID and contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \varnothing$ (cf. Agrawal et al. 1993). This is a typical problem in many businesses and finding a solution for it can be worth a lot of money. If a company knows for example that people who have bought some product A also often have bought some product B in the past, they can act on this knowledge and suggest to new customers who have just bought product A that they might also be interested in buying product B, which given the association between A and B often results in higher sale rates of

B. Staying in this context for a minute, we may illustrate the general approach on the basis of a simple example. Let us suppose the owner of a grocery store has collected the following information about recent transactions involving any of the four products beer, crisps, vodka, and cookies. Our fictional shopkeeper is thinking about re-arranging his products and conjectures that it would be best if items associated with one another were presented in the same area of the shop. Table 16 presents the type of data he has collected (*1* indicates purchase, *0* indicates non-purchase):

Table 16: Data base for association rule mining (example)

Case (transaction ID)	property A/ product A (beer)	property B/ product B (crisps)	property C/ product C (vodka)	property D/ product D (cookies)
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Given a data set like this—which of course ideally would include some additional cases, the task to be solved now is to detect those rules that express interesting relationships among the variables, i.e. the products. The tricky part now is to find a reliable means of obtaining this knowledge. As many different approaches are conceivable, we should not be surprised to find a rather large number of algorithms suggested in the literature. The one opted for here, *k-optimal pattern discovery*, is one of the more sophisticated procedures, whose search for interesting rules goes beyond searching for frequent patterns, i.e. patterns that occur above some pre-specified threshold level. While such a "brute force" approach may be sufficient for certain purposes, it tends to focus on rules that relate properties that have high token frequencies (and hence high joint probabilities). However, sometimes we are interested in events that—even though they are relatively rare—are strongly associated with each other (a typical example in market basket analysis is caviar and vodka, which are purchased rather infrequently, but if purchased tend to be obtained together).

We will refrain from offering the attempt of a comprehensive overview of the different approaches to association rule mining and their respective algorithms (but cf. Zhu and Davidson 2007 for an approachable introductory overview). However, we will have a look at a number of standard indices that figure in the expression of the degree of "interestingness" of a given rule.

The first concept we need to introduce is that of *support* of an itemset. The value for support of an itemset X simply is the proportion in the data set which contains the itemset.

For our example in Table 16 this means that the set {beer, crisps} has a support of 2/5 = 0.4, i.e. it occurs in 40% all transactions. From this concept of support, we can derive more informative properties, starting with a property termed *confidence*. The confidence of a rule is defined as follows:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support } X}$$

So, if X is {beer, crisps} and Y is {vodka} we get for the example given in Table 17 a value for the confidence of the rule {beer, crisps} ⇒ {vodka} of 0.2/0.4 = 0.5, which means that for the transactions containing beer and crisps the rule is correct 50% of the time.[18] Hence, confidence can be interpreted as the conditional probability Pr($Y|X$).

The second index of interest—slightly more sophisticated—is the "lift" of a rule, which is the ratio of observed confidence to that ratio expected by chance or:

[18] Sometimes a distinction is made between support and coverage. If this distinction is made, *coverage* is derived from the number of cases that satisfy the LHS, whereas support pertains to cases that satisfy the LHS and the RHS of the rule. To prevent any kind of confusion, I will always present results in an explicit disambiguated fashion.

$$\text{lift}\ (X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}\, X * \text{support}\, Y}$$

If applied to our example, i.e. the rule {beer, crisps} ⇒ {vodka}, we get 0.2/ 0.4 * 0.4 = 1.25.

The "strength" of a rule is the proportion of examples covered by the LHS that are also covered by the RHS. Hence, it indicates the probability that a case will satisfy the RHS if it satisfies the LHS. For example, suppose that the LHS covers 20 examples and the RHS covers 5 of these examples. The strength then is 5/20 = 0.25. Hence, the closer the value for strength approximates 1, the stronger the rule.

Our last—and most important—concept is the *leverage* of a rule. Leverage measures the number of additional cases/transactions that an interaction involves above and beyond those that should be expected if one assumes statistical independence. This is easier to understand if we represent it schematically as:

$$\text{leverage}\ (X \Rightarrow Y) = \text{support}\ (X \Rightarrow \text{Y}) - \text{support}\ (X) * \text{support}\ (Y)$$

Throughout all ARM applications, we will—in our discussions of the results—focus on the leverage value of a rule because it may be considered closest to "the ultimate measure of interest to the user such as the magnitude of the profit" (Webb & Zhang 2005:36) or—mutatis mutandis—the degree of entrenchment of a particular regularity. High leverage implies high support and—contrary to optimizing lift—optimizing leverage guarantees a certain minimum support.

The association rule mining technique used here, *k-optimal pattern analysis*, uses the OPUS algorithm (**O**ptimized **P**runing for **U**nordered **S**earch; Webb 1995, 2000), which is particularly effective in cases, where the amount of data is limited so that a "brute force" approach of looking for the most frequent patterns is not likely to deliver meaningful results.

For a discussion of how the *k*-optimal pattern approach handles the usual problems of data mining techniques (false discoveries, spurious rules, redundant rules) see Webb (2000, 2006).

We may now turn to the application of the technique to our first linguistic domain, i.e. the domain of non-finite RCCs. For this first application of both the ARM and the CFA techniques, we will have a look at the variables presented in the preceding section (namely EMBEDDING, INTERNAL ROLE, TYPE OF NON-FINITE RC, and MEDIUM) and—to make matters a little more interesting—add to those factors two variables that also contribute to the overall complexity of a construction, namely the valency of the main clause and the relative clause, respectively.[19] All conditions were measured as binary variables, except TYPE OF NON-FINITE RC, which was treated as a three level factor. These rather coarse-grained contrasts were used to ensure that the data complexity does not overstress the capacities of the pattern detection procedures, especially those of the CFA. As a rule of thumb, the appropriate sample size for a CFA should be $N \geq 5 * 2^d$, with d being the number of dimensions and assuming that all dimensions are expressed as two-level factors (cf. Krauth & Lienert 1995:34). Given the number of factors here, we should hence aim at a sample size of n = 350. We have at our disposal a data set consisting of 285 cases, so the recommended sample size is not met exactly. However, firstly, the difference between suggested and actual sample size is not too large and secondly—and more importantly—small sample sizes will first and foremost affect our ability to detect so called *antitypes*, i.e. pattern that occur significantly less often than we would expect (cf. § 4.2.2). However, as our goal is the detection of patterns that occur with above chance-level frequencies, we may stretch the limits of the techniques and see what we can achieve on the basis of the available data. Basically, we can help ourselves to the assumption that those configurations that we can detect using the data we have at our disposal certainly do constitute interesting patterns. There greater the mismatch between the recommended and the actual sample size, the harder the detection of significant patterns (and hence the greater likelihood of missing some interesting configurations). The ARM technique

[19] All CFA calculations were conducted using *hcfa 3.2,* a script for R for Windows®. All ARM computations were conducted with *Magnum Opus 3.0* (demo version).

focuses on the detection of surprisingly frequent patterns and is a lot less restricted with respect to required sample sizes. We will nevertheless use the same data for all procedure so as to ensure maximal comparability of their respective results/outputs.

4.2.1.1 Method & Results: Association rule mining

The complete subset of non-subject RCCs (N= 285) was submitted to the analysis. The dataset to be mined consists of descriptions of the respective usage events with the following six factors used in the characterization.

	VARIABLE	FACTOR LEVELS
(1)	**Type of non-finite RC**:	*-ing* participle, *-ed* participle, *to*-infinitive
(2)	**Internal syntactic role**:	subject versus non-subject
(3)	**Type of embedding**:	right versus center
(4)	**Medium**:	spoken versus written
(5)	**Valency of main clause predicate**	1 argument versus 2 arguments
(6)	**Valency of RC predicate**	1 argument versus 2 arguments

This gives us a type of data matrix very similar to the one presented as Table 16, which is insinuated in Table 17. The column labeled "add" specifies the location of the corresponding RCC in ICE-GB R2.

Table 17: ARM type of input data

case	add	finite.type	synR.int.bi	embedding	medium	valency.MC	valency.RC
1	<ICE.GB:S1A.009#085:1:B>	ingp	subj	right	spoken	arg2	arg2
2	<ICE.GB:S1A.009#201:1:B>	edp	subj	right	spoken	arg1	arg1
3	<ICE.GB:S1A.010#231:1:A>	to.inf	Nsubj	right	spoken	arg1	arg1
4	<ICE.GB:S1A.012#082:1:E>	to.inf	Nsubj	right	spoken	arg1	arg2
5	<ICE.GB:S1A.012#242:1:A>	to.inf	Nsubj	right	spoken	arg1	arg1
6	<ICE.GB:S1A.018#136:1:A>	to.inf	Nsubj	right	spoken	arg2	arg2
7	<ICE.GB:S1A.019#084:1:A>	edp	Nsubj	right	spoken	arg2	arg2
8	<ICE.GB:S1A.019#084:1:A>	to.inf	Nsubj	right	spoken	arg2	arg2
9	<ICE.GB:S1A.023#365:1:B>	edp	Nsubj	right	spoken	arg2	arg1
...	...	...	...	...	...	...	...
285	<ICE.GB:S1A.029#202:1:B>	edp	subj	right	spoken	arg2	arg2

The association rule procedure was set to search for rules by leverage and filter out rules that

are insignificant at a critical value of 0.05. The maximum number of attributes permitted on the LHS was set to 13—corresponding to all available factor levels except those defining modality—and the permitted values on the RHS were restricted to include only those factor levels specifying the value for modality, i.e. SPOKEN and WRITTEN. All remaining options were left at their respective defaults so that the following values were used:

MINIMUM LEVERAGE =	-1.0
MINIMUM COVERAGE =	0.0
MINIMUM COVERAGE COUNT =	1
MINIMUM SUPPORT =	0.0
MINIMUM SUPPORT COUNT =	0
MINIMUM LIFT =	0.0
MINIMUM STRENGTH =	0.0

The search for association rules detected 19 rules that satisfy the specified constraints. For purposes of exposition, only the most interesting rule—as expressed in terms of leverage—will be presented here in its explicit form. For each detected rule, we get the following type of result:

DETECTED RULE 1: {SUBJECT & CENTER } ⇒ {WRITTEN}

SUBJECT & CENTER	# properties used in the description available for LHS inclusion
are associated with WRITTEN	# one of two possible values for RHS
with STRENGTH = 0.945	# strength of the rule
COVERAGE = 0.446:	# 127 cases satisfy the LHS
SUPPORT = 0.421:	# 120 cases satisfy both the LHS and the RHS
LIFT 1.45:	# the strength is 1.45 times greater than the strength if there were no association
LEVERAGE = 0.1302	# the support is 0.1302 (37.1 cases) greater than if there were no association

Table 18 presents an overview of the 19 rules found ordered by leverage. The complete output containing all the information given for Rule 1 can be found in the appendix).

Table 18: Significant rules detected for non-finite RCCs

Left Hand Side	Right Hand Side	Leverage	Strength
subject role & center	Written	0.1302	0.945
non-subject role	Written	0.1231	0.879
subject role	Written	0.1231	0.813
subject role & arg2	Written	0.1192	0.860
right embedding	Spoken	0.1145	0.572
center embedding	Written	0.1145	0.886
subject & arg2 & center	Written	0.1136	0.980
right & nonSUBJ	Spoken	0.1073	0.925
to.inf	Spoken	0.1037	0.722
subject & *ed*.p	Written	0.1003	0.861
right & to.inf	Spoken	0.0999	0.806
subject & *ed*.p & center	Written	0.0966	0.976
ed.p & center	Written	0.0932	0.955
ed.p & arg2	Written	0.0901	0.867
ed.participle	Written	0.0901	0.832
arg2MC & nonSUBJ[20]	Spoken	0.0675	0.968
arg2MC & center	Written	0.0673	0.967
arg2	Written	0.0256	0.687
arg1	Spoken	0.0256	0.446

[20] The expression *arg2MC* indicates that the main clause of the RCC is bi-valent. In case there is no *–MC* suffix in the term, we are dealing with a feature of the RC proper.

4.2.1.2 Discussion: Association rule mining

The first thing to note is that most rules specify conditions that pertain to the written medium. This is certainly at least partially caused by the contingent fact that the written non-subject RCCs outnumber the spoken ones almost 2:1 (186:99, see above). As p-values are dependent on samples sizes, an associative relationship must be a lot more pronounced with small sample sizes in order to be judged significant. So, given the limitations of our data set, we need to accept the fact that we cannot detect all interesting relationships true of the population. This caveat extends to all phenomena of interest in this study. However, we can find comfort in the fact that whatever gets identified certainly is an interesting relationship and as such most noteworthy. Another thing that we can immediately read off Table 19 is that the judgments made by the ARM technique are compatible with the results we have obtained from our bivariate investigations, which rested on a chi-square technique. That is to say that the ARM technique did recognize the same positive associative relationships, viz. the preferences of written discourse towards a) reduced relatives (*–ed* participle forms), b) internal subject roles, and c) center embedding. In contrast, spoken discourse is associated with a) *to*-infinitival RCs, b) non-subject roles, and c) right embedding. So, even though these results may not be perceived as being particularly surprising at this point, they are still pleasing as they document the positive outcome of our sanity test of the method. In regard of the new variables concerning the valency (or arity) of the predicates of the respective clausal constituents, the ARM technique associates more complex RCs (arg2) with written discourse and, conversely, less complex RC variants (arg1) with spoken discourse. No statistically meaningful claim is made with respect to valency differences of the main clause across modalities (at least not for the case of the LHS consisting of a set with just a single member). At the next higher level of LHS complexity (two properties), we observe that written language gets associated with the feature combination {subject role, center embedded}, which in fact is the most interesting rule (=highest leverage value). We also obtain the rules {subject role, arg2}, {subject role, *-ed* participle}, {*-ed* participle, center embedded}, {*-ed* participle, arg2}, and {arg2 MC, center embedded}. As suggested by these results—though by no means necessary—we also find a yet more complex rule, taking us from the triple {subject role, *-ed* participle, center embedded} to {written}. The remaining three rules pertain

to spoken language and do not exceed the complexity of a two-property LHS. Two of them make reference to the preferred type of embedding, i.e. {right embedded, *to*-infinitival} and {right embedded, non-subject role}, which is not particularly surprising given the strong correlation between the properties of being *to*-infinitival and serving a non-subject role. Notice, however, that the rule {*to*-infinitival, non-subject role} ⇒ {written} is not significant and thus not part of the rules considered noteworthy. Finally, the last rule holds that spoken discourse is characterized by the combination {arg2MC & non-subject role}, which also is plausible form a linguistic point of view in light of the observed preference for right embedding in spoken language and the fact that the presence of a second overt argument increases the chance of right embedding. Of course, the correlation is not perfect as relative clauses in English can also modify elements of the clause that do not have argument status, as exemplified in (73).

(73) *He lives in Berlin, which is the capitol of Germany*.

So, in the light of the results of the *k*-optimal pattern discovery technique, we learn that a typical non-finite RCC in spoken language differs substantially from that predominantly used in written discourse. The respective typical forms are exemplified in (74) and (75).

(74) The high loading rates **applied** to anaerobic ponds **ensures** [sic] that oxygen is utilized more rapidly than it is replaced by atmospheric diffusion. [W2A-021 #066]
Written type$_{schematic}$: $_{NP}$[N-bar[$_{RC}$[∅ VP$_{RC}$[__ -*ed* PRT]]]] VP$_{MC}$

(75) Oh well obviously he has large shoes to fill. [S1A-018 #136:1]
Spoken type$_{schematic}$: NP$_{MC}$ VP$_{MC}$ [V $_{NP}$[N-bar $_{RC}$[*to*-inf __]]]

So, language users appear to not shy away from complex forms in written discourse (heavy

subjects, center embedding and a high number of arguments—which are almost exclusive realized as full lexical NPs). Spoken language appears to prefer simpler variants in (virtually) all tested domains—the non-subject preference with *to*-infinitival RCs arguably being the exception to the rule (cf. § 3.5.4).

Before we now apply the second pattern recognition technique in our arsenal, let us briefly return to the issue that many of the variable investigated here do correlate with each other: It certainly is a viable question to ask why this particular set of variables was chosen in the first place (as it usually considered good practice in statistical modeling to avoid highly correlated descriptors). To this very sensible comment I would like to reply that the goals of this study are somewhat different from what is usually asked for in statistical approaches to natural language. The goal here is not to construct elegant mathematical models that account for the variation of a given data set. Instead, the primary goal of this study is to identify those forms that co-occur with above chance level frequency so as to suggest that their repeated usage makes them relatively easier to process. Consequently, this study does not entertain a functional division of variables into independent variables and dependent variables and then explain the behavior of the response variable on the basis of the predictors. What we are interested in is the identification of entrenched patterns. And as long as correlating variables are not encoded on the same morpheme they serve as accumulative cues to the interpreter, each of which helps to detect and anticipate larger structures. And even though the association rule technique with its conditional format does suggest a design similar to that of a regression analysis, this is not what we should take it to mean. The change from thinking in terms of functional roles played by certain variables to explain another to thinking in terms of holistic configurations, and to not distinguish variables at a functional level (dependent – independent), will become clearer once we turn to the configural frequency analysis.

4.2.2 Configural frequency analysis

As already indicated in the above, *configural frequency analysis* (CFA) is yet another method suitable for the task at hand, i.e. entrenched pattern detection. CFA is a statistical technique that used in the analysis of categorical (i.e. nominal or ordinal) variables (Lienert 1969, peril 1985, Krauth 1993, Krauth & Lienert 1973/1995, von Eye 1990, Lautsch and von Weber

1995, Gries 2008). Such variables are characterized by the fact that the states they can assume are mutually exclusive and therefore lend themselves well for a cross-classification as expressed in a contingency table. The dimensionality of such tables is of course dependent on the number of factors crossed so that a set of *d* variables results in a *d*-dimensional contingency table.

Typically, the focus of statistical procedures that are used in the analysis of contingency tables is on the disclosure of the relationships among the variables that make up the table. It is asked whether there exists an associative relationship between the variables and the statistical procedure will output some coefficient that can be interpreted in analogy to those produced by correlational techniques. The approach taken by CFA techniques is somewhat different in so far as it focuses on the description of groups of individuals (in the logical sense). To illustrate the difference, let us again consider a non-linguistic example discussed in von Eye (1990), which involves three categorical variables, extraversion (E), criminal behavior (C) and intelligence (I) each of which has two levels (so that: E1 = extraverted, E2 = introverted, C1=presence of criminal record C2 = absence of criminal record, I1 = highly intelligent, I2 = less intelligent). Consequently, we get 2 x 2 x 2 = 8 cells in our contingency table. Figure 82 presents the resulting possible configurations graphically.

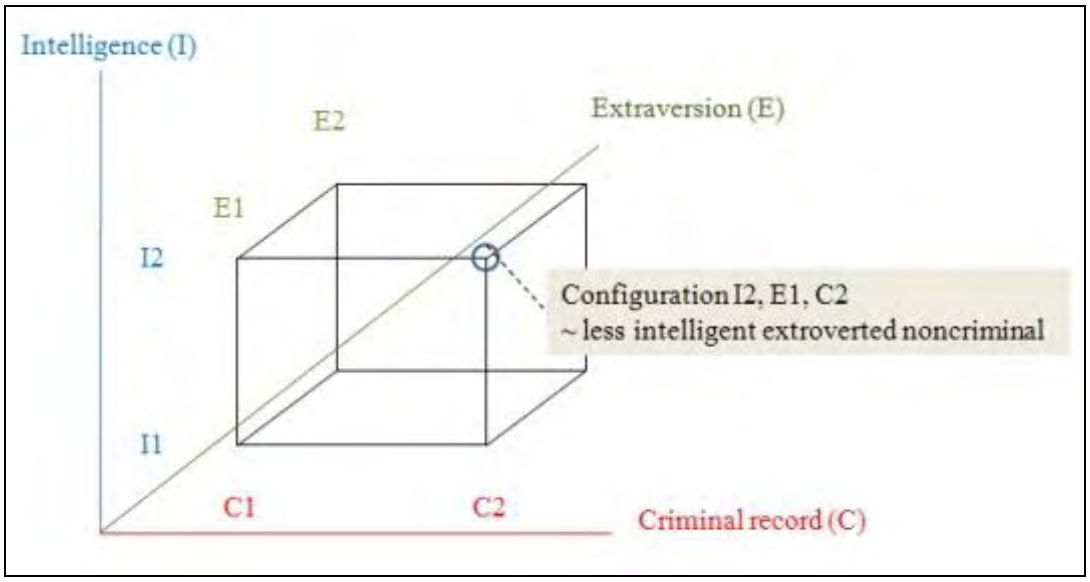

Figure 82: Illustration contingency cube

Each corner of the cube designates a position in a three dimensional space corresponding to a particular combination of properties so that the encircled corner corresponds to the set of properties { [- INTELLIGENT], [+ EXTROVERTED],[-CRIMINAL] }. Having collected an adequate amount of data, we could aim at detecting correlative relations between intelligence and criminal record for example using techniques like log-linear modeling. It is important to note that a correlation, if attested, would then be assumed to hold for all levels of intelligence and criminal record, respectively. Such approaches focus on variables and their interrelationships. The approach underlying CFA, however, focuses not on the variables but rather on the individuals or groups of individuals that share a particular pattern (or *configuration*) of properties. So, generally speaking, instead of focusing on disclosing relations among a set of variables {A, B, C}, the CFA approach focuses on individual cells in the contingency table and aims at identifying those cells that are, statistically speaking, special. Being special translates either in a particular pattern being statistically less frequent than expected under the assumption of statistical independence (=H0)—such patterns are labeled as *antitypes*— or, and this is a lot more relevant for our present purposes, a pattern could be statistically special because it occurs more often than expected under H0, in which case it is called a *type*. In the example above, a possible outcome of some socio-psychological study could be that the number of individuals belonging to the group of highly intelligent extroverted criminals is far greater than expected, which is to say that there is a group that is characterized by a particular combinations of attributes, whose co-occurrence is statistically speaking surprisingly frequent.

Now, obviously the present study is not about persons that share certain characteristics. However, even though CFA techniques are in fact mainly used in the fields of differential psychology, psychiatry, and medicine (cf. Krauth & Lienert 1973/1995), the application of CFA is of course not restricted to an analysis of groups and individuals of a particular type, say humans. The configurations of interest here are feature bundles that can be used in the description of relative clause construction. Specifically, CFA techniques will be used to detect types of relative clause constructions that occur more frequently than expected and which can therefore be taken to have a privileged status in the linguistic system of an (idealized) language user. A big advantage of the configural approach is that it does not

distinguish between dependent and independent variables, which really is not helpful if we have no clear ideas in place regarding which variables come first in a causal path of explanation. In the configural view, all variables are put on the same level (cf. von Eye & Pena 2004 for a discussion).[21]

4.2.2.1 Methods & Results: Configural frequency analysis

The same data set that was subjected to the association rule mining procedure, 285 non-finite RCCs, was subjected to the CFA. The analysis focuses on maximally specified configurations, i.e. configurations that have specified values for all the variable slots as these correspond to the most detailed patterns. Our goal is to learn what set of patterns must be considered to be entrenched in the minds of (idealized) language users. The CFA is thus geared to pick out cells that are a) maximally specific and b) contradict the random (=base) model.

The examination of the cells in the resulting multidimensional contingency tables, i.e. the evaluations of the departure of the observed frequency of a configuration from its expected value under H0, was computed conservatively on the basis of binomial tests, i.e. exact hypothesis tests, which were corrected via the Bonferroni correction so as to safeguard against falsely significant results (cf. Abdi 2007). Table 20 presents an excerpt of the obtained results and is restricted to those configurations that occur with a frequency greater than 5. The results are—as the watchful reader may have already noticed—ordered by configural frequency (again the complete table can be found in the appendix).

Table 19: Results CFA non-finite RCC (excerpt)

Med	type	internal role	embed	valency MC	valency RC	Freq	Exp	P.adj.bin	Dec	Q
Written	edp	subject	center	arg1	arg2	41	13.91	7.41E-08	***	0.1
Written	edp	subject	center	arg2	arg2	34	12.17	8.51E-06	***	0.08

[21] This general perspective is shared by other approaches to domains that involve bio-psycho-social variables, e.g. research using the Experience Sampling Method (cf. Delespaul 1995).

Spoken	to.inf	nonSUBJ	right	arg2	arg2	20	1.117	6.78E-17	***	0.067
Written	edp	subject	right	arg1	arg2	16	14.41	35.49791	ns	0.006
Spoken	to.inf	nonSUBJ	right	arg1	arg2	14	1.277	7.84E-09	***	0.045
Written	edp	subject	right	arg2	arg2	12	12.61	48.52607	ns	0.002
Written	ingp	subject	center	arg2	arg1	10	1.881	0.002414	**	0.029
Written	ingp	subject	center	arg1	arg2	9	6.128	15.75074	ns	0.01
Spoken	edp	subject	right	arg1	arg2	9	7.668	34.61616	ns	0.005
Written	ingp	subject	center	arg2	arg2	7	5.362	27.98398	ns	0.006
Spoken	to.inf	nonSUBJ	right	arg1	arg1	6	0.448	0.000701	***	0.02
Spoken	to.inf	nonSUBJ	right	arg2	arg1	6	0.392	0.000329	***	0.02
Written	ingp	subject	right	arg2	arg2	6	5.554	46.19282	ns	0.002
Spoken	ingp	subject	right	arg2	arg1	5	1.037	0.399559	ns	0.014
Spoken	to.inf	nonSUBJ	center	arg1	arg2	5	1.233	0.812199	ns	0.013
Written	edp	subject	center	arg2	arg1	5	4.269	40.69121	ns	0.003

4.2.2.2 Discussion: Configural frequency analysis

A straightforward way of interpreting the results sketched in Table 20 is to look for configurations that exhibit an observed frequency significantly greater than expected. From this set we should start our discussion with those configurations that exhibit the largest values of Q, which is a *coefficient of pronouncedness* and as such quantifies the magnitude of the cells departure from the base model. Looking at the results from this perspective, we observe that the configuration which ranks highest among all candidates is a type characteristic of written language characterized by the following features {*-ed* participial, center embedded, subject relative, bivalent RC, monovalent MC}. The second most entrenched written type is characterized by the very similar feature set {*-ed* participial, center embedded, subject relative, bivalent RC, bivalent MC}. The two dominant patterns of non-finite RCC in written discourse, which jointly already account for roughly 40% (75/186) of the written cases, thus differ in only a single feature, namely the valency of the main clause predicate. An example of this pattern is given in (76).

(76) The subject offered must correspond to those approved in Appendix I.

[W2D.007 #038:1]

With respect to spoken language we observe that the two most characteristic patterns also differ in just a single parameter and again this difference concerns the valency of the main clause predicate. If we abstract away from this factor, the most entrenched type of non-finite RCC in spoken language are described by the feature set {*to*-infinitival, right embedded, non-subject relative, bivalent RC predicate} as exemplified in (77).

(77) Yeah that would be an interesting one to do. [S1A-053 #090:1]

So, given the dimensions of contrasts used in the characterization here, we observe that the preferred patterns non-finite RCCs are orthogonally different across modalities.

At a methodological level, we observe that there is a strong agreement across techniques, i.e. the results obtained from the association rule mining technique are fully compatible with the results obtained from the CFA. We should note, however, that the ARM technique did not output rules complex enough to count as direct correspondents of the CFA types. This may very well due to the mechanics of the ARM technique and the significance filter (recall that rules that did not reach the .05 significance level were filtered out). Since more specific patterns are instantiated by fewer examples in the data set—simply because they impose more constraints of the data, and p-values are dependent on sample sizes, it is harder for such rules to reach the .05 level. Nevertheless, the high level of convergence across methods should give us some confidence in asserting that (at least) the detected patterns enjoy a cognitively prominent status.

4.2.3 Identifying exemplar clusters: RCC-similarity in configural space

The association rule mining technique and the configural frequency analysis have disclosed certain patterns that occur with above-chance frequencies. Now, given our theoretical commitments regarding the relationships of a) frequency of occurrence and mental representation and b) mental representation and processing difficulty, these results suggest

that those patterns that are identified as types (or significant rules with high leverage values) are deeply entrenched, hence easy to access, and hence easy to process. Now, given the nature of the statistical procedures employed here and the rather limited data set available for analysis, it appears worthwhile to take all attested patterns into account (even those that are not statistically significant) and relate these to each other with respect to their degree of similarity. If we structure our constructions this way, we should be able to derive processing hypotheses not only for entrenched patterns but also for all other patterns by means of relating those other patterns to the entrenched ones. Following the logic behind exemplar based models, we can (at least roughly) estimate the expected processing difficulty of an arbitrary construction C, by assessing C's distance to the nearest (cluster of) entrenched pattern C_E in some n-dimensional state space S:

CATEGORIZATION AS DISTANCE TO ENTRENCHED PATTERNS:

The smaller the distance between C_i and C_E in S,

the easier should it be to categorize C_i.

A promising way to structure objects with respect to their degrees of (dis)similarity is using a cluster analytical technique. We shall follow this line of thinking here and start our discussion of this issue with a brief introduction into these methods and—as their faithful application requires a solid understanding of their workings—the choices that led to the specific technique used for the present purposes.

4.2.3.1 Relating patterns by similarity

Cluster analysis can be conceived of as a family of techniques that aim at allocating objects to groups (or clusters) on the basis of their (dis)similarity. A clustering procedure starts with a data set that contains information about a sample of objects and seeks to organize these objects into homogeneous groups so that objects that are judged by the clustering algorithm to be very similar are allocated to the same group and dissimilar objects are put in different

groups. The variant of cluster analytic techniques that is made use of here, hierarchical agglomerative clustering analysis (HACA), starts off with as many clusters as there are objects to be classified and then iteratively classifies these objects into ever larger groups until all objects have been assigned some group membership. Other clustering techniques partition the data into a user-specified number of clusters (=partitioning) or are divisive, i.e. they start with a single cluster and then split up the aggregate until all objects are in different groups. Everitt (1993) provides an approachable introduction presenting many different clustering techniques. A HACA involves two parts: first, the calculation of a (dis)similarity matrix, which contains for each pair of objects an expression of the degree of (dis)similarity between the members of that pair. Second, once we have a similarity assessment of all pairs of object in place, we need to find a way to link them all together to complete the hierarchical structure and include all objects (=*amalgamation*). There are numerous ways to measure a) the (dis)similarity between objects and b) how to link them and when.

Hence, the difficulties for clustering techniques involve questions regarding what measure should be used to perform these two tasks and as each choice will impose a different type of structure on the data, the choices are not without their consequences. The next section will disclose the algorithmic choices opted for here and present the results for the clustering of non-finite RCCs.

4.2.3.2 Method & Results: Clustering Non-finite RCCs

In order to structure the attested patterns on the basis of their similarity a hierarchical agglomerative cluster analysis was performed. Degrees of (dis)similarity were expressed in terms of the metric distance between the objects in the configurational state space so that similarities between non-finite RCC types correspond to metric distance between objects in that space. This required recoding the data into corresponding numerical values. Quite generally, this was done by substituting each factor level of a given variable against a positive integer starting with 1 assigned to the value that corresponds to the variant that is commonly associated with fewer difficulty (e.g. right for embedding). The association of small numbers and lesser amounts of processing difficulty is only a mnemonic device. Of course, as long as the recoding is consistent nothing hinges on these choices. Each consecutive factor level was

re-labeled accordingly until the most complex level was assigned the largest number. Consequently, each RCC is identified with a coordinate vector, whose elements specify the Cartesian coordinates of the endpoint, i.e. the position of the object in configurational space. In this case, five of the six factors whose values serve as coordinates in the state space are binary with a sixth variable, type of RC, allowing three levels. At this point, we must decide on the issue of standardization as those variables with a greater range have a greater impact on the (dis)similarity assessment. Because type of RC—the only three-level factor—was considered to deserve a prominent role in the description of the construction, the implicit variable weighting was accepted as a desirable property of the similarity assessment. Following many approaches in exemplar-based representation, Euclidean distance was chosen to express degrees of (dis)similarity (cf. Gower 1985 and Everitt 1993 for discussions of this possible alternatives). Expressing (dis)similarity in terms of distance in some metric space entails that similarity is conceptualized as a symmetric notion, since the distance $d(x,y) = d(y,x) \geq 0$. This symmetry assumption, however, is controversial (cf. Tversky 1977). To what extent this choice is relevant for the present study is considered an open empirical question that future research may address. Quite generally, the overall strategy employed in this work was to always opt for more conservative choices unless there are pressing reasons to do otherwise.

The amalgamation procedure was conducted using the *neighbor.joining tree estimation* algorithm (Saitou & Nei 1987). This method was judged to be more adequate than potential alternative measures because in contrast to those alternatives, it produces unrooted (phylogenetic) trees. These trees do not culminate in a binary split, which is not always desirable when it comes to clustering linguistic data. Simply put, it is not always sensible to impose a particular kind of structure on ones data (cf. Cysouw 2005 for a discussion in a typological context whose argumentation is also valid in the present context). A solution of a clustering method is typically represented as a tree structure (or dendrogram). Figure 83 presents the unrooted tree. Table 21 serves as a legend enabling a more convenient interpretation.

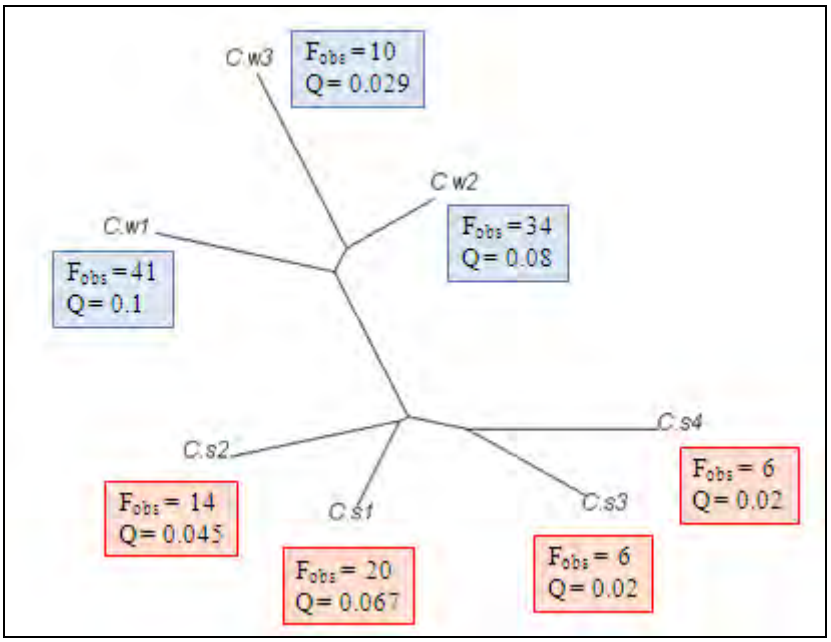

Figure 83: Unrooted phylogenetic tree (nonfinite RCC types)

Table 20: Description of top configurations

label	Med	type	internal role	embed	valency MC	valency RC	Freq	Q
C.s1	Spoken	to.inf	nonSUBJ	right	arg2	arg2	20	0.067
C.s2	Spoken	to.inf	nonSUBJ	right	arg1	arg2	14	0.045
C.s3	Spoken	to.inf	nonSUBJ	right	arg1	arg1	6	0.02
C.s4	Spoken	to.inf	nonSUBJ	right	arg2	arg1	6	0.02
C.w1	Written	edp	subject	center	arg1	arg2	41	0.1
C.w2	Written	edp	subject	center	arg2	arg2	34	0.08
C.w3	Written	ingp	subject	center	arg2	arg1	10	0.029

Figure 83 shows the structure identified for the set of nonfinite types in the data set, i.e. it shows only those patterns that occur significantly more often than expected on the basis of chance. Each terminal node represents one of these types and is labeled in such a way that we can read off quickly the modality of that type (s = spoken; w = written) and also the rank of that type within that modality. The initial 'C' simply indicates that the object is a configuration. So, *C.w1* is the top-ranked configuration from the written modality, *C.w2* the second most pronounced pattern in that modality and so on. For each type, the respective box provides information about the observed absolute frequency and the coefficient of

pronouncedness of that pattern, which allows us to directly see which type not only is significantly more frequent than expected but also which pattern is numerically dominant. High values for Q, i.e. highly pronounced patterns, require relatively high token frequencies as only those allow the contribution to the global chi square sum to be high enough to result in high values for Q. In virtue of the underlying amalgamation procedure, Figure 83 represents the degree of (dis)similarity in terms of branch length, we can think of the branches of the network as representing pathways that one must travel to get from point A, i.e. an arbitrary terminal node, to point B, i.e. another arbitrary terminal node. The longer one has to travel to get from A to B, the greater the degree of dissimilarity of the objects that A and B stand for. In this case, we can see that there is no route from an arbitrary spoken node to any given written one that is shorter than the longest route between any two points from the same modality. Simply put, the spoken types are clearly separated from all written types suggesting that there really are different preferred patternings across modalities. To get a more complete picture of the patternings in our data, we can apply the very same procedure to all attested patterns in the data and relate them to each other in exactly the same way. Figure 84 presents the results.[22]

[22] Unfortunately, identical configurations are plotted on top of each other making it hard to read the respective labels. An alternative, informationally equivalent representation is given in the appendix.

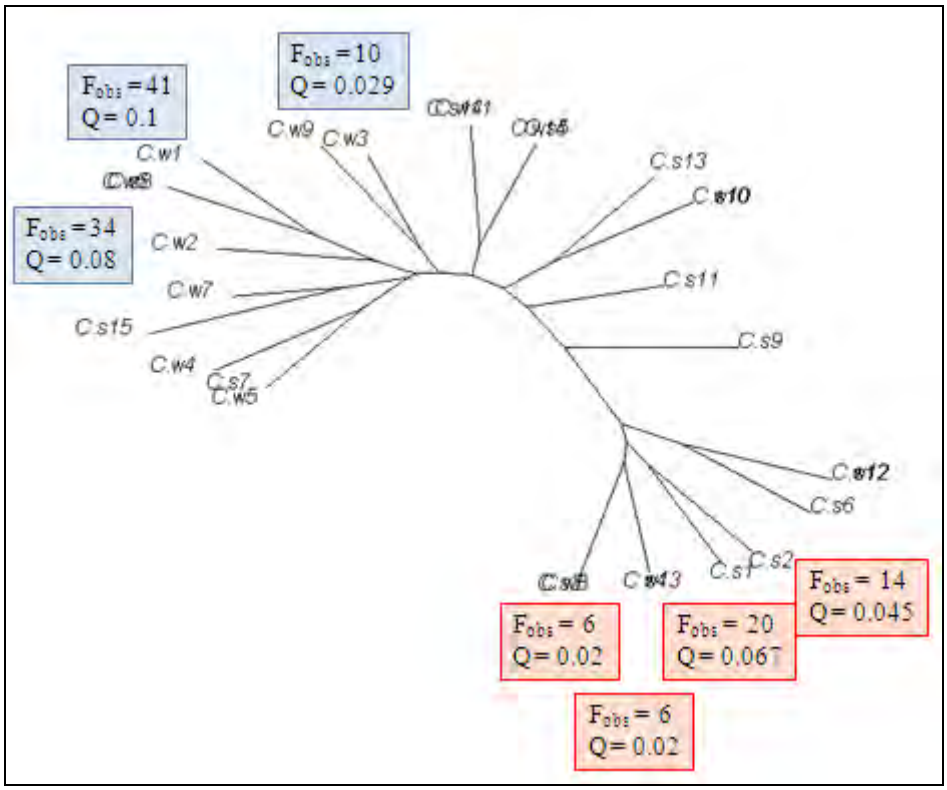

Figure 84: Unrooted phylogenetic tree (all attested non-finite RCC)

Again the red (for spoken) and blue (for written) boxes help us find the types detected for the respective modalities. This richer tree allows us to appreciate the rather distinct patterning of non-finite RCCs across modalities even more.

In summation, using the descriptor variables TYPE OF RC, RELATIVIZED ROLE, TYPE OF EMBEDDING, and TRANSITIVITY OF MC AND RC, we were able to detect pronounced differences across modalities. The two methodologies, the association rule mining technique and the configural frequency analyses produced very similar, highly compatible results, but while the ARM techniques was—on the basis of the available data—only able to detect rather simple rules with a maximum of three specified properties on the LHS, the CFA was capable of identifying seven entrenched patterns. All of the results of the numerous bivariate analyses could be replicated by the multivariate approach, but only the latter approach was able to show the interdependencies of the main effects disclosed in § 4.1. This is viewed a huge advantage of the multivariate approach taken here. In the (exemplar-based) view presented here processing difficulty is viewed as a recognition process in which chunks of features of a

pattern are detected and then used to anticipate the overall form of a perceived input. Less emphasis is put on the conceptual properties and nature of the variable themselves. Focusing on the impact of a single variable on the overall difficulty of a sentence suggests that it is the intrinsic complexity of a construction that determines its processing demand (e.g. working memory load). In the view taken here, however, there is nothing above and beyond the frequency of past processing to determine the ease of future processing. The frequency of a pattern in turn is crucially determined by the usefulness of the communicative function in the discourse that is associated with a particular structure/form.

The clustering technique, employing Euclidean distance as an expression of similarity and a neighbor joining algorithm, helped to reveal that there good reasons to believe that written and spoken discourse are characterized by very different preferred patternings: while spoken discourse is characterized by a heavy use of right embedded *to*-infinitival RC, we observe a strong preference for *–ed* participial reduced relatives and center embedding for the written modality. The analysis furthermore revealed that valency patterns of both RC and MC do not discriminate between the modalities in any particularly interesting way.

4.3 Finite RCC

As already indicated, our discussion of non-finite RCC was primarily geared to familiarize the reader with the methodological approach on the basis of a less complex construction. The focus of this study clearly lies on finite RCCs because a) their inherent complexity emphasizes the utility of the configurational approach advocated here and b) the psycholinguistic literature has clearly concentrated on finite patterns thus offering a far greater body of results, which can—and will—be compared to the results of this study.

We will start our discussion with an analysis of subject RCC (§ 4.3.1) and discuss non-subject RCC thereafter (§ 4.3.2). During these explications, it is helpful to keep in mind that from the theoretical stance assumed here, we may derive the following general prediction:

GENERAL PREDICTION (AND IMPLICATIONS OF THE PATTERN SEARCHES)

If p, degrees of entrenchment corresponds to processing difficulty, and if q, the methodology employed here can provide good enough approximations of degrees of entrenchment, then the detected patterns should mirror the patternings of processing difficulty typically observed in experimental studies.

Finally, we will have a look at specific phenomenon occurring with certain types of non-subject RCC, namely the omission of non-obligatory relativizers. This will serve as a last test of the methodology proposed (§ 4.3.3).

4.3.1 Finite subject RCCs

For the analysis of finite subject RCCs the following descriptors were chosen:

- Head type, i.e. the syntactic type of head (lexical/pronominal)
- Additional modification, presence or absence of a AdjP or PP modifier
- Definiteness of the head (y/n))
- Concreteness of the head, referent perceivable by the senses (y/n)
- R-type, i.e. type of relativizer (*that*, *which*, *who*)
- Type of embedding (right/center)
- Valency main clause predicate (one argument, two arguments or more)
- Valency relative clause predicate (one argument, two arguments or more)
- Medium (spoken/written)

Following the procedure applied to the nonfinite RC-constructions, the analysis proceeded by way of first searching for strong association rules and then searching for CFA-types. As we will see shortly the obtained results led to the introduction of an additional step in the analysis, namely the introduction of hierarchical CFA models. But let us first have a look at

the results and why these called for this methodological elaboration.

4.3.1.1 Association rule mining: Finite subject RCCs

The rule mining technique was applied using the same conservative settings described before. A total of 386 finite subject relatives were submitted to the analysis. On the basis of these data the ARM technique detected 47 interesting associative relationships that satisfied the constraints, i.e. allowing only modality values on the RHS of a given rule and all but those values on the LHS. Again most of the detected rules are quite simple rarely specifying more than three properties on the LHS. We will confine ourselves to a presentation of the most complex rules (a complete list of all rules can be found in the appendix).

Rules detected for written discourse (selection)

{ ABSTRACT.HEAD & CENTER & ARG2MC } ⇒{ WRITTEN }
with strength = 0.900
lift 1.62: the strength is 1.62 times greater than the strength if there were no association
leverage = 0.0267: the support is 0.0267 (10.3 cases) greater than if there were no association

{ LEXICAL & ABSTRACT.HEAD & ARG2MC & ARG2RC } ⇒{ WRITTEN }
with strength = 0.877
lift 1.57: the strength is 1.57 times greater than the strength if there were no association
leverage = 0.0473: the support is 0.0473 (18.3 cases) greater than if there were no association

{ ABSTRACT.HEAD & CENTER & ARG2RC } ⇒{ WRITTEN }
with strength = 0.867
lift 1.56: the strength is 1.56 times greater than the strength if there were no association
leverage = 0.0361: the support is 0.0361 (13.9 cases) greater than if there were no association

Rules detected for spoken discourse (selection)

{ RIGHT & PRONOMINAL } ⇒{ SPOKEN }
with strength = 0.815
lift 1.84: the strength is 1.84 times greater than the strength if there were no association
leverage = 0.0260: the support is 0.0260 (10.0 cases) greater than if there were no association

{ THAT & ARG1RC & RIGHT } ⇒{ SPOKEN }
with strength = 0.771
lift 1.74: the strength is 1.74 times greater than the strength if there were no association
leverage = 0.0298: the support is 0.0298 (11.5 cases) greater than if there were no association

We observe that more complex rules have been found for the written modality, which is either a reflex of the smaller sample size for spoken language (170 spoken versus 215 written cases) or is an indication of a higher degree of heterogeneity among the set of spoken patterns, or a combination of the two. As we will see in our discussion of the CFA results, there are good reasons to believe that the hypothesized heterogeneity plays a greater role for these findings. Concerning the properties identified as being associated with written discourse, we find a preference for higher degrees of valency of both the RC and the embedding MC, a preference for abstract referents and—similar to what we disclosed for non non-finite RCCs—a preference for center embedding. The association of center embedding and written language is weaker for finite subject RCCs than it is for non-finite RCC though. The rule {center} ⇒ {written} for non-finite RCCs is stronger (strength = 0.886) and also has more leverage (0.1145) than the corresponding rule for finite subject RCCs (strength = 0.706, leverage = 0.0619; cf. appendix). For spoken language, the most important finding concerns the detected preference for pronominal heads. The fact that the ARM technique restricts its output to rather simple rules has its pros and cons. We shall return to this issue in our discussion of the CFA results.

4.3.1.2 CFA: Finite subject RCCs

Let us now turn to the CFA results. Since our contingency table is quite complex, allowing 1024 possible configurations (i.e. factor level combinations), we will look at only those patterns here that exhibit an observed frequency greater than four. This threshold level is completely arbitrary and motivated only by very pragmatic considerations. While only few patterns occur more than five times in the data, there just happened to be a rather large number of configurations occurring four times. These were considered worthy of presentation. Of the 1024 possible patterns only 183 were actualized in the data set and only 25 occurred with a token frequency of four or greater. Table 21 presents an overview of these results.

Table 21: Finite subject RCCs - CFA results

medium	head type	add mod	def head	conc head	R type	val RC	emb	val MC	Freq	Exp	P.adj.bin	Dec	Q
spoken	lexical	no.mod	indef	conc	who	arg1RC	right	arg2MC	6	0.5833	0.03	*	0.014
spoken	lexical	no.mod	def	conc	who	arg1RC	right	arg1MC	5	0.6606	0.61	ns	0.011
spoken	lexical	no.mod	def	abs	that	arg1RC	center	arg1MC	4	1.0676	23.78	ns	0.008
spoken	lexical	add.mod	def	abs	that	arg1RC	right	arg1MC	4	0.8678	12.10	ns	0.008
spoken	lexical	no.mod	def	conc	who	arg2RC	right	arg1MC	4	0.7965	9.07	ns	0.008
spoken	lexical	no.mod	indef	abs	which	arg1RC	right	arg1MC	4	1.4618	62.04	ns	0.007
spoken	lexical	no.mod	indef	abs	which	arg1RC	right	arg2MC	4	1.3315	47.11	ns	0.007
spoken	lexical	no.mod	def	abs	that	arg1RC	right	arg1MC	4	1.5079	67.86	ns	0.006
spoken	lexical	no.mod	def	abs	that	arg2RC	right	arg1MC	4	1.8181	113.86	ns	0.006
written	lexical	add.mod	indef	abs	which	arg2RC	right	arg2MC	8	1.1617	0.03	*	0.018
written	lexical	no.mod	indef	abs	which	arg2RC	right	arg2MC	8	2.0185	1.14	ns	0.016
written	lexical	no.mod	indef	abs	which	arg1RC	right	arg2MC	7	1.6741	1.71	ns	0.014
written	lexical	add.mod	indef	abs	which	arg1RC	right	arg2MC	6	0.9635	0.49	ns	0.013
written	lexical	no.mod	def	abs	that	arg2RC	center	arg1MC	6	1.6184	6.39	ns	0.011
written	lexical	no.mod	def	conc	who	arg2RC	center	arg1MC	5	0.709	0.83	ns	0.011
written	lexical	no.mod	indef	abs	that	arg2RC	right	arg2MC	6	2.0185	17.36	ns	0.01
written	lexical	no.mod	def	abs	which	arg2RC	center	arg2MC	5	1.4742	17.58	ns	0.009
written	lexical	no.mod	indef	conc	who	arg2RC	center	arg1MC	4	0.6873	5.47	ns	0.009
written	lexical	add.mod	def	abs	that	arg1RC	center	arg1MC	4	0.7725	8.18	ns	0.008
written	lexical	add.mod	def	abs	which	arg2RC	center	arg1MC	4	0.9314	15.29	ns	0.008
written	lexical	add.mod	def	abs	which	arg2RC	center	arg2MC	4	0.8484	11.22	ns	0.008
written	lexical	add.mod	indef	abs	that	arg2RC	right	arg1MC	4	1.2753	41.37	ns	0.007
written	lexical	no.mod	def	abs	that	arg1RC	center	arg2MC	4	1.2227	36.37	ns	0.007
written	lexical	no.mod	indef	abs	that	arg2RC	right	arg1MC	4	2.216	187.76	ns	0.005
written	lexical	no.mod	def	abs	which	arg2RC	right	arg1MC	4	2.286	202.11	ns	0.004

As shown in Table 21, the CFA could only detect two fully specified configurations that occurred significantly more often than expected. For the spoken modality this pattern is characterized by a rather simple (univalent) right embedded RC introduced by *who* with an indefinite, lexical, concrete head that is no further specified, i.e. there is no modifying AdjP or PP. An example of this pattern is given in (78):

(78) <Uh> the History of Art department has one <uh> member of staff who's on an academically-related scale. (S1B-075 #021:1)[23]

This pattern is expected to occur 0.5 times in the data and did actually occur 6 times ($p_{adjusted\ binomial} = 0.03$). The other significant configuration is a written RCC type characterized by a more complex (bivalent+) right embedded RC introduced by *which* with a indefinite lexical head accompanied by an additional modifier and referring to an abstract entity. An example is given in (79):

(79) The first method is a Best-Fit Analysis technique which calculates the best-fit point, line and plane through a three-dimensional data array. (W2A-036 #093:1)

This configuration is expected to occur 1.1 times and did actually occur 8 times ($p_{adjusted\ binomial} = 0.018$). When we look at the dimensions where these patterns contrast, we find that all differences go in the direction predicted by the processing account, i.e. all less complex properties are associated with spoken discourse (additional modifier > no additional modifier, abstract > concrete, 2 arguments in RC > 1 argument in RC). However, a closer look at Table 21 reveals some curious details. Notice that there is not a single pattern in the list exhibiting a

[23] The expression *member of the staff* was treated as a coherent unit and not analyzed as an analytic post-modified structure.

pronominal head. This surely is surprising—at least for those who spent their days reading through large amounts of relative clause constructions. This innocent fact has a lot to tell us. First, it can tell us that it is indeed worthwhile to make use of more than just a single method, since the rule mining technique in fact detected two rules with that specification ({PRONOMINAL HEAD → SPOKEN} and {RIGHT EMBEDDING & PRONOMINAL HEAD → SPOKEN}). Second, it shows us the limits of what the CFA technique applied so far can reveal about the available data set and thus, third, it suggests that it may be advisable to also look out for less than fully specified patterns.

4.3.1.3 Interlude: Variability in typical patterning: gains and pains of hCFA

In order to counter some of the undesirable effects of the poor ratio of state space complexity and size of the data set, the next step in the analysis involved the employment of a hierarchical CFA so as to bring to light more schematic significant patterns. As less specific constructions impose fewer constraints on their instantiations, a decrease in specificity is expected to result in a greater number of detected types. The general idea behind a hierarchical CFA is easily illustrated by way of a brief glimpse at the form of its actual input. Consider Table 22:

Table 22: Illustration of hCFA procedure

medium	syn type	add mod	definiteness	concreteness	R type	val RC	emb	val MC	Freq
written	.	.	.	.	.	.	.	.	215
spoken	.	.	.	.	.	.	.	.	171
.	lexical	.	.	.	.	.	.	.	341
.	pronominal	.	.	.	.	.	.	.	45
.	.	no.mod	.	.	.	.	.	.	245
.	.	add.mod	.	.	.	.	.	.	141
.	.	.	definite	.	.	.	.	.	196
.	.	.	indefinite	.	.	.	.	.	190
.	.	.	.	abstract	.	.	.	.	239
.	.	.	.	concrete	.	.	.	.	147
.	.	.	.	.	that	.	.	.	139
.	.	.	.	.	which	.	.	.	139
.	.	.	.	.	who	.	.	.	99
.	.	.	.	.	no	.	.	.	9
.	.	.	.	.	.	arg2RC	.	.	211
.	.	.	.	.	.	arg1RC	.	.	175
.	.	.	.	.	.	.	right	.	226
.	.	.	.	.	.	.	center	.	160
.	.	.	.	.	.	.	.	arg1MC	202
.	.	.	.	.	.	.	.	arg2MC	184
written	lexical	.	.	.	.	.	.	.	200
spoken	lexical	.	.	.	.	.	.	.	141
spoken	pronominal	.	.	.	.	.	.	.	30
written	pronominal	.	.	.	.	.	.	.	15
written	.	no.mod	.	.	.	.	.	.	128
spoken	.	no.mod	.	.	.	.	.	.	117
written	.	add.mod	.	.	.	.	.	.	87
spoken	.	add.mod	.	.	.	.	.	.	54

A hierarchical CFA (hCFA) provides a means to systematically exclude those factors that contribute little (or nothing) to the contribution of types (and antitypes), while the non-hierarchical form used so far uses all factors simultaneously. Table 22 is meant to illustrate the logic behind hCFA. An hCFA starts off by evaluating first the observed frequencies of all possible states of each single factor in isolation. Once this is done, all two-factor configurations are evaluated, followed by all three-factor configurations and so on until the maximum level of complexity is reached and all factors have entered the analysis. The merits of this capacious approach are of course easily appreciated—the analysis is considerably widened in its scope. But it also does render a subsequent inspection of the results a little more difficult: Integrating all possible configurations into the scope of the analysis in this case results in a total of 32,804 evaluated configurations. A result of this magnitude prevents

any reasonable attempt of serious manual inspection. In order to cope with these rather large amounts of data, the following strategy was applied.

First, all configurations with an observed token frequency smaller than the expected one (including all non-attested patterns) were deleted reducing the number of patterns to 11,355. Of these remaining patterns, 10.494 were judged to be not significant leaving us with 861 patterns that are at least marginally significant ($0.1 > p > 0.05$). If we drop the marginally significant cases as well, we end up with 691 patterns that are significant at the conventional level of significance of $\alpha = .05$. From this set, all those patterns that were not specified for modality were deleted further reducing the set to a total of 409 patterns, which constitutes a more manageable set for manual investigation.

The interpretation of the results turned out to be trickier than one might expect. The most sensible way of interpreting the results would have us sort the patterns by their pronouncedness (/conciseness), i.e. their values for Q. However, at closer inspection of the results, I was dissatisfied with the rankings resulting from this approach. Sorting by Q, leads us to focus on rather schematic types, because more schematic patterns (lower level configuration) are more likely to be frequent and more frequent patterns have a greater chance of scoring high on Q (Krauth & Lienert 1973). However, the patterns we are looking for ideally are deeply entrenched and complex, so ordering by Q does not really lead us to where we want to be. The less-than-optimal consequence of focusing on Q values is best illustrated by looking at the type of result we get. Table 23 presents the top 20 types sorted by Q:

Table 23: Top 20 types of finite subject RCC (sorted by Q)

medium	emb	R type	syn type	add mod	definite	concrete	val RC	val MC	Freq	Exp	Cont chisq	Dec	Q
written	.	which	lexical	.	.	abstract	.	.	82	42.35	37.1249	***	0.115
written	.	.	.	.	.	.	.	.	215	193.00	2.5078	*	0.114
written	.	which	.	.	.	abstract	.	.	84	47.94	27.1289	***	0.107
written	.	.	lexical	.	.	abstract	.	.	142	117.60	5.0615	*	0.091
written	.	.	lexical	.	.	abstract	arg2RC	.	93	64.29	12.8263	**	0.089
spoken	.	who	.	.	.	concrete	.	.	49	16.70	62.4557	***	0.087
written	.	which	lexical	.	.	.	.	.	96	68.40	11.1403	**	0.087
spoken	right	.	.	.	.	.	.	.	124	100.12	5.6961	*	0.084
written	.	which	lexical	.	.	abstract	.	arg2MC	51	20.19	47.0318	***	0.084
written	right	.	lexical	.	indefinite	abstract	.	.	63	33.89	24.9981	***	0.083
spoken	.	who	.	no.mod	.	concrete	.	.	41	10.60	87.1696	***	0.081
written	center	.	lexical	.	definite	.	.	.	68	39.98	19.6441	***	0.081
written	center	.	.	.	definite	.	.	.	73	45.25	17.0144	***	0.081
written	center	.	.	.	.	.	.	.	113	89.12	6.3992	*	0.08
written	center	.	lexical	.	.	.	.	.	103	78.73	7.482	*	0.079
written	.	which	.	.	.	abstract	.	arg2MC	51	22.85	34.6749	***	0.078
written	.	which	lexical	.	indefinite	abstract	.	.	49	20.85	38.0267	***	0.077
written	.	which	lexical	.	.	abstract	arg2RC	.	51	23.15	33.5069	***	0.077
written	right	.	.	.	indefinite	abstract	.	.	65	38.37	18.4912	***	0.077
written	.	.	.	.	.	abstract	arg2RC	.	97	72.77	8.0689	*	0.077

Each line represents a finite subject RCC type. The first nine columns specify the factors under investigation. A '.'-symbol indicates that the level for that factor was not specified in that type. The highlighted types illustrate why I consider an interpretation solely based on Q somewhat problematic. If we sort the results by Q, the second most pronounced/concise configuration detected for the written modality is simply {WRITTEN}, which occurs 215 times (F_{exp} = 193), contributes a mere 2.5 units to the $\chi 2$ sum and is statistically significant only at α = 0.05. Similarly, the second largest value for patterns in the spoken modality (Q=0.084) is assigned to the configuration [SPOKEN & RIGHT EMBEDDED], which occurs 124 times in the data (F_{exp} = 100.12), contributes a mere 5.7 to the total $\chi 2$ sum is statistically significant only at α = 0.05. The point I am trying to make here is rather blunt: such patterns (and their statistics) are not exactly spine-tingling. That is to say that even though the Q values of these types are very high (relatively speaking of course), the other statistics do not exactly suggest that we are dealing with particularly interesting configurations. The Q-coefficient does not seem to fully capture what we consider interesting here and particularly so, if we compare Q across configurations of different degrees of schematicity. For this reason, it appears worthwhile to look at the data from different angles, e.g. from the perspective a pattern's contribution to $\chi 2$, which is closely related to the idea behind Vogel's deviation-from-independence coefficient V (see von Eye & Rovine 1988). Table 24 presents a list of the top

20 types, if we rank by that quantity:

Table 24: Top 20 types of finite subject RCC (sorted by contribution to chi square))

medium	syn type	add mod	definite	concrete	type	val RC	emb	val MC	Freq	Exp	Cont chisq	Dec	Q
written	lexical	.	indefinite	abstract	which	.	right	arg2MC	29	5.82	92.3742	***	0.061
spoken	.	no.mod	.	concrete	who	.	.	.	41	10.60	87.1696	***	0.081
written	.	.	.	concrete	who	.	center	.	36	8.70	85.5914	***	0.072
spoken	.	no.mod	.	concrete	who	.	right	.	29	6.21	83.7013	***	0.06
written	.	no.mod	.	concrete	who	.	center	.	27	5.52	83.473	***	0.056
spoken	.	no.mod	.	concrete	who	arg1RC	right	.	18	2.81	81.9526	***	0.04
spoken	.	no.mod	indefinite	concrete	who	arg1RC	right	arg2MC	8	0.66	81.5859	***	0.019
spoken	pronominal	no.mod	.	concrete	who	arg2RC	.	.	8	0.68	79.4062	***	0.019
spoken	pronominal	no.mod	.	concrete	who	.	.	.	11	1.24	77.1403	***	0.025
written	.	.	indefinite	abstract	which	.	right	arg2MC	29	6.59	76.2885	***	0.059
spoken	pronominal	no.mod	.	concrete	who	.	right	.	8	0.72	73.1703	***	0.019
spoken	pronominal	no.mod	definite	concrete	who	arg2RC	right	.	4	0.20	71.8821	*	0.01
spoken	.	no.mod	definite	concrete	who	.	.	.	25	5.38	71.4893	***	0.052
written	lexical	.	.	concrete	who	.	center	.	31	7.69	70.6605	***	0.062
spoken	.	no.mod	definite	concrete	who	.	right	.	18	3.15	69.9534	***	0.039
spoken	.	no.mod	.	concrete	who	arg1RC	right	arg2MC	11	1.34	69.5457	***	0.025
written	lexical	add.mod	indefinite	abstract	which	.	right	arg2MC	14	2.13	66.3518	***	0.031
spoken	.	no.mod	definite	concrete	who	.	right	arg1MC	12	1.65	64.9591	***	0.027
spoken	pronominal	no.mod	definite	concrete	who	arg2RC	.	.	5	0.34	63.2293	**	0.012
spoken	lexical	no.mod	.	concrete	who	arg1RC	right	.	15	2.49	62.9964	***	0.033

So, in what way is what we see in Table xxx different from what we have seen in Table 23? Comparing these tables, we observe that the top 20 patterns in Table 24 are almost exclusively highly significant (18/20 types are marked as '***', i.e. they exhibit $p \leq 0.001$), while at the same time being more specific (recall that it is generally harder for infrequent types to be judged statistically significant). The mean degree of specificity of the patterns—expressed in terms of how many slots are unspecified—is far greater. As indicated by the highlighting, eight of the top twenty types specify all but two slots (or less). The Q-list in Table 23 does not list a single type of that specificity. However, the mean token frequency is lower for the types in Table 24 (Arithmetic means: 18.95 for contribution to chi-square-ranking versus 85.5 for the Q-ranking). Conceptually speaking, the 'contribution to chi-square'-index may be said to express how surprising it is to find the observed elevated frequency of a pattern.

In order to prevent potential misunderstandings at this point, I would like to emphasize here that I am not suggesting that the contribution to $\chi 2$ is more important than Q. What I am suggesting, rather, is that these different indices highlight different aspects, i.e. different properties, of a given pattern and that a single index cannot capture all aspects that

interest us here.

4.3.1.4 Results: Finite subject RCC across modalities

We may now finally turn to a comparison of the patterns that are typical for written modality and those typical for the spoken discourse. In the attempt to find the best solution to the interpretation problem discussed above, it was decided to focus on more complex patterns while permitting some degree of variability. Specifically, to arrive at faithful characterizations of the patterning differences across modalities, the focus was put on those patterns that have a maximum of one variable slot (complete list of all types (N=409) can be found in the appendix). Table 25 presents what is considered to be the most revealing sub-set of the 409 types detected by the hCFA ordered by modality and Q.

Table 25: Results hCFA - top-ranked finite subject RCC (w/ up to 1 unspecified slot)

medium	name	syn type	add mod	definite	concrete	R type	Val RC	embed	Val MC	Freq	Dec	Q
written	c.w1	lex	.	indef	abs	which	arg2	right	arg2	16	***	0.033
written	c.w2	lex	y	indef	abs	which	.	right	arg2	14	***	0.031
written	c.w3	lex	n	indef	abs	which	.	right	arg2	15	**	0.03
written	c.w4	lex	.	indef	abs	which	arg1	right	arg2	13	**	0.027
written	c.w5	lex	n	.	conc	who	arg2	center	arg1	9	**	0.02
written	c.w6	lex	y	indef	abs	which	arg2	right	arg2	8	*	0.018
written	c.w7	.	y	indef	abs	which	arg2	right	arg2	8	*	0.017
spoken	c.s1	lex	n	def	conc	who	.	right	arg1	9	*	0.02
spoken	c.s2	lex	n	indef	conc	.	arg1	right	arg2	10	*	0.02
spoken	c.s3	.	n	indef	conc	who	arg1	right	arg2	8	***	0.019
spoken	c.s4	lex	n	indef	conc	who	arg1	right	.	8	*	0.018
spoken	c.s5	lex	n	.	conc	who	arg1	right	arg2	8	*	0.018
spoken	c.s6	lex	.	indef	conc	who	arg1	right	arg2	7	*	0.016
spoken	c.s7	.	n	indef	conc	who	arg2	right	arg1	7	*	0.016
spoken	c.s8	lex	n	indef	conc	who	arg1	right	arg2	6	*	0.014
spoken	c.s9	pron	n	def	conc	.	arg2	right	arg1	5	*	0.012
spoken	c.s10	pron	n	def	conc	who	arg2	right	.	4	*	0.01

To evaluate the conciseness of those types it is helpful to know the mean Q value of the detected types from a given modality. For written language, the average value for Q is (sum of all Q values / number of configuration =) 0.04; for the spoken modality the arithmetic mean of Q is 0.03. Hence, we are looking at configurations that are below average in that respect. However, considering their complexity these values must still be considered

remarkably high.

The characterization of the structure of the RCC types in Table 25 can be assessed on the basis of the structure assigning procedure that has been introduced in our discussion of non-finite RCCs. Again, the clustering was conducted using Euclidean distance as an expression of (dis)similarity among all patterns and the neighbor joining algorithm was applied to handle the amalgamation of clusters. Figure 85 presents the resulting dendrogram.

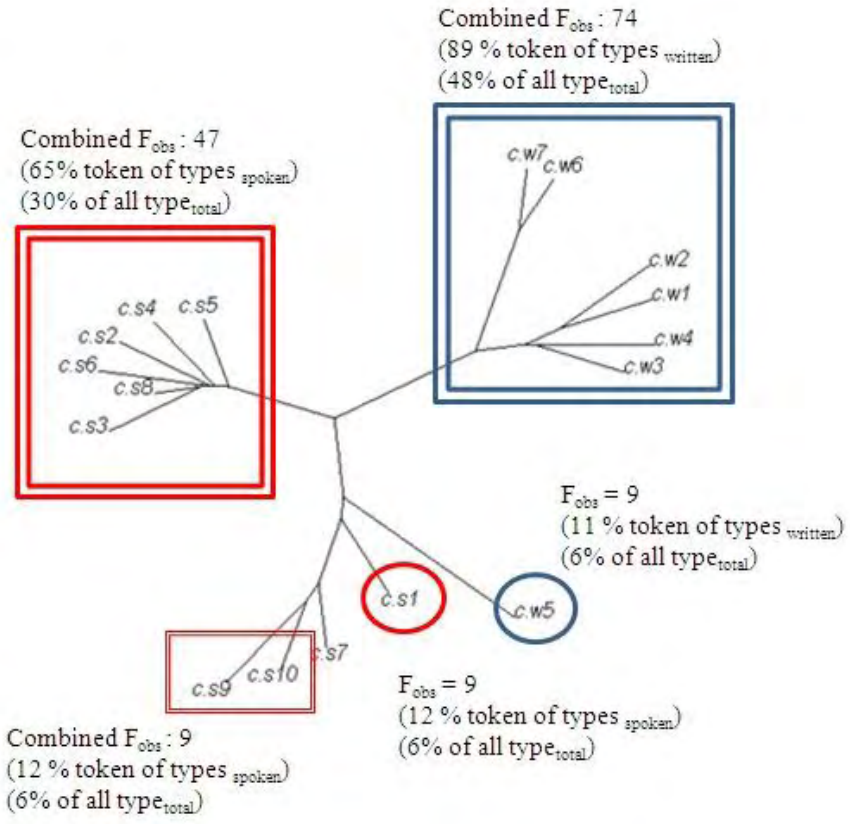

Figure 85: Dendrogram of top ranked finite subject RCC

The first thing to be observed is the clear grouping of the spoken types in the upper left area of the figure incorporates 6 out o 9 spoken types (and no written type). These types jointly provide 65% of the combined token-set of all spoken types and 30% of the tokens of all types. The group is characterized by the properties { $\text{LEXICAL}_{\text{HEAD}}$*, $\text{INDEFINITE}_{\text{HEAD}}$*, $\text{CONCRETE}_{\text{HEAD}}$, ¬ ADD MODIFIER, SINGLE ARGUMENT RC, RIGHT EMBEDDING, R_{WHO} }. The '*' symbol here indicates that the respective property is not shared by all members of the group

but is in fact absent in exactly one pattern. A second major cluster—in the blue double-lined box—characterizes the dominant written patterning. The member of this group make up 89% of the token-set consisting of the types from the written modality and 48% of the total set. They share the properties {LEXICAL$_{HEAD}$*, INDEFINITE$_{HEAD}$, ABSTRACT$_{HEAD}$, BIVALENT+ MC, RIGHT EMBEDDING, R$_{WHICH}$}. The difference between the patterns is thus most pronounced when it comes to the type of entity which gets further described by way of an RC. In spoken language, there the NP dominating the head noun refers to a concrete thing that—considering the morphological information on the relativizer—is likely to be human (but at least animate). In contrast, there is a tendency to use finite subject RC to attribute some set of properties to non-human, abstract entities. Also, the embedding MC tends to be more complex in written discourse than in the spoken discourse. Interestingly, the top-ranked spoken pattern (in the red circle) is not member of the spoken-group just described but is judged by the clustering method to be somewhat in between.[24] The main difference of this pattern to the other ones discussed is its definite head, which could either mean that the NP is quantified by a definite determiner or that the head is a proper name. To the extent that the head-slot is realized by proper names, we may take this pattern to reflect that spoken discourse shows a greater amount of non-defining RCs as proper names are typically sufficient to enable the comprehender to fix reference. As the ratio of definite description to proper names is not revealed to us here, this last hypothesis is rather speculative at this point. The patterns labeled as *C.s9* and *C.s10* (in the small double-lined box) represent both RCCs which are headed by a definite pronoun, say *something* or *somebody*. These patterns represent also the improvement of the hCFA approach to the full CFA approach used earlier as RCC types with pronominal heads could not be detected with the latter technique. Another configuration that would have escaped our attention is the written pattern labeled *C.w5* (in the blue circle). This patterns is the only one characterized by a center embedded RC.

[24] Referring to that type as the top-ranked pattern may be a little misleading as the conciseness of that pattern (Q) is only minimally greater than that of the next three patterns. The top-ranked pattern thus does not enjoy a very privileged position among the other top-ranked typed.

In summation, the application of association rule mining, CFA and hCFA has allowed us to detect a set of robust characteristics of finite subject RCC and the distributional differences of their preferred sub-types across modalities. We have observed that finite subject RCC in written discourse written discourse tend to be more complex than their cousins from the spoken domain: they are usually headed by an indefinite lexical NP that refers to an abstract entity. Its clausal constituents are characterized by relatively higher valency values, and among the entrenched type are even some that incorporate additional modifiers (AP, PP) or exhibit a center embedded RC (even though the combination of these features is not attested in any statistically significant pattern). The spoken types can be grouped into two groups: one group is characterized by indefinite lexical NP postmodified by a simple right embedded RC that refers to an animate entity. The second group exhibits a similar feature set but is endowed with a definite pronominal head (e.g. *somebody*, *something*).

While these findings corroborate the idea that the processing limitations impose certain restrictions on what type of pattern can become routinized and used in spoken language whereas no such constraints are binding for the written modality, we will make an attempt to explain (some of) the patterning in terms of the discourse functions that have been suggested in the literature for a number of RCC types. Some of these functions may very well be more typical of genres associated with spoken discourse while others are characteristic for genres that dominate written language. But before we turn to such form-function mappings, let us continue our discussion with an analysis of non-subject RCC.

4.3.2 Finite non-subject RCCs

We are now approaching the final and configurationally most complex domain of interest, namely the domain of finite non-subject RCCs. As the class of finite non-subject RCC is associated with a well-studied variation phenomenon—the omission of non-obligatory relativizers—we will widen our scope again and use our patterns detection techniques with different sets of variables. The set of variables used in the characterization of finite non-subject RCCs includes the factors shown in Table 26.

Table 26: Descriptors for finite non-subject RCC

Factor	Levels
type of RC subject	lexical/pronominal
definiteness of the head	definite/indefinite
animacy of the head	concrete/abstract
contentfulness of the head	general/specific
Presence of a uniqueness adjective	present/absent
Embedding	right/center
Medium	spoken/written
Relativizer	present/absent

The data set used for the analysis comprises of 329 constructions. The size of the data set comes with a limit of the number of variable that should be included in the design. Recall that in order to be able to run a thorough exhaustive CFA we should—as a rule of thumb—have as many as $N = 5 * 2^d = 5 * 2^9 = 2560$ cases. Given the fact that our data set is considerably smaller, the goal cannot be an exhaustive analysis. Rather what we are aiming at is a subset of types that can be detected even on the basis of this limited data set and which hence corresponds to the subset of constructions that we assume is most deeply entrenched. Looking at these variables we recognize some "usual suspects", i.e. variables we have already made use of earlier (SYNTACTIC TYPE OF HEAD, DEFINITENESS OF THE HEAD, EMBEDDING). Some variables have been refined a little: Instead of investigating the concreteness of the (referent of the) head, we have now settled on a more specific variable, namely animacy. The feature + ANIMATE does of course imply + CONCRETE. However, no inferences about concreteness can be drawn if we observe – ANIMATE. Animacy has been shown to modulate processing difficulty for finite non-subject RCCs (e.g. Mak et al. 2002, Pu 2007; cf. § 3.2). Also, instead of asking whether or not there is any additional modification of the head noun (viz. in addition to the RC), we now ask for a specific type of modification, namely one that presupposes the uniqueness of the referent in the context of utterance (cf. Fox and Thompson 1990). An example of such a modifier would be an AdjP like *only* in $_{NP}$[*the only man* $_{RC}$[…]]. As finite non-subject RCC introduce an additional referring expression in the RC proper—

the subject constituent—, we will also have a look at the morphosyntactic realization of that NP (cf. § 3.5.4). Following the work on mental accessibility and its relation to linguistic form (Chafe 1976, Ariel 1990) and, more generally, standard assumptions about the givenness and salience of a referent (representation) in information structural accounts (e.g. Lambrecht 1994, Büring 1995, Erteschik-Shir 2007), we will work from the assumption that pronominal NPs are easier to process than lexical ones. More precisely we may say that the corresponding referent representation of a pronominal NP, i.e. the mental representations corresponding to the entity referred to by the linguistic expression, is easier to access (or activate) for one of two reasons: first and second person personal pronouns refer to speaker(s) and hearer(s) respectively, and the corresponding referent representations can safely be assumed to enjoy low activation thresholds and high degrees of salience in any conceivable context. Third person pronouns indicate low processing cost, because we can infer from the fact that their use has been sanctioned in the discourse—i.e. from the fact that the speaker has apparently used the form felicitously—that the referent representation has been accessed earlier in the discourse via either a definite description or a proper name (or some non-linguistic means such as pointing). So, we assume that—ceteris paribus—a clause with a pronominal subject is easier to process than a clause with subject that is realized by a lexical NP.

Finally, as we are interested in typical patternings across modalities and the conditions that are characteristic of relativizer omission, these two variables (MEDIUM, PRESENCE OF RELATIVIZER) have been added as well.

4.3.2.1 ARM: finite non-subject RCCs

Again we will start our analysis with an application of the association rule mining technique and subsequently approach the task of detecting entrenched patterns on the basis of (hierarchical) configural frequency analyses. The ARM technique was slightly modified so as to allow values for both MEDIUM and PRESENCE OF A RELATIVIZER on the right hand side of the rule, i.e. on the consequent side, while disallowing them on the antecedent side.

Using these specifications of rule form and, again, the default values on rule detection described earlier, the procedure outputs 40 rules that satisfy the constraints. We will present

these results abbreviated as Table 27 (the complete output with all the relevant statistics has been moved to the appendix).

Table 27: Association rules detected for finite non-subject RCCs

Rank (by leverage)	written	spoken	present	absent
1	contentful head & no unique A	pronominal RC subject	contentful head & no unique A	pronominal RC subject
2	lexical head & no unique A	pronominal RC subject & generic head	lexical head & not unique A	pronominal RC subject & generic head
3	contentful head	generic head	contentful head	generic head
4	lexical head & no unique A & center	pronominal head & pronominal subject	right & no unique A	pronominal RC subj & center
5	cententful head & no unique A & center	pronominal head	lexical head & indef head	pronominal RC subj & unique A
6	contentful head & center	unique A & pronominal subj RC	lexical head & right	pronominal RC subj & generic head & center
7	definite head & no unique A	unique A & pronominal subj RC	no unique A	unique A
8	lexical head	unique A	contentful head & indef head	pronominal head
9	no unique A & center		lexical head	center
10	no unique A		right	pronominal head & center
11	lexical head & indef head & center			indef head & center

Table 27 presents an overview of the 40 rules detected by the ARM technique. The columns denote the RHS of a rule while all cells in these columns specify a particular LHS. Hence, the first cell specifies the probabilistic rule { + CONTENTFUL HEAD & ¬ UNIQUENESS ADJECTIVE } → { WRITTEN }. That is to say that this rule associates the presence of a semantically rich head and the absence of a uniqueness adjective with written language. The rules are ordered by leverage so that for each column a high position in that column indicates a high rank of that rule in that group. If we inspect the rules in Table 27 more closely, we learn that the characteristics detected for spoken discourse are shared by the patterns exhibiting an omitted R element: the first three most interesting rules in both cases associate the respective group

with pronominal subjects of the relative clause, a semantically vague (generic) head and the conjunct of these properties. Both sets, i.e. the set of spoken RCCs and RCCs that have no overt R element, are further characterized by a tendency to prefer the presence of uniqueness adjectives, pronominal generic heads, and pronominal RC subjects. In addition to these properties, the R-omission set is further associated with center embedding and to some degree with indefinite heads. A similar correspondence can be observed for the set of written RCCs and the set of RCCs that exhibit an overt relativizer: both show a preference towards semantically richer (contentful), lexical heads and the absence of a uniqueness adjective. The only dimension where the sets differ noticeably concerns the associated types of embedding—with written discourse showing a general preference for center embedded RC and overt R patterns generally leaning towards right embedding.

All these results fit nicely into the picture drawn so far: given the restrictions of linguistic communication via the auditory channel (no external representation, only short term buffering, fast decay of memory traces), we would expect language users to employ a greater number of forms that facilitate processing in that modality. The co-occurrence of such forms can then be exploited in on-line processing: A form F_i frequently co-occurring with a subsequently perceived form F_j will eventually lead to F_i being treated as signaling F_j. The signal character of such pairs (or higher order tuples) allows for a better anticipation of yet unperceived material. Forms that strengthen the anticipatory character of language processing make linguistic communication easier, and are thus frequently employed. From a comprehension perspective, they allow for the faster recognition/comprehension of linguistic strings as it is easier to mentally access routinized patterns. They also make things easier for the producer as a routinized pattern allows faster mappings from thought to linguistic form and also require less effort to be spent on the phonological articulation of the constitutive forms (cf. the *phonetic reduction effect* mentioned in § 2.2.1). The ARM results have provided us with a set of rather simple but robust associations.

4.3.2.2 (H)CFA: finite non-subject RCCs

For the non-hierarchical configural frequency analysis the same set of 329 constructions was used. The first step in the analysis focused on the detection of maximally specified patterns.

The application of the CFA procedure resulted in an evaluation of 768 patterns, of which only seven were judged to be statistically significant. Table 28 presents their descriptions and the relevant statistics (the complete results from the CFA can be found in the appendix).

Table 28: Finite non-subject RCCs - Fully specified types

label	med	syn head	uni A	content	animate	definite	RC subj	R	emb	Freq	Exp	Q
c.s1	spo	prn.h	no	gen	no	no	Prn	no	rig	18	1.13	0.051
c.s2	spo	prn.h	no	gen	no	no	Prn	no	cen	16	0.73	0.047
c.s3	spo	lex.h	yes	gen	no	yes	Prn	no	cen	15	1.08	0.042
c.s4	spo	prn.h	no	gen	no	no	Prn	yes	rig	12	1.25	0.033
s.w1	wri	lex.h	no	con	no	yes	Lex	yes	cen	17	1.25	0.048
s.w2	wri	lex.h	no	con	no	no	Lex	yes	rig	14	1.3	0.038
s.w3	wri	lex.h	no	con	no	yes	Lex	yes	rig	14	1.94	0.037

The patterns listed in Table 28 are ordered by modality, revealing four significant spoken types and three written types, and subsequently by their conciseness (Q). We see that each of the types occurs roughly ten times more often than expected on the basis of chance (all $p_{\text{adjusted binomial}} < 0.001$). The first two spoken types in that list share all properties except the value for type of embedding. Their description tells us that they are characterized by the feature combination {INDEFINITE HEAD & PRONOMINAL HEAD & NO UNIQUENESS ADJECTIVE & GENERIC CONTENT & INANIMATE REFERENT}. Examples for each pattern are given in (80) and (81).

(80) That's **all I've done.**

(label: C.s1) [S1A-087 #073] # right embedded

(81) **All I'm saying** is we'll have might have to take a bit extra time to get there.

(label: C.s2) [S1A-100 #221] # center embedded

The examples exhibit the same head, *all*, but the feature list {–ANIMATE, –DEFINITE, +PRONOMINAL, -CONTENT} also allows certain usages of *anything*, *one*, and *something* as possible head realizations in this pattern. The pattern in fourth place (C.s4) is easily

recognized as a variant of the top two patterns just described and differs from C.s1 only in so far as it exhibits an overt relativizer. An example is given in (82):

(82) **It's something that** we were doing two or three weeks ago.
(label:C.s4) [S1B-013 #110]

The example in (82) is the archetype of a cleft (or cleaving) construction. Since the other spoken types are considered to be closely related to that pattern in form and function, we may start our discussion with some comments on the cleft-construction. Its form is characterized schematically in (83).

(83) [*It* + BE + FOCAL ELEMENT + subclause introduced by *that*|*who*|∅]

The cleft pattern can be used to give focal prominence to phrases and clausal constituents. It has been argued that a string '*It* BE NP {*that*|*who*|*which*} VP' is ambiguous between a proper (integrated) RCC and *it*-cleft construction. Huddleston and Pullum do in fact assign a different structure to *it*-clefts, which is given in Figure 86.

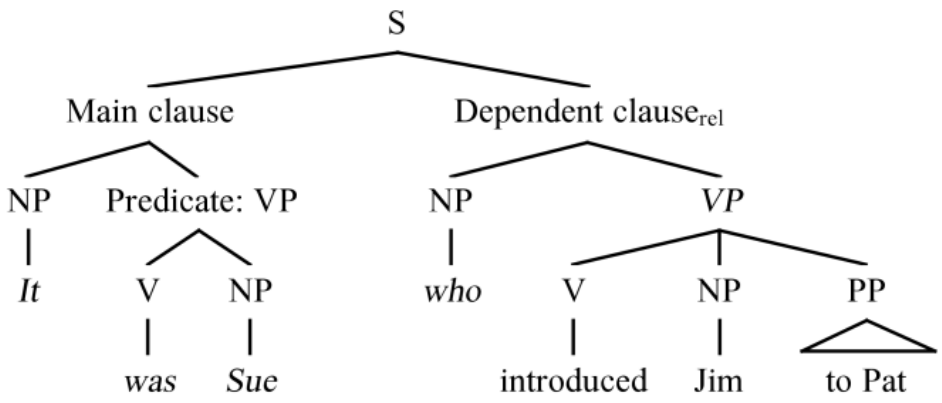

Figure 86: *It* cleft syntax (Huddleston & Pullum 2002)

Huddleston and Pullum argue that *Sue who introduced Jim to Pat* is not a constituent and that

it in these cases is not pronoun but a dummy subject. It should be pointed out that the idea to relegate *it* to being expletive and semantically inert in these contexts is an influential position (cf. Huddleston 1984, Lambrecht 2001), even though competing analyses have been proposed, which we need not discuss at this point (for a discussion cf. Akmajian 1970, Gundel 1977, Davidse, K, 2000). So, if the string *It was Sue who introduced Jim to Pat* were used to answer the question *Was it Sue or Mary who introduced Jim to Pat*?, it would count as an instance of a cleft-construction and would have the syntax given in Figure 86. In contrast, if it were used to answer *Who was that woman who just said 'Hi'?,* it would be considered an ordinary RCC, where the RC would be dominated by the main clause VP. This strikes me as odd (and is somehow reminiscent of the (tiresome) discussion about the syntactic representation of defining and non-defining RC). We will not attempt to provide the pro and cons of competing analyses here, nor will we try to disambiguate the respective patterns and divide them into cleft constructions and ordinary RCCs. Suffice it to say that even if a IT IS X THAT Y pattern is not a proper focus construction in the sense sketched above, it certainly is very similar (in fact formally indistinguishable) to that pattern and, hence, is likely to evoke some of the discourse-functional potential of these constructions. Schachter (1973) compares (restrictive) relative clause constructions to focus constructions and points out that these exhibit striking formal and functional similarities in many languages. So, let us return to our description of the pattern labeled C.s4. The focal element in this construction is universally realized by a pronominal expression which by itself contributes very little to the semantics of the referring expression. In fact, the expressions *something* or *all* are about as general in their semantics as linguistic expressions can possibly be and it is this semantic generality that in some sense signals an upcoming RC so that it can provide the missing semantic content. The semantic properties of cleft constructions are widely acknowledged and much can be said about their meaning and use (cf. Quirk et al 1985 for a quick overview). At this point I would like to draw attention to only some of these properties. The vagueness of the head expression requires that the descriptive feature required for successful reference resolution are provided by some other element that has a part in the complex referring expression, which raises the likelihood of a RC. Linguistic material that is highly probable in a particular context is easier to anticipate and hence easier to process

under real-time constraints. The only feature discriminating the pattern C.s1 from C.s4 is the explicit relativizer. As we have indicated earlier—and will also return to in § 4.3.3—, optional relativizers are by hypothesis produced if the construction currently being produced is somewhat difficult, i.e. they signal some kind of processing difficulty. Correspondingly, deeply entrenched patterns should not cause such troubles. We have further noted already that the limited data set makes it relatively hard for our methods to detect complex deeply entrenched types. Therefore, we may assume that those patterns that occur (highly) significantly more frequent than expected in spite of this limitation certainly count as deeply entrenched patterns. But if this is true, how come we find a pattern with an overt relativizer among the top four types? Notice that this pattern scores high in terms of its relative frequency (12 out of 230 investigated spoken cases are instances of that construction), it is about ten time more frequent than expected and judged to be highly significant ($p< 0.001$). The detection of a spoken type with an explicit R-element is in some sense surprising here as the general logic of the present work presents the occurrence of optional relativizers as an indication of low degrees of entrenchment of the corresponding pattern and thus as an indication of potential processing difficulty. The answer I would like to propose here to resolve this apparent conflict is grounded in the discourse function of the pattern and the recognizability of that function. The "*It is* X *that* Y"-pattern, an archetype of a focus construction, is more recognizable with an overt *that*. I would hence argue that *that* is retained to make the communicative point associated with cleft constructions more recognizable and that this appearance of *that* should in this case not be taken as a signal of potential processing difficulty. So, instead of viewing it as a true optional element we may say that in certain constructional environments, such as the one under discussion, an alleged non-obligatory *that* may actually serve a function, namely to reinforce the function associated with the cleaving syntax. We may further hold that the extensive use of this construction has led to deep cognitive routinization, which decreases the amount of structure of the expression (it becomes more unit-like). The example in (82) is thus an instance (of a particular

elaboration) of a RCC schema, which can be accessed as a whole. The *that*-relativizer is a salient component of that unit.[25]

A slight lexical variation of the schematic form given for (82) yields a pattern that can account for cases like (80). This variation is given as (84).

(84) [DEM PRN + be + FOCAL ELEMENT + subclause introduced by *that*|*who*|∅]

Diessel (2004) following Lambrecht (1988) has discussed this sub-construction of RCC, which is often referred to as a *presentational relative.* Like the cleft construction these constructions have the RC modify the predicate nominal of the MC but the main clause subject is typically a demonstrative pronoun (hence the label *presentational*). Diessel describes their function as establishing "a referent in focus position making it available for the predication expressed in the relative clause" (Diessel 2004:132). So, Diessel argues that it is the dependent clause that provides nearly all of the propositional content of the pattern. Diessel (2004) follows Lambrecht (1988) in characterizing the copular MC as 'propositionally empty'. In their view, a sentence like (a), *This is the sugar that goes here*, is a paraphrase of a sentence (b), *The sugar goes in here*. To call (a) a paraphrase of (b) effectively means that (a) and (b) mutually entail each other. This, however, seems incorrect. However, what is important here is that the RC predication is certainly heavier, i.e. informationally richer, than the MC predication. Once we acknowledge the general discourse pragmatic function of the pattern exemplified in (80) and (82), it appears natural to also look at the center embedded type in (81) from this perspective as well. For convenience of

[25] The arguably a little clumsy formulation "instance of an elaboration of a schema" is meant as an attempt to capture the hierarchical nature of schematic construction type. A maximally specified sentence like *That's all I've done* is of course still a type. An utterance of that type would be an instance of that type but it would also be an instance of more general types, say [DEM PRN]*'s all I've done, or* [DEM PRN]*'s all I* [IP], and so on.

exposition, let me repeat the example in (81) here as (85).

(85) **All I'm saying** is we'll have might have to take a bit extra time to get there.
(label: C.s2) [S1A-100 #221] # center embedded

A brief glimpse at (85) should suffice to acknowledge that it certainly is not the relative clause that carries the main predication. In fact, the construction is quite typical in its information structuring as it encodes the new (and asserted) information on the main clause VP, which is located at the end of the sentence. The early RC is very formulaic permitting almost no lexical variation at the head position and is hence best viewed as discourse marker, that is employed for rhetoric effect. We can help ourselves to a first explication of the function of this marker and postulate that it used to put focus on the proposition expressed by the MC while at the same time de-emphasizing any potential propositions that have been expressed so far by that speaker in the ongoing discourse on the topic at hand.

This brings us to our last construction in the quartet of identified spoken types, i.e. pattern C.s3. This type differs noticeably in its formal specification from the other spoken types: it is characterized by a definite lexical head that is pre-modified by a uniqueness adjective (recall that all other heads were pronominal). An example is given in (86).

(86) The only thing you could do is is is is is do something for money here […]
(label:C.s3) [S1A-035 #043]

However, like the other types its head expression—albeit lexical—is very general semantically. This and the center embedding make it similar to the type just discussed (type C.s2). In fact, a lexical head such as *thing* surely is general enough to prevent successful reference resolution on the basis of its descriptive content. The use of the definite article asserts uniqueness of the intended referent in the context of utterance, which is further

emphasized by the presence of the adjective *only*. Given these conflicting pieces of information—asserted uniqueness and extremely broad descriptive meaning of the nominal—, a post-modifier is necessary in order to give the hearer any sensible chance to fix reference. And relative clauses are the ideal means to do so. Qua being clausal constituents, RCs allow for a much more flexible and richer description of the referent's properties than any other possible type of postmodifier. A post-posed PP (or AdjP for that matter) can only express a sub-set of potentially applicable properties. The set of predicates expressible by PP (or AP) constituents is by necessity smaller than the set of properties that a clausal constituent can encode, as clauses can contain any number of PPs (and APs) but not vice versa. Consequently, a RC constituent is highly predictable at the point *thing* has been perceived. As the overall RCC type is so much more frequent than expected, we can assume that it too has received unit status making it easy to process. The discourse-functional relatedness to the other types can be disclosed, if we consider minimal pairs like (87) – (88):

(87) The only thing you could do is YP

(88) All you could do is YP

We may say that—in virtually all contexts of utterance—a speaker of (87) could also have used (88) instead (and vice versa) without risking any serious change in the perceived discourse function (let alone descriptive meaning). So, we may conclude and say that the detected types of RCC characteristic for spoken discourse are unified by certain discourse functional properties, namely to give focal or contrastive prominence to the entity described by the NP dominating the RC. The high relative frequencies and conciseness of the patterns mark both their privileged cognitive status and their principal discourse functional status. Note that the four top patterns account for 61/329 = 18.54 % of the total data and for 61/230 = 25.52 % of the spoken data.

So, what about the written types? Table 25 shows that there are indeed very noticeable differences in the preferred patterning across modalities. The example given in (89) presents

an instance of the most pronounced written pattern (C.w1):

(89) The matter [on which Mr Pitkin had required advice] was the estate.
(**label: C.w1**) [W2F-011 #038]

As the pattern C.w1 is not specified for a particular internal role, we should abstract away from the fact that the example in (89) exhibits pied-piping. The remaining two detected types are quite similar differing only in the type of embedding and the type of determiner used. What all three detected written types have in common is an overt relativizer and a lexical, inanimate, semantically rich head and the absence of uniqueness adjective. Interestingly all types exhibit lexical RC subjects. Again the results corroborate the general processing hypothesis as the written types are characterized by more difficult factor levels whenever there is a difference in preferred patterning.

Again, we can represent the similarity-based network of constructions on the basis of a hierarchical agglomerative cluster analysis (with with Euclidean distance and NJ amalgamation, cf. § 5.2.3). Figure 87 presents the resulting structure.

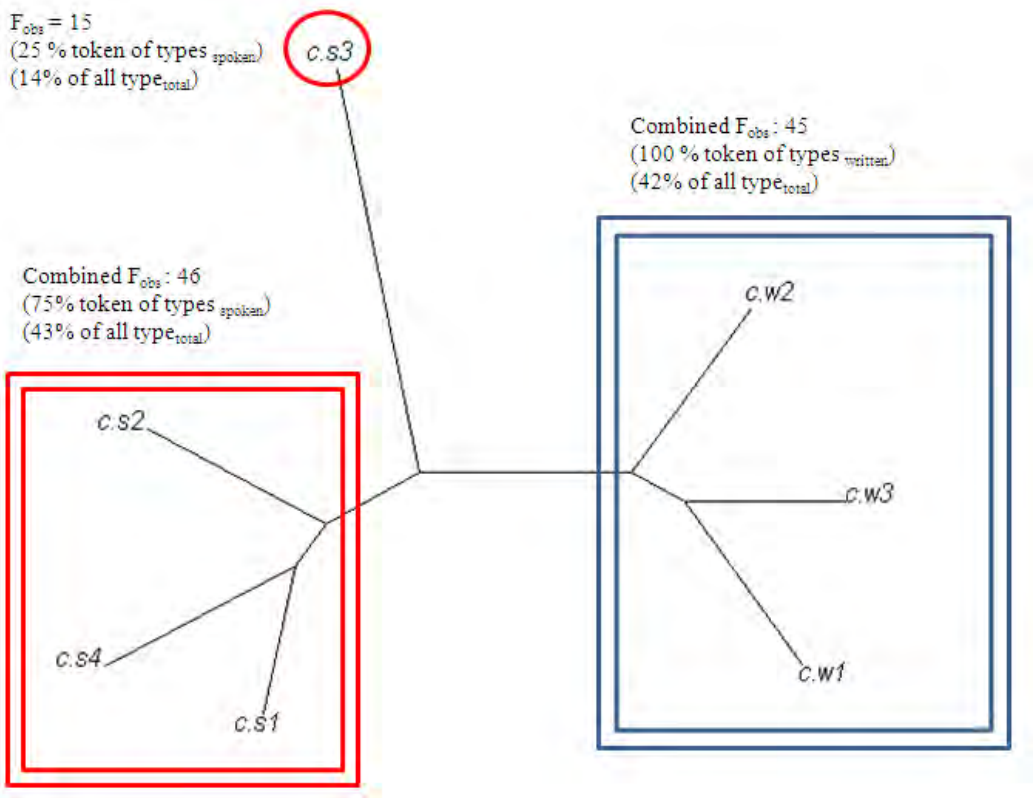

Figure 87: Dendrogram for fully specified types

The dendrogram in Figure 87 provides us with a convenient graphical display of the results just discussed: the types from both modalities cluster together quite nicely and in their alignment demonstrates the difference in preferred patterning in spoken and written discourse. Table 29 lists the patterns presented in Table 28 without their statistics and may help us interpret the tree.

Table 29: Top seven fully specified types (description)

label	med	syn head	uni A	content	animate	definite	RC subj	R	emb
c.s1	spo	prn.h	no	gen	no	no	Prn	no	rig
c.s2	spo	prn.h	no	gen	no	no	Prn	no	cen
c.s3	spo	lex.h	yes	gen	no	yes	Prn	no	cen
c.s4	spo	prn.h	no	gen	no	no	Prn	yes	rig
s.w1	wri	lex.h	no	con	no	yes	Lex	yes	cen
s.w2	wri	lex.h	no	con	no	no	Lex	yes	rig
s.w3	wri	lex.h	no	con	no	yes	Lex	yes	rig

We will refrain at this point from a more detailed discussion of the tree and postpone any conclusions that we may draw from it until we get to see the results from the hierarchical CFA. At this point, however, we may recall that following the logic behind exemplar based models discussed in § 2.3, we can now (at least roughly) estimate the expected processing difficulty of an arbitrary construction C_i, by assessing C's distance to the nearest (cluster of) entrenched pattern C_E in some n-dimensional state space S:

PROCESSING DIFFICULTY OF PATTERNS IN CONFIGURAL SPACE:

The smaller the distance between C_i and C_E in S,

the easier the processing of C_i

Applied to this example, we may thus assume that the closer a given patterns is to the patterns enclosed in the (red or blue) double-lined boxes, the more typical the pattern, and hence the easier its processing.

Having applied both the ARM technique and the CFA procedure—and having related their outputs via HACA—, we may now proceed just like we did in the analyses before and allow for a certain degree of variability in the pattern detection. This will enable us to see which factors might be less relevant for the construction of types. So again, we will extend our analysis by also computing the hierarchical variant of the CFA. The to-be-evaluated configurations were again restricted to those with a single variable slot so that each configuration is specified for seven properties and an indication of the donor modality (spoken/written discourse) resulting in (4096 + 1 =) 4097 patterns. Deleting all patterns whose observed frequency is smaller than or equal to the expected frequency effectively reduced the set to 453 constructions, of which 68 were statistically significant. Seven patterns were deleted because their value for medium was unspecified, thereby preventing any comparison of spoken and written discourse. Table 30 presents the remaining 61 types ordered by their degree of conciseness (a table with all statistics is given in the appendix).

Table 30: hCFA (one variable slot) finite non-subject RCCs

Label	syn head	uniA	content	animacy	definite	RC subj	R type	embed	Freq
C.wS1	lex.h	noA	con	ina.h	def.h	lex	yes	.	31
C.wS2	lex.h	noA	con	ina.h	.	lex	yes	rig	28
C.wS3	lex.h	noA	con	ina.h	.	lex	yes	cen	20
C.wS4	lex.h	noA	con	ina.h	def.h	.	yes	cen	22
C.wS5	lex.h	.	con	ina.h	def.h	lex	yes	cen	18
C.wS6	lex.h	noA	con	ina.h	def.h	lex	.	cen	18
C.wS7	lex.h	noA	con	.	def.h	lex	yes	cen	17
C.wS8	lex.h	noA	.	ina.h	def.h	lex	yes	cen	18
C.w1	lex.h	noA	con	ina.h	def.h	lex	yes	cen	17
C.wS9	.	noA	con	ina.h	def.h	lex	yes	cen	17
C.wS10	lex.h	noA	con	ina.h	ind.h	lex	yes	.	17
C.wS11	lex.h	noA	con	ina.h	ind.h	.	yes	rig	18
C.wS12	lex.h	noA	con	.	ind.h	lex	yes	rig	14
C.w2	lex.h	noA	con	ina.h	ind.h	lex	yes	rig	14
C.wS13	lex.h	.	con	ina.h	ind.h	lex	yes	rig	14
C.wS14	.	noA	con	ina.h	ind.h	lex	yes	rig	14
C.wS15	lex.h	noA	con	.	def.h	lex	yes	rig	14
C.w3	lex.h	noA	con	ina.h	def.h	lex	yes	rig	14
C.wS16	lex.h	noA	.	ina.h	ind.h	lex	yes	rig	14
C.wS17	lex.h	.	con	ina.h	def.h	lex	yes	rig	14
C.wS18	.	noA	con	ina.h	def.h	lex	yes	rig	14
C.wS19	lex.h	noA	con	ina.h	ind.h	lex	.	rig	14
C.wS20	lex.h	noA	con	ina.h	def.h	lex	.	rig	15
C.wS21	lex.h	noA	.	ina.h	def.h	lex	yes	rig	14
C.sS1	prn.h	noA	gen	ina.h	ind.h	prn	no	.	34
C.sS2	prn.h	noA	gen	ina.h	ind.h	prn	.	rig	30
C.sS3	lex.h	yesA	gen	ina.h	def.h	prn	no	.	21
C.sS4	lex.h	yesA	.	ina.h	def.h	prn	no	cen	20
C.sS5	prn.h	noA	gen	.	ind.h	prn	no	rig	18
C.sS6	prn.h	.	gen	ina.h	ind.h	prn	no	rig	18
C.s1	prn.h	noA	gen	ina.h	ind.h	prn	no	rig	18
C.sS7	prn.h	noA	gen	ina.h	ind.h	.	no	rig	18
C.sS8	prn.h	noA	gen	ina.h	.	prn	no	rig	19

C.sS9	lex.h	yesA	gen	ina.h	def.h	prn	.	cen	18
C.sS10	prn.h	noA	.	ina.h	ind.h	prn	no	rig	18
C.sS11	lex.h	.	gen	ina.h	def.h	prn	no	cen	21
C.s2	prn.h	noA	gen	ina.h	ind.h	prn	no	cen	16
C.sS12	prn.h	noA	gen	ina.h	ind.h	.	no	cen	16
C.sS13	prn.h	noA	gen	ina.h	.	prn	no	cen	17
C.sS14	prn.h	noA	gen	.	ind.h	prn	no	cen	16
C.sS15	prn.h	.	gen	ina.h	ind.h	prn	no	cen	16
C.sS16	lex.h	yesA	gen	.	def.h	prn	no	cen	16
C.sS17	.	yesA	gen	ina.h	def.h	prn	no	cen	16
C.sS18	prn.h	noA	gen	ina.h	ind.h	prn	.	cen	16
C.sS19	prn.h	noA	.	ina.h	ind.h	prn	no	cen	16
C.sS20	.	noA	gen	ina.h	ind.h	prn	no	rig	20
C.s3	lex.h	yesA	gen	ina.h	def.h	prn	no	cen	15
C.sS21	lex.h	yesA	gen	ina.h	def.h	.	no	cen	15
C.sS22	lex.h	yesA	gen	ina.h	.	prn	no	cen	15
C.sS23	.	noA	gen	ina.h	ind.h	prn	no	cen	17
C.sS24	prn.h	noA	gen	ina.h	ind.h	.	yes	rig	13
C.sS25	prn.h	noA	gen	.	ind.h	prn	yes	rig	12
C.s4	prn.h	noA	gen	ina.h	ind.h	prn	yes	rig	12
C.sS26	prn.h	.	gen	ina.h	ind.h	prn	yes	rig	12
C.sS27	lex.h	yesA	.	ina.h	def.h	prn	no	rig	14
C.sS28	prn.h	noA	gen	ina.h	ind.h	prn	yes	.	12
C.sS29	prn.h	noA	gen	ina.h	.	prn	yes	rig	13
C.sS30	prn.h	noA	.	ina.h	ind.h	prn	yes	rig	12
C.sS31	lex.h	.	gen	ani.h	def.h	prn	no	rig	7
C.sS32	.	noA	gen	ani.h	def.h	prn	no	rig	6
C.sS33	lex.h	noA	gen	ani.h	.	prn	no	rig	6

The discussion of these results is more sensible once they have been grouped by similarity. To that end a cluster analysis was run on the hCFA output. Again, all factor levels were translated into positive integers such that each arguable more difficult level was assigned a value of “3”, the simpler level was re-coded as “1” and all unspecified levels were assigned the value “2”. Again dissimilarity was expressed in terms of metric distance in Euclidean

space and the neighbor joining algorithm was used for the amalgamation. The dendrogram representing the structure detected for these data (on the basis of the just mentioned algorithmic choices) is given as Figure 88 (as the graphical representation of the unrooted tree structure shows a lot of overlapping in the labels of the terminal nodes, a rooted version is provided in the appendix).

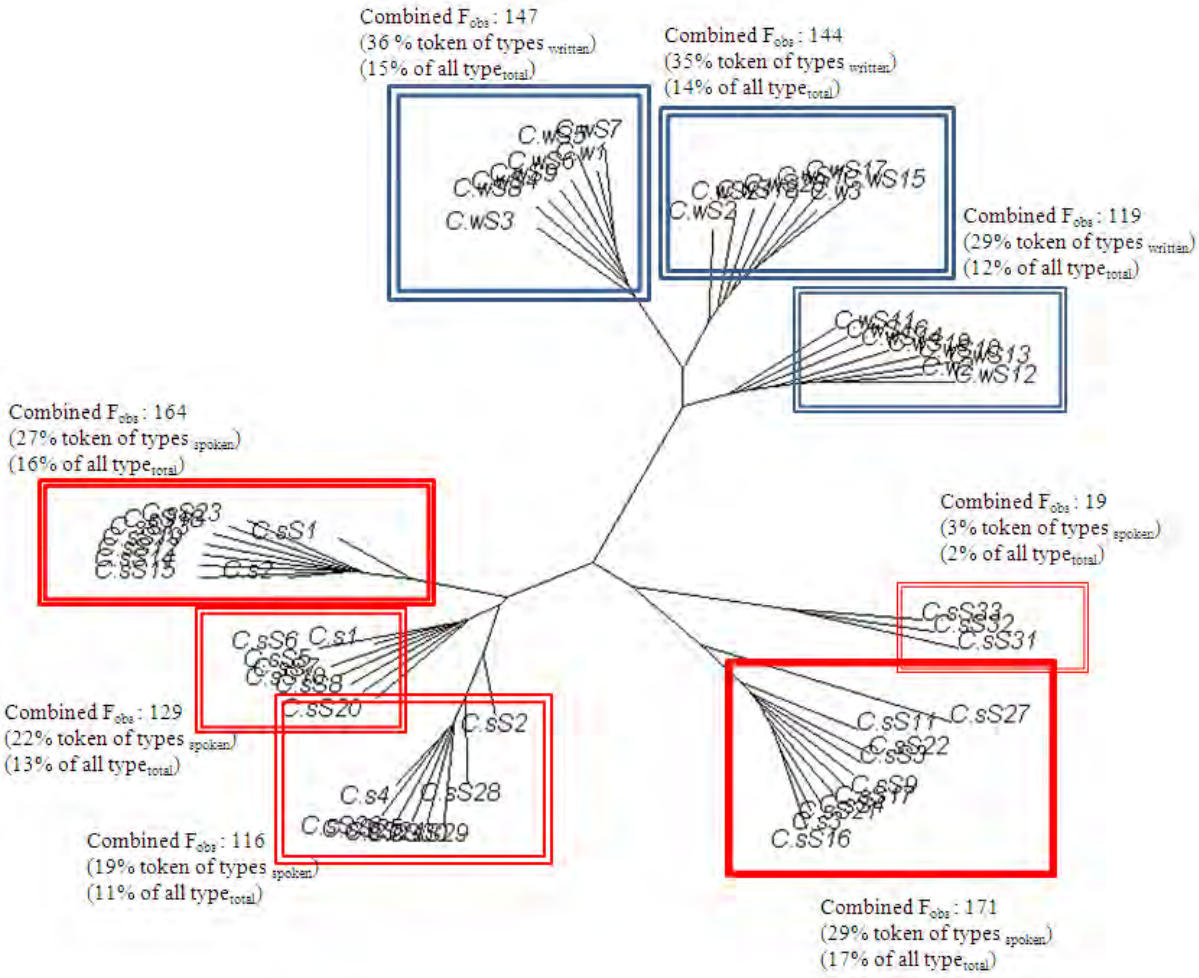

Figure 88: Dendrogram for variable types of finite non-subject RCC

The labeling of the types is similar to the format introduced earlier: "C.s1" refers to the top-ranked configuration from the spoken modality. The capitol 'S' in "C.sS1" denotes the

property of BEING SCHEMATIC, so that "C.sS1" labels the top-ranked schematic configuration from the spoken modality.[26] The colored box around a given cluster encodes two things: (a) the modality of the members of that group (red=spoken; blue=written) and also (b) the relative dominance of that cluster: the thicker the boundary, the stronger is that cluster supported by the data. This information is also given numerically for each cluster, i.e. for each group there is an indication of the combined frequency (F_{obs}) and frequencies relative to the set of spoken types and relative to the total number of types identified. Due to the fact that all but seven of the configurations are partially schematic, there is a certain amount of double counting involved. If a cluster contains both a fully specified configuration and one (or more) schematic ones, so that the former is a proper subset of the latter, the observed frequency given for that cluster would actually be greater than the number of actual cases attested in the data. This, however, was not considered problematic as there are no reasons to believe a priori that this will affect the outcome in any undesirable way.

The first thing to be observed is that there is a nice split between the spoken types and the written ones. There is not a single cluster that includes both spoken and written types. Secondly, the similarity between an arbitrary pair of clusters from a given modality is higher than any the similarity between any two cluster from different modalities (this can be read off from the length of the branches that we need to "travel along" to get from node A to node B). Thirdly, each fully specified pattern discussed before has its own cluster and groups around itself all sufficiently similar types. The strongest cluster (F_{obs} = 171) incorporates the type labeled C.s3 repeated here as (90)

(90) The only thing you could do is is is is is do something for money here […]
(label:C.s3) [S1A-035 #043]

[26] Of course, in some sense all the patterns discussed here are schematic as they are defined in terms of a set of rather abstract properties. However, the configurations labeled as schematic here exhibit one slot that is not specified and are thus more abstract than the fully specified patterns.

Recall that this pattern was in a way most distinct from the other types detected by the CFA due to its different head properties (definite lexical semantically vague head noun plus presence of uniqueness adjective). The move to de-emphasizing less important features by way of including more abstract patterns (CFA ⇒ hCFA) has changed the relative importance of that type from third to first place. This is a nice result because it fits very well into the results reported in previous research on the processing difficulty and its relation to *that* omission. We will discuss this in more detail in § 4.3.3.

The second strongest group includes C.s2 repeated here as (91).

(91) **All I'm saying** is we'll have might have to take a bit extra time to get there.

(label: C.s2) [S1A-100 #221] # center embedded

Again, the output of clustering technique is very encouraging as the features unifying the members of that group are really distinctive of the type C.s2 ({INDEFINITE PRONOMINAL HEAD & CENTER EMBEDDING}). The next two groups, i.e. those with combined frequencies of F_{obs} = 129 and F_{obs} = 116, are organized around the patterns C.s1 and C.s4 respectively, whose only difference is the absence or presence of an overt relativizer. Again the examples illustrating these patterns are repeated here as (92) and (93).

(92) **It's something that** we were doing two or three weeks ago.

(label:C.s4) [S1B-013 #110]

(93) That's **all I've done.**

(label: C.s1) [S1A-087 #073]

The examples we have picked to illustrate C.s1 and C.s4 differ in their lexical realization of head. The indefinite pronoun *something* is used together with the overt relativizer pattern,

while the covert variant is exemplified with *all* filling the head slot. Notice that this need not be the case as the respective patterns do not differ in their descriptions of the head slot. If we look more closely, i.e. if we cross the factors PRESENCE OF RELATIVIZER and HEAD TYPE for this subset, we can observe that these associations are not completely random but reflect a general tendency in the data. Table 31 gives data in tabular form.

Table 31: HEAD x RELATIVIZER for patterns C.s1 and C.s4

		relativizer		
		absent	present	
head	*all*	10	2	12
	anything	2	2	4
	one	1	1	2
	something	5	7	12
		18	12	30

Figure 89 presents the corresponding mosaic plot.

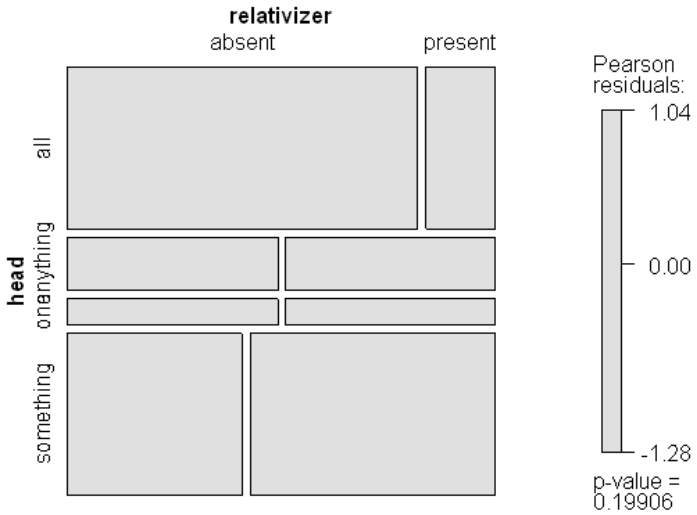

Figure 89: HEAD x RELATIVIZER for patterns C.s1 and C.s4

Even though the association is not significant ($\chi^2 = 4.653$, df = 3, p-value > 0.199, $p_{\text{Fisher Exact}}$ > 0.15), we observe that *all* occurs predominantly with prefer the zero variant, while

something is more balanced but leans slightly towards *that*.[27] If we consider all cases of finite non-subject RCCs from the spoken modality (n = 42), the preference for *all* towards zero is rather striking. Table 32 presents the distribution.

Table 32: HEAD x RELATIVIZER (all finite non-subject RCCs)

		relativizer		
		absent	present	
head	*all*	27	2	29
	something	5	8	13
		32	10	42

Focusing on the distributions of *all* and *something* across the two variants yields a clearer picture (χ^2 = 14.774, df = 1, p $_{\text{chi square}}$ < 0.00012, p $_{\text{Fisher exact}}$ < 0.00037). In the light of these results it seems plausible to assume that we are dealing with lexically specific constructions we may represent as shown in (94) and (95),

(94) (PRN$_{\text{DEM}}$ | *It*) BE ***all*** $_{\text{RC}}$[∅ NP$_{\text{PRN}}$ VP']

(95) (PRN$_{\text{DEM}}$ | *It*) BE ***something*** $_{\text{RC}}$[***that*** NP$_{\text{PRN}}$ VP']

The omitted R-element fits very well to the general hypothesis that optional relativizers are dropped when the RC is easy and/or predictable. But what is it that has *something* co-occur so often with an overt *that*? In our discussion of the patterns above, we have already suggested that it is the *it*-cleft schema that is responsible for the retained *that*. But why do we not observe a similar effect with *all*? The reason for this may lie in the semantics of these terms. The expression *something* may be better suited to occur in the X slot of the "*It is* X *that* Y"-pattern, because its semantics is in some sense more modest. That is to say, as "*It is*

[27] I we drop all *one* and *anything* cases to reduce the degrees of freedom the p = value drops below .1, but does still not reach conventional levels of significance.

something that S" imposes fewer constraints on the truth conditions associated with the underlying proposition than "*It is all that* S" does, it and can be used in a wider area of contexts, leading to a potentially stronger entrenchment of the form. Future research may assess the discourse-functional differences that may have arisen from this semantic difference.

We will continue our discussion of finite non-subject RCC with a closer look at the phenomenon of relativizer omission, which in light of the preceding discussion presents itself a natural focal point of attention.

4.3.3 Constructional schemas and relativizer omission

The characterization of finite non-subject RCCs also allows us to put the general model to a more local empirical test and focus on a grammatical phenomenon that has received a huge amount of attention in the literature: the omission of optional relativizers. The goal is to assess to what extent the pattern oriented view produces results that converge on what is usually observed in experimental studies. I hope to be able to show that the methodology proposed in this study is a fruitful corpus-linguistic complement to experimental techniques in the study of language processing.

Although we have already alluded to the topic of relativizer omission at various points in our discussion, it is convenient to start off with an example. Consider the pairs of sentences in (96) to (98):

(96) The man$_{i\ RC}$[$_{R}$[that|who|whom] you saw __$_{i}$ on the plane] carried a concealed explosive

(97) The man$_{i\ RC}$[∅ you saw __$_{i}$ on the plane] carried a concealed explosive

(98) The man$_{i\ RC}$[$_{R}$[that|who] __$_{i}$ walked into the cockpit] had a weird look in his eyes.

(99) *The man$_{i\ RC}$[∅ __$_{i}$ walked into the cockpit] had a weird look in his eyes.

As indicated by the gapping, we are looking at a pair of object relatives in (96) and (97) and a pair of subject relatives in (98) and (99). In English it is permissible to not produce a relativizing element—say *that*—with all non-subject relatives except genitives. In some varieties of English, the starred type in (99) is in fact grammatical as well (for a global synopsis of morphological and syntactic variation in English cf. Kortmann and Szmescanyi 2004). The literature on relativizer omission (and similarly complementizer omission) is vast and has uncovered a large number of factors that influence this variation in the English grammar (cf. Bolinger 1972, Elsness 1984, Ferreira and Dell. 2000, Gibson 1998, Grodner et al. 2002, Jaeger and Wasow 2005, Tottie 1995, Zwicky and Zwicky 1986, inter alia). Most accounts highlight the role of processing factors and in the attempt to account for this variation, prior research has identified a number of factors including the grammatical weight and length of the dependency domain (e.g. Gibson 1998, Hawkins 2004), conceptual accessibility (Bock and Warren 1985, Prat-Sala and Branigan 2000, Ariel 2001), ambiguity avoidance (Temperley 2003) and the interactants' attention to information flow (Fox and Thompson 1990). With few exceptions (e.g. Tottie 1995, Jaeger and Wasow 2005, Wiechmann 2007), most studies have focused on a single or very few factors and have hence little to say about the relative weight of these factors or possible interactions that may hold between these factors. In order to help remedy this situation, Wiechmann (2007) utilized binary logistic regression modeling to test a set 39 variables for their power in predicting relativizer omission in finite non-subject relatives (an overview of these factors and their leveling can be found in the appendix). The data set used in that study comprised (200 overt *R* + 200 *R*-less =) 400 usage events extracted from the spoken part of the ICE-GB, specifically from the subset of private direct conversations.[28] As 400 data points are of course

[28] In order to arrive at a data set of this size from this particular part of the ICE-GB did of course prevent that the data were cleaned in the fashion described in section 3.1. The direct conversation subset includes a total of 1660 RC, 651 of which exhibit an covert relativizer. However, the same subset must also be considered the ecologically most valid data set within that corpus so that all we should expect to observe effects as they occur in natural conversation (as opposed to effects observed in some artificial experimental context).

not sufficient to test the impact of some 39 variables, they were subdivided into smaller subsets and tested in an iterative fashion, so as to never permit more than five variables be entered into a single regression model. The set of factors tested was successively cycled with deletion of factors that did not result in significant model improvements until a 5-factor minimal adequate model was identified (cf. Crawley 2007: Ch. 9) for a detailed description of that approach to statistical modeling). The most powerful predictors reported in that study are given in Table 33:

Table 33: Best predictors of relativizer omission (Wiechmann 2007)

Factor	Levels
Definiteness of the head	definite/indefinite
Concreteness of the head	concrete/abstract
Contentfulness of the head	high/mid/low
Presence of a uniqueness adjective	present/absent
Accessibility of RC subject	high/mid/low

While the present approach purposely departs from the statistical approach opted for in that study (recall that a global assessment of factors entails the problematic assumption that a correlation, if attested, would be assumed to hold for all levels; cf. § 4.2.2), it agrees with that study in assuming that the identified variables are important determinants of the grammatical alternation. For the present analysis the number of factors was raised so as to include a total of 13 factors (including the factor PRESENCE OF RELATIVIZER). Table 34 presents the extended list.

Table 34: Factors used in CFA model

Factor	Levels
Definiteness of the head	definite/indefinite
Concreteness of the head	concrete/abstract
Animacy of the head	animate/inanimate
Contentfulness of the head	high/low
Type of head (~accessibility)	lexical/pronominal
Presence of a uniqueness adjective	present/absent

Type of RC subject (~accessibility)	lexical/pronominal
Definiteness of RC subject	definite/indefinite
Animacy of RC subject	animate/inanimate
Embedding	right/center
Formality of context	formal/informal
Medium	spoken/written
Relativizer	present/absent

Apart from an elaboration of investigated head features (i.e. the inclusion of animacy and morphosyntactic realization) and the expansion of the features encoded on the RC subject (animacy, definiteness, morphosyntactic type), the extended list also contains information about the type of embedding of the RC, as well as information about the register and the degree of formality. Finally, the last variable describes the presence or absence of the non-obligatory relativizer. The design certainly stretches the limits of what we can hope to discover on the basis of a mere 329 data points. A permutation of all factor levels yields 27,648 possible (fully-specified) patterns. With a type/token ratio like this, it is awfully hard—statistically speaking—to detect significant types at all. Interestingly, the CFA procedure still judged seven types to be (highly) statistically significant. These types are presented as Table 35.

Table 35: CFA - relativizer omission

		omi.c1	omi.c2	omi.c3	omi.c4	omi.c5	omi.c6	omi.c7
FACTORS	medium	spoken	spoken	spoken	spoken	spoken	written	written
	formality	informal	formal	informal	formal	formal	formal	formal
	head.type	h.prn	h.lex	h.prn	h.prn	h.prn	h.lex	h.lex
	uni.A	no.unique.A	unique.A	no.unique.A	no.unique.A	no.unique.A	no.unique.A	no.unique.A
	content.h	h.gen	h.gen	h.gen	h.gen	h.gen	h.con	h.con
	animacy.h	h.ina	h.ina	h.ina	h.ina	h.ina	h.ina	h.ina
	def.h	h.indef	h.def	h.indef	h.indef	h.indef	h.indef	h.def
	ms.rcs	RCs.prn	RCs.prn	RCs.prn	RCs.prn	RCs.prn	RCs.lex	RCs.lex
	ani.rcs	RCs.ani	RCs.ani	RCs.ani	RCs.ani	RCs.ani	RCs.ina	RCs.ina
	def.rcs	RCs.def	RCs.def	RCs.def	RCs.def	RCs.def	RCs.def	RCs.def
	conc.rcs	RCs.con	RCs.con	RCs.con	RCs.con	RCs.con	RCs.abs	RCs.abs
	embedding	right	center	center	center	right	right	center
	relativizer	absent	absent	absent	absent	present	present	present
STATISTICS	Freq	12	10	8	7	7	5	4
	Exp	0.3369	0.3428	0.2229	0.2284	0.3812	0.0131	0.0133
	Cont.chisq	403.7634	272.0581	271.3472	200.7643	114.9226	1898.41	1195.0208
	P.adj.bin	7.44E-11	1.10E-07	3.15E-06	0.000137235	0.00433987	8.63E-08	3.50E-05
	Dec	***	***	***	***	**	***	***
	Q	0.036	0.03	0.024	0.021	0.02	0.015	0.012

Table 35 presents the detected types as column vectors specifying the respective factor level combinations. The only reason why the data have been transposed here, and hence depart from the way the CFA results have been presented so far, is that this format requires less space. In terms of the information presented, it is equivalent to what we have seen in previous sections. So in the lower box, we find all the statistics associated with a given configuration, i.e. its observed (Freq) and expected frequency under H0 (Exp), its contribution to the Chi square sum(Cont.chisq), the adjusted p-value of the binomial test (P.adj.bin), the significance level (Dec), and the coefficient of determination (Q).

A good way to start our discussion of the results is having a look at the values for Q. To allow for an easier interpretation, Figure 90 presents the results graphically.

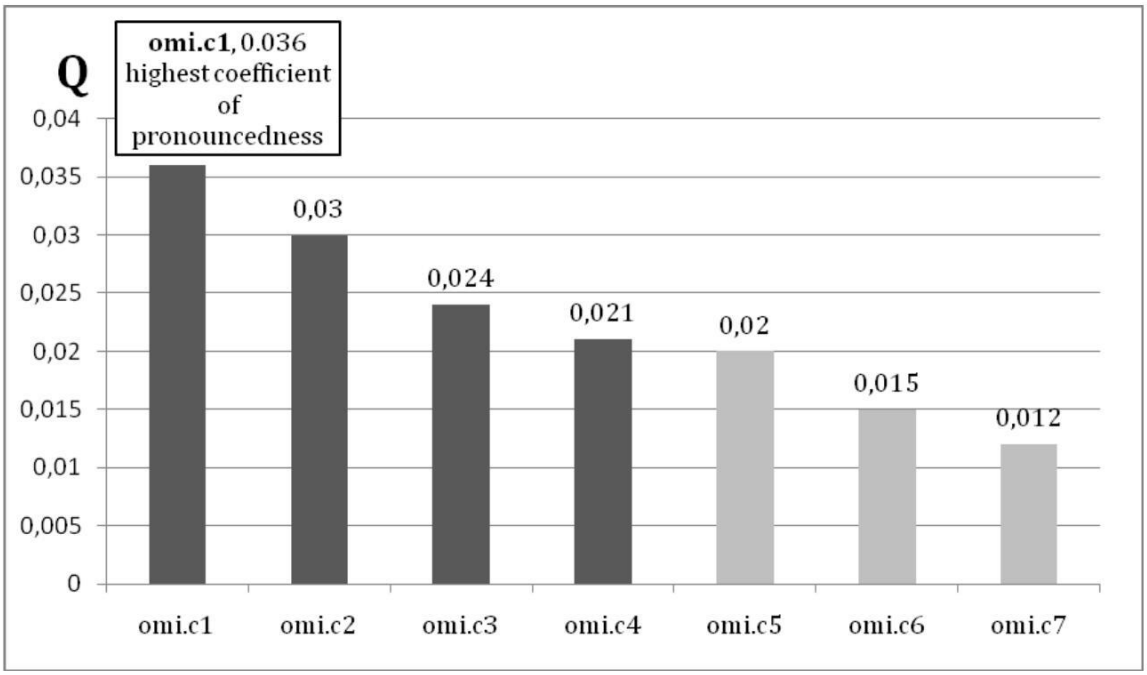

Figure 90: CFA results ordered by Q

We observe a smooth cline from the types labeled as *omi.c1* to *omi.c7* with the former being almost three times as pronounced as the latter. The shading of the bars indicates the value for relativizer: dark shading indicates the absence of a relativizer, dark shading indicates its presence. The strongest type associated with relativizer omission—*omi.c1*—is thus characterized by the feature combination {spoken, informal, h.prn, no.unique.A, h.gen, h.ina, h.indef, RCs.prn, RCs.ani, RCs.def, RCs.con, right, absent}. That is to say that non-obligatory R-elements tend to be omitted in informal spoken RCCs with a right embedded RC that exhibits a pronominal, indefinite head which denotes a general class of inanimate referent objects followed by a pronominal, definite RC subject that refers to a animate (and hence concrete) referent. A schematic representation certainly helps computing this information:

omi.c1: spoken discourse, informal contexts

syn:	[SUBJ	V	$NP_{prn,indef\ RC}$	[∅	$NP_{prn,\ def}$	VP_{RC}]]
sem	[ARG1	ACTION	$ARG2_{generic,\ conc,\ ani}$		[$ARG3_{conc,\ ani}$	$ACTION_{RC}$]

Instead of going through each of the patterns individually (and risking a considerable amount of repetition in our description), it is convenient to look at the results of the HACA technique that allows us to assess the underlying structural (dis)similarities. Figure 91 shows the annotated unrooted tree which represents the results of the clustering. All algorithmic choices remain as discussed in § 4.2.3.

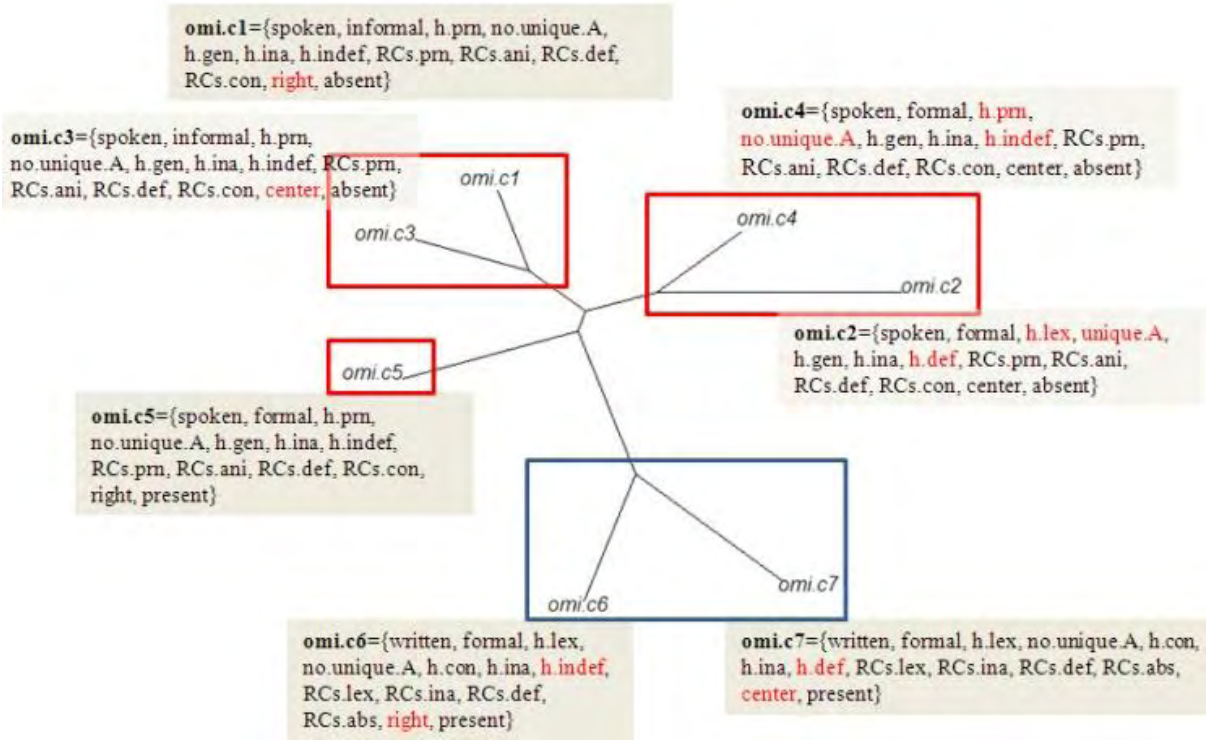

Figure 91: Results HACA of CFA types

Each node in the tree represents one of the seven configurations that occur significantly more often than expected under the assumption of statistical independence. For each node, we find a characterization in the form of a feature list. The boxes group together immediate neighbors in the constructional network and the coloring allows us to distinguish spoken types (red) from written ones (blue). Red coloring within a feature list indicates that the members of the corresponding group (=box) assume different values on that dimension. So, omi.c1 and omi.c3 are highly similar and are distinguished only by a single factor, namely the type of embedding. This high degree of similarity is reflected by the short path between the nodes.

The closest neighbor of omi.C1 can thus be represented schematically as.

omi.c3: spoken discourse, informal contexts

syn:	[$SUBJ_{prn,indef}$	[∅	$NP_{prn, def}$	VP_{RC}]]	VP_{MC}
sem	[$ARG1_{generic, inanimate}$	[	$ARG2_{conc, ani}$	$ACTION_{RC}$]	$ACTION_{MC}$

Table 35 tells us that these two patterns jointly account for (20 out of 44 ~) 45% of the instances of the spoken types. They also constitute 20 out of 500 instances of the spoken data, which means that one in 25 RCs in the spoken part in the complete sample is an instance of one of the two patterns. This is a striking number given that there are 27,648 possible factor level combinations that could have contributed instances to the sample.

If we look at the 20 instances of the two patterns, we can observe that in fact all instances of the center embedded variant have *all* as their lexical head realization, whereas the right embedded variants are a more variable and comprises of the set {all (n=8), anything (n=1), something (n=2), one (n=1)}, which is consistent with our considerations in the preceding section. For both patterns the transitivity value of the RC without exception is monotransitive, which strongly suggests that the internal role of the head without exception is that of DO, and both clausal constituents are in the active voice in all examples. Examples (100) – (105) are representative for the respective patterns:

(100) [All he'd want to do] was sit down and listen to this. [S1A.014#129:1:C]
(101) [All I had to do] was heat it up [S1A.020#290:1:C]
(102) […] [all I was saying originally] is that […] [S1A.037#139:1:B]

(103) That's [all I've done] [S1A.087#073:1:A]
(104) This is [something I still just occasionally wear] [S1A.022#224:1:D]
(105) You could have {anything you wanted] [S1A.021#151:1:C]

All these constructions are highly formulaic and they all serve specific discourse functions, which we have already discussed in the preceding section. Notice also that the verbs in the respective RCs are very common, high frequency verbs {*say, do, have, want, wear*} or light verb constructions {*have to do, want to do}*. This is characteristic of the detected patterns. Given the high frequency of the respective predicates and the generality of the heads, the resulting composite concepts expressed by the respective complex NPs are all good candidates for being lexical concepts, i.e. concepts for which there is a single word in a language. The concepts denoted by the strings *all I have done* or *all I have to do* are arguably frequently entertained in western cultures and so—following Zipf's considerations about the principle of least effort (Zipf 1949; cf. § 1.1.2)—, we should expect such concepts to be expressed with a minimum of linguistic material. At this point we can collect the fruits of assuming a sign-based view on grammar. From a construction grammar perspective, we are entitled to say that the forms in (100) to (105) are actually not too far away from this presumably appropriate minimal form. Once we have dismissed the ontological difference between words and (analytic) syntactic structures, we can appreciate the quasi word-like behavior of these RC-types and treat them as partially filled constructions. The omission of the relativizer makes the overall form of the complex NP shorter and this shortening can be seen as a iconic grammatical reflex of a high degree of coherence and unit status of the expression.

When we turn to the closest neighboring cluster, we observe that the pattern omi.c4 is nearly identical to omi.c3. The only feature distinguishing the two patterns refers to the level of formality of the donor discourse genre. We may thus conclude that the just described schematic constructions retain their productivity across genres. The last pattern, omi.C2, which is the second most frequent and also the second most pronounced pattern, is the type we have discussed in the preceding section as C.s3. It is characterized by a center embedded RC modifying a semantically vague lexical head as shown in (106) to (108).

(106) [The second thing [I want to say]] […]

[S1B.036#091:1:E]

(107) [The only way [you can make police forces genuinely accountable]] […] [S1B.033#032:1:D]

(108) [The best [she could hope for]] […]

[S1B.062#168:1:C]

The pattern in (106) is a straightforward example of the type we have already discusses in the preceding section. We have argued for the idea that while the expression types *the only thing* and *all* certainly are rather dissimilar from a morphosyntactic point of view, they are very similar in terms of their semantics (cf. § 4.3.2.2). The pattern is (107) is interesting because it presents strong evidence against the idea that relativization on deeper roles invariably results in greater processing difficulty. The RC in (107) does not even relativize on a core argument but on an adverbial role and yet it is among the most pronounced of all R-omission patterns. This is another example of why methods that look out for general tendencies may lead us to somewhat over-simplistic theories that misrepresent the item-specificity of human language processing. The example in (108) is particularly interesting because it fits very well to the characterization of the movement towards unit-status of frequently used complex NPs. One may argue that the example should not be counted as an instance of the pattern as it does not exhibit uniqueness adjective and a juxtaposed vague nominal head. However, the nominalised superlative form still retains some of its adjectival features as its denotation is only restricted by the quality of being (maximally) good. The speaker might have used an extra nominal (*the best thing* […]) to express the exact same thought. By not producing a possible generic head (thing) the speaker utilizes a form that is even shorter and is thus even more iconic.

The pattern omi.c5 seems a little odd at first glance as it meets all the typical characteristics of R-omission yet still retains the relativizer. However, it appears plausible that the relativizer is retained for reasons other than processing difficulty. All the instances of this construction can be viewed as cleaving structures and we have argued in the preceding section that the relativizer may be retained so as to make the discourse function of such patterns easier to recognize. This treatment may seem a little ad hoc though. It is a serious

problem of functional linguistic argumentation that it is always possible to present some post hoc explanation of why what has been found has in fact been found: This is due to the fact that functional explanations often make reference to antagonistic forces that act of a linguistic phenomenon (►competing motivations, cf., e.g., Kirby 1997). Consequently, a functional linguist can always "tell a story" about any conceivable finding by making reference to either motivation/force A or some antagonistic motivation/force B. As Newmeyer has discussed in considerable depth, it is a tell-tale sign of a weak theory if it still lends itself to a neat explanation, when we imagine a reversal of the facts (cf. Newmeyer 1999). The here proposed explanation of why certain entrenched pattern exhibit an overt relativizer may thus raise some concern from this direction. We can, however, shield us from such attacks and defend the idea that omi.c5 is in fact somewhat divergent by looking closely at the adjusted p-values of the types in Table 35. Consider Figure 92.

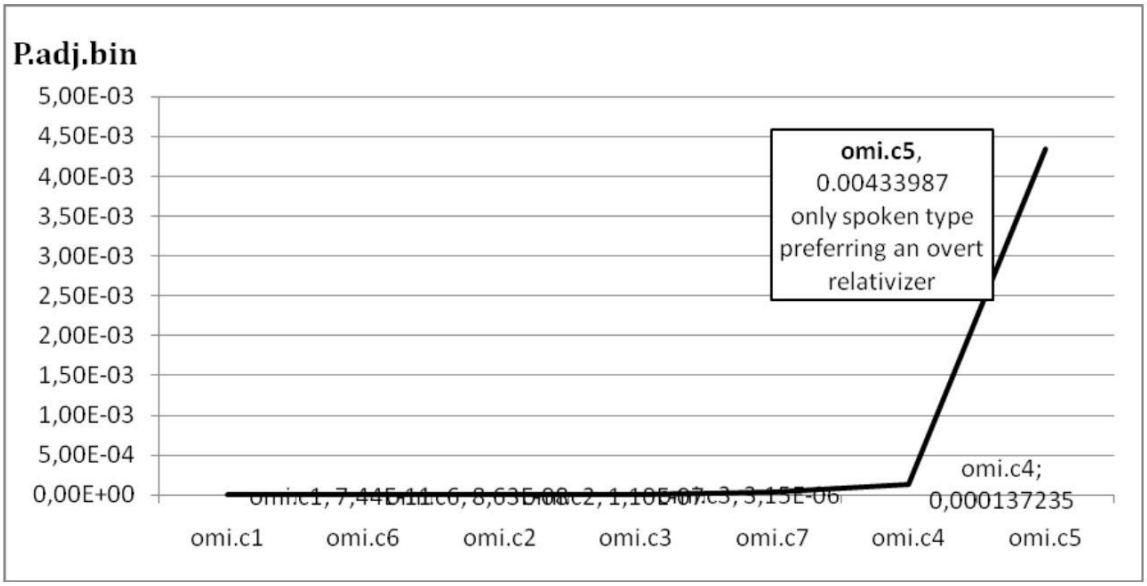

Figure 92: Adjusted p-values from binomial tests (R-omission patterns)

We can immediately see that the p-value of omi.c5 is considerably higher than all other patterns including the written ones, which even result from much smaller sample sizes. The considerably higher p-value tells us that there is considerably less evidence against H0 for this pattern than it is the case for all other detected types. While this difference of course does not provide evidence for the idea that that is retained for reason of recognizability of the cleft

structure (and its discourse function), it certainly shows that the pattern is different and the recognizability hypothesis is a possible explanation for divergent status.

This leaves us with the two remaining patterns that characterize overt relativizer patterns. Both patterns occur predominantly in formal written discourse and differ from the patterns discussed so far in that they exhibit lexical RC subjects that denote abstract entities. The respective RC heads are lexical (non-pronominal) and denote informationally richer (+contentful), abstract concepts. The key-difference between these two types lies in their attachment site, i.e. the type of embedding and the definiteness of the respective head. Consider the examples in (109) and (110).

(109) $_{NP}$[The extent $_{RC}$[to which their populations were new and transitory is apparent]] […]
[W2B-002 #068:2]

(110) He or she has $_{NP}$[no mechanism $_{RC}$[by which semantics could be taught as a sequence of instructions]] [W2A-032 #022:1]

The difference in definiteness is hardly surprising as it follows from general information structural constraints: center embedded RC modify the subject constituent which in languages like English typically encodes given (presupposed) information and given entities are usually (but not necessarily) expressed with definite NP (cf. Lyons 1999: 232). Conversely, we expect the second argument in a monotransitive pattern to express new information and hence expect it to be realized as an indefinite NP. Simply put, the definiteness contrast between the two written patterns is likely to be due to more general principles of information flow in English grammar and as such less relevant for our present discussion. What is interesting, however, is that 7/9 instances of the two constructions exhibit pied-piping. It has been argued that pied-piping reduces the processing demand of a pattern relative to the alternative of stranding the preposition, i.e. leaving the preposition in situ. Consider the examples in (111) to (114).

(111) He has $_{NP}$[no mechanism$_{i}$ $_{RC}$[$_{PP}$[by which$_{i}$] semantics could be taught __]]
(**pied piping**)

(112) He has $_{NP}$[no mechanism$_{i}$ $_{RC}$[which$_{i}$ semantics could be taught $_{PP}$[by [__$_{i}$]]]
(**stranding**)

(113) He has $_{NP}$[no mechanism$_{i}$ $_{RC}$[∅ semantics could be taught $_{PP}$[by [__$_{i}$]]]
(stranding with omission)

(114) *He has $_{NP}$[no mechanism$_{i}$ $_{RC}$[$_{PP}$[by ∅] semantics could be taught __]]
(pied piping with omission)

Stranded variants are usually considered more complex and hence harder to process presumably because it is easier to extract the whole PP than material from within that PP (for detailed proposals why this should be the case cf. Ross 1986, Roeper 2003, Hawkins 2004, Hoffmann 2005 and references therein). If we accept the hypothesis that pied-piping is in fact the easier of the two variants, then we may assume that the producer of (111) chose the simpler variant to minimize processing efforts, which in turn prevented the omission of the relativizer categorically. In such cases it seems questionable to consider such examples as involving optional relativizers in the first place. As both cases, stranding and pied piping, are highly infrequent in the present data set, we cannot examine the relationship between relativizer omission and preposition positioning any further. However the relationship may turn out to be exactly, it is likely to involve a lot of item-specificity. On the experience-based view presented here, certain cases of stranding are expected to be easier than their pied-piped counterparts. The contrast between *the guy I was talking to* and *the guy to whom I was talking* is a case in point. While a quick-and-dirty Google-query for the latter produces merely six hits, the former can be matched about 31,700 times. So, if frequency has anything to do with processing difficulty at all, stranding does not seem to be more difficult in all cases.

In summation, we may say that our pattern detection was very successful in identifying typical conditions of R-omission, but was not really able to identify clear patterns that preferred an overt relativizer. At least, it could not identify typical patterns that make the production of a truly optional relativizer more likely. In one of two cases, the overt relativizer cannot be considered optional (pied-piping); in the second it was argued that the overt R-element is produced to enhance the recognizability of a cleft structure and its associated

discourse function. This can be seen as a result of the relatively small data set. In order to determine exactly what conditions make an optional relativizer more likely, the CFA would require much larger data sets so as to be able to detect antitypes, i.e. configurations that occur significantly less frequent than expected. However in light of the fact that the method was capable of detecting numerous scenarios that make R-omission highly probable, we have every reason to book the results on the credit and not the debit side.

With all the results in place, we may now ask if there is a unifying explanation for all these findings. In the present account, it is presumed that non-obligatory grammatical material is dropped if the donor construction is frequent enough to lose part of its structure so as to assume a more unit-like form. This is true for a number of center embedded RC type (e.g. *All you need to do is* X -> Function: guiding focus/attention)) and also certain right embedded one (e.g. *Yesterday I met the guy you like so much* -> anchor new discourse participants).

An influential processing related account of relativizer omission has come to be known as the *predictability hypothesis* (Jaeger et al. 2005, Wasow et al. to appear), which we have in fact alluded to at numerous times in our discussion.

PREDICTABILITY HYPOTHESIS

> "*In environments where an NSRC* (finite **n**on-**s**ubject **r**elative **c**lause: DW) *is more predictable, relativizers are less frequent*"
>
> Wasow et al. (to appear: 5)

The term *environment* here denotes the linguistic material of the complex NP that precedes the RC, i.e. potential determiners and adjectives as well as the head noun. The predictability of a NSRC in an environment of one of these words is measured by "the percentage of the NPs containing that word that also are modified by an NSRC" (Wasow et al. to appear: 5). This particular measure of predictability may be a little crude, but the important idea is that optional relativizers (and by extension non-obligatory elements that do not result in a change

of meaning of the overall construction) are dropped, if the relative clause is highly predictable. Wasow and colleagues argue that the production of an optional relativizer facilitates processing in both directions: it facilitates sentence production because it provides the speaker with extra processing time (cf. Race and MacDonald 2003) and it facilitates sentence comprehension because the relativizer signals an upcoming clausal constituent to the parser (cf. e.g. Hawkins 2004).

The present account is very sympathetic to the predictiveness hypothesis (PH), at least to the general rationale of the way it portrays the impact of processing preferences on the organization of linguistic structures. However, as it stands, it appears to be unclear whether it should be viewed as a principle of on-line processing or if we should take it to describe a process of language change, or maybe even both. The best way to incorporate the idea behind the PH into the view presented here is portray it as a *as-if* description of a process that leads to the shaping of grammar over historical time (as opposed to a description of what is going on in on-line processing at a computational level). The reason for allocating the causal powers of the PH at this level is grounded in the idea that language processing here is presented as being rather holistic (and correspondingly less compositional) in nature. It has been argued that complex sentential constructions, bi-clausal RCCs, are bearers of rather specific discourse-functions. It is assumed that these discourse-functions are known to the interlocutors allowing language users to employ them as tools in communicative situations. Processing these complex patterns has been characterized as accessing high-level schematic constructions (complex units, routines) from memory. This picture de-emphasizes the role of low-level on-line composition. The PH portrays the production of *that* as an ad hoc choice made by the system on the basis of the difficulty of the sentence. In the here advocated view, the production of *that* is rather dependent on whether or not that is part of the constructional template (schema) that the speaker wishes to employ in the discourse. By viewing *that*-omission this way, we can account for the frequent use of patterns exemplified in (92) and repeated here as (115).

(115) **It's something that** we were doing two or three weeks ago.

As we have argued at various occasions earlier, there is no reason to believe that sentence types like the one underlying (115) are particularly hard to process thus requiring extra production time and/or explicit parsing instructions. In fact, when we look at the underlying factor levels of the RCC descriptor variables (e.g. type of embedding, type of RC subject, head type, etc.), it rather seems to be the case that we are dealing with the simplest of conceivable patterns. So why would an overt *that* figure in such a pattern, which occurs with such high above chance-level frequency? Furthermore, we should note that the indefinite head, *something*, which we have shown to be characteristic of the pattern, certainly adds to the predictability of the RC and the PH clearly predicts *that* omission here. In consequence, the PH—if viewed as an on-line principle—is unable to predict the empirical reality of the frequency of the construction.

Another thing that we may want to address in the context of the predictability hypothesis concerns the apparent rational behavior of human linguistic choices. Theories of rational behavior, like Game Theory, Decision Theory, and Information Theory, have attracted researchers of language for many decades and have had important impacts on the philosophy of language, e.g. the notion of conversational implicature (Grice 1969, Levinson 2000) and linguistic pragmatics, e.g. the notion of *relevance* (Sperber and Wilson 1995, Merin 1997). The general idea is that language use like any other kind of human behavior can be explained in terms of rationality (Searle 2003 presents an approachable introduction to the technical philosophy behind this issue). The rational behavior underlying the PH is that speakers use non-obligatory material if it is helpful in the processing of the overall construction, which can be viewed as a means to optimize cost/benefit-ratios during on-line processing. While it may play some role in on-line processing, I believe it is more adequate to assume that such principles are operative at the level of conventionalization of linguistic forms, i.e. at the level of language change over historical time. Following Croft (2008), we may presume that it is evolutionary processes that determine the constructional repertoire of a language. Such evolutionary processes do not require more than a tiny processing advantage to develop more adaptive forms over time. The exact nature of the relation between on-line

processing and language change remains a fascinating puzzle to be solved.

Apparent rational choices like these have led to the postulation of various default strategies. One of these maxims, which also pertains to relative clauses, is the idea of an *ambiguity avoidance*-principle (cf. Temperley 2003). While this certainly sounds like a good strategy to apply in communicative contexts, both experimental (e.g. Arnold et al. 2004) and corpus-based studies (Roland et al. 2006) rather suggests that humans do not behave in accordance with such maxims during on-line processing. But other recent proposals that employ the idea of rational behavior in message formulation have been supported by empirical evidence. One of these proposals that directly pertains to relativizer omission is Roger Levy and Florian Jaeger's *Uniform Information Density* (UID) hypothesis (Levy and Jaeger 2007). The UID hypothesis states that speakers prefer choices that avoid peaks and troughs in the information transmitted per unit, where information of an event is defined in its basic information theoretic sense, i.e. as the negative log of the probability of the event. Information density then is the amount of information per unit comprising the utterance. Applied to relativizer omission the UID-principle predicts that speakers insert *that* when the first word in the RC would be high in information density (e.g. *I*/*you*) , whereas they would omit the optional formal when the first word is low in information density (e.g. a proper name). Jaeger and Levy (2007) report a self-paced reading experiment conducted to test the processing related predictions of UID by comparing the actual distribution of *that* against its impact on processing complexity. The experiment was designed as follows: 24 representative relative clause constructions were extracted from the Wall Street Journal Corpus. Twelve of these patterns contained typical overt that RCs while the remaining twelve cases exemplified typical R-omitted types. To test speakers' reactions to changes in constructional choice the reading times of the actual choices in the corpus in (116) and (117) were compared to their grammatically permitted but less natural alternative in (118) and (119), respectively.

(116) The way $_{RC}$[**that** we've been managing …] (originally with that)

(117) The ball $_{RC}$[∅ he hit] wasn't a strike. (originally without that)

(118) The way $_{RC}$[∅ we've been managing …] (originally with that)

(119) The ball $_{RC}$[**that** he hit] wasn't a strike. (originally without that)

The analysis of the reading time behavior corroborated the UID hypothesis: despite the fact that an overt relativizer always improves processing, speakers tend produce that only when the RC onset would exhibit a high information density. Future research may assess the relationship of the UID hypothesis and the entrenchment hypothesis guiding the present account.

4.4 General discussion and concluding remarks

This work has presented a corpus-based approach to the processing of complex constructions, specifically English relative clause constructions. At the most general level, the goal of the investigation was to help bridge the gap between linguistic and psycholinguistic research. The concept that carried most of the theoretical weight in this attempt was the idea that the processing difficulty of a linguistic structure above all is determined by the degree of entrenchment of a pattern and thus by the experience that language users have with that (and similar) pattern(s). For lexical units, frequency effects have long been acknowledged, but the received view in much of psycholinguistics is still characterized by completely different conceptions of lexis and grammar. In the theoretical part of this paper I have thus tried to motivate a constructionist view on language, which entails that linguistic knowledge is best described as a large assembly of symbolic structures, i.e. constructions, of various degree of specificity. A huge part of this knowledge is made up of schematic constructions, i.e. abstractions or generalizations of recurrent experiences of linguistic events that share certain formal and functional properties. This conception of linguistic knowledge was tied to an exemplar-based view on language representation and processing, which promises to provide an appropriate psychological underpinning of that conception. Processing a linguistic structure in this view is classifying a number of constituent units that jointly allow for the categorization of a complex construction type, say a sentence. A sentence-level construction, e.g. an RCC, is a natural unit in this process as it constitutes the element in the discourse that

expresses a complete thought—in the sense of Frege. On the basis of an analysis of the usage patterns of biclausal relative clause constructions, it was argued that the frequency of a pattern in the ambient language—and hence its frequency in a language user's linguistic experience—is closely tied to the functional role that the form plays in the language in question, i.e. how the form is used in linguistic communication. Drawing on ideas proposed in research into discourse structure and interlocutor's attention to information flow, it was argued that it is the utility of a form in the discourse that ultimately determines its frequency. The frequency of a construction E can thus be viewed as an expression of the degree of familiarity a language user has with E and it is this familiarity that accounts for the relative difficulties that has been observed in experimental studies. Consequently, the explanation of processing difficulty is conceived of as being only indirectly related to the intrinsic structural properties of E, say the position of an RC-structure in the syntactic tree or the relativized role within the RC. This is to say that in the causal chain of explanation proposed here, the observable correlation between structural complexity and processing ease is interpreted in a manner that departs from more traditional treatments. Mechanistically speaking, the most direct cause for the difficulty of E is identified with the ease of mental activation of the corresponding cognitive routine. Deeply entrenched units (of arbitrary complexity and schematicity) are easy to activate (access/categorize) due to principles of routinization (§ 2.2.1). Once we accept the idea that frequency information—in some sense—is all we need in order to felicitously predict (and hence in some sense explain) human language processing behavior, we need to explain why certain patterns occur with above-chance probabilities in the ambient language. In other words, why is it that speakers habituate themselves to particular RCC-patterns and not others? Or more generally: why do speakers conventionalize schematic constructions of particular types and not others? The answers to such questions are probably quite multifaceted, but it seems plausible that certain types are dominant simply because the discourse functions they encode are prominent in a given genre. These discourse-functions include anchoring new referents into the discourse (-> object relatives and transitive subject relatives; Prince 1981, Fox 1987, Fox & Thompson), marking focus (-> cleft-like relatives, Schachter 1973), channeling attention (-> presentational relative; Diessel 2004), or adding an iconically shaped secondary predication to the main clause predication (-

> center embedded *–ed* participial RC, informally discussed in Granger 1983). The intrinsic properties of a construction type are viewed as being relevant for determining which forms are preferred (in a particular context) in the fulfillment of the associated function. The expression of thought via language involves speakers having to make certain formal choices. Linguistic communication involves the utilization of mappings from conceptual structure to formal structure and often there is more than one way of mapping language to thought. Such situations require that speakers weigh certain competing motivations, most prominently factors pertaining to the complexity of E and its degree of explicitness. This is arguably true both at the level of individual utterances, which are subject to architectural constraints from language production, and also at the social level, which embodies processes of conventionalization. It is at this point that ideas like Hawkins *performance-grammar correspondence* hypothesis figure in the explanation of on-line processing behavior (Hawkins 2004, but cf. also Dahl 2004).

We may sketch the explanation proposed here as depicted in Figure 93:

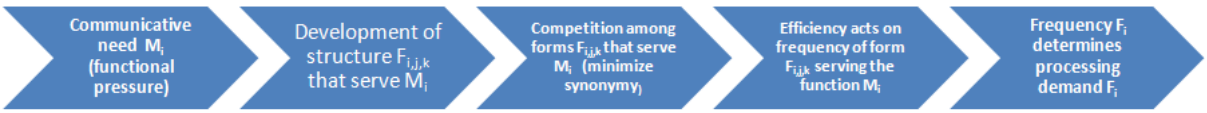

Figure 93: Communicative need, frequency, and processing difficulty

The starting point in the chain of explanation depicted in Figure 93 is the hypothesis that there is a functional pressure to develop certain communicative tools. This is in agreement with a *belief-desire-intention* model of human psychology, i.e. a particular stance that we can assume to explain human behavior in reference to the agent's future directed intention and thus certain actualized beliefs and desires (cf. Dennett 1981). At certain points in the (continuous) development of a language system, we may find a set of structures {F_i, F_j, …, F_k} competing for application in the expression of a particular thought, or satisfaction of a particular function. I presume that a great deal of explaining why certain forms survive while

others eventually die out can be expected from research into evolutionary processes in language change (Christiansen and Kirby 2003, Hurford 2007, Kirby 2007, Croft 2008 inter alia). Evolutionary processes may be held responsible for a principle we may call *minimize synonymy* (cf. Bolinger 1968, Goldberg 1995) and eventually lead to an optimal repertoire of formal means to satisfy the social/communicative needs. Forms with the right characteristics will be used most frequently and this enhanced frequency directly impacts processing difficulty.

In the present view, processing a natural language sentence is activating a series of representations that jointly constitute a cognitive routine, which as a whole bears a particular (albeit rather general) function. From this view it is natural to think not only of words but also of sentence types as tools that can be used to bring about certain communicative effects. Learning a language in this view is acquiring a toolkit that can be employed for social purposes, viz. act on other people's mental states. In any given domain (register > genre > situation), some functions will be more prominent than others and those tools that serve these functions will thus be employed more often than other tools. Over time language users will develop a remarkable proficiency in using these tools and they will excel in using those tools that they have worked with most frequently.

The more traditional alternative conception, which in fact must be considered the default view, has its roots in the very beginnings of cognitive science and can be traced back to Miller's seminal paper on the on the *magical number seven* (Miller 1956). Miller proposed that human short term memory was a resource limited storage device with a capacity to store 7 +/- 2 units at any given time. The size of these units could change so that larger structures eventually could count as units in this metric by way of a process termed *chunking*. Inspired by these ideas, a huge amount of research into adult language comprehension, specifically into the processing of relative clauses, has tried to account for the observed processing difficulties by way of assessing the memory demands of a given structure (Chomsky and Miller 1963, Wanner and Maratsos 1978; Ford 1983, Gibson 1998, Just and Carpenter 1992, inter alia). More complex structures, say center embedded structures or relativization on lower roles, are considered to consume more resources than simple structures, thereby

imposing greater demands on short term memory. We may illustrate this on the basis of the difficulty observed for object relatives (relative to subject relatives).

(120) The reporter that __ attacked the senator admitted the error.

(121) The reporter that the senator attacked __ admitted the error.

The typical empirical finding with sentences like these is that object relatives are harder to process. Disagreement in the literature typically arises when it comes to pinpoint the exact nature of the increase in complexity. The following hypotheses can be seen as representatives of the resource-based approaches that have been proposed to account for the processing asymmetry of subject and object relatives.

I The *parallel function hypothesis* (Sheldon 1974)

II The *perspective-shifting hypothesis* (MacWhinney 1977)

III The *active filler strategy* (Frazier & Flores d'Arcais 1989)

IV The *accessibility hypothesis* (Keenan & Comrie 1977)

V The *storage/integration cost hypothesis* (Gibson 1998)

VI The *confusability/similarity-based hypothesis* (Gordon et al. 2001, 2004)

The first hypothesis in our list, the parallel function hypothesis (PFH), predicts that RCCs in which the head plays the same grammatical role in both clausal constituents are easier than RCCs where the head plays different roles. Since we have discussed the PFH in some detail in § 3.5.2 and because it is closely tied to the account given as II, namely the perspective shifting account (PSA), we may recapitulate the general idea underlying both accounts on the basis of the latter one. The PSA holds that the subject of the MC determines the perspective from which the described scene is viewed. According to the PSA, constructions are harder to

process when they require that the comprehender switches the perspectives from which the (complex) event is described. Changes in perspective are assumed to impose greater demands on working memory. Applied to the examples in (120) and (121) this means that (121) is harder because the reporter is the agent of the MC-event but a patient in the RC-event, while in (120) (s)he plays the same role in both events.

Clifton and Frazier's active filler strategy (AFS) links the resource consumption idea to syntactic complexity. The principle makes reference to the distance from the filler—here the head of the RC—to its gap, i.e. its canonical position in the RC. Memory load is then a function of that distance, so that the greater the distance, the more resources are consumed in the processing of the structure. This account has been refined in a number of ways over time. One important improvement, which has been pointed out by John Hawkins, has is that for such a principle to make reasonable predictions for languages that are typologically different from English, it is more felicitous to consider the distance between the filler and its subcategorizing element, i.e. the verb of the RC (cf. Hawkins 2004: 169ff.). But again, the AFS should be viewed as a representative of many accounts that essentially express the same idea (Wanner and Maratson 1978 may be credited with intellectual ancestry). Keenan and Comrie's accessibility hypothesis can be viewed as a close cousin of the AFH, which essentially formulates a similar idea at a level of grammatical functions (cf. § 1.1.2).

The accounts in V and VI focus on integration or storage cost of integrating incomplete head-dependencies (Ford 1983, Gibson 1998). Perhaps the best-known and best-developed example from this group is Ted Gibson's *Syntactic Prediction Locality Theory* (SPLT). SPLT has two main components: an integration cost component and a memory cost component. The general idea is that on-line processing requires the integration of perceived material into the current discourse representation (or situation model). Each word has to be integrated into the structure currently being processed and this process requires the consumption of certain resources. In addition to the units themselves, the system also has to keep in memory a set of predictions, usually based on dependencies (e.g. verb requires an object), and holding these predictions in memory also consumes resources. Object relatives are predicted to be harder than subject relatives that employ the exact same lexical array,

because they require that the predictions associated with the head have to be kept in memory for a longer time. The model is constantly improved and continues to integrate more constraints from more informational sources, most notableby the impacts from of referential processing (-> givenness value) associated with the respective NP (cf. Gibson 2000, Hsiao and Gibson 2003, Grodner and Gibson 2005, Warren and Gibson 2002 for recent characterizations). On the basis of similar considerations and assumptions, Gordon and colleagues have argued that unintegrated NPs can interfere with each other, especially when they are formally/semantically similar (Gordon et al. 2001, 2004). Object relatives like the one in (121) present two consecutive NPs (*the reporter*, *the senator*) that cannot be integrated into the discourse model, because a crucial component, namely the verbal expression, has not yet been perceived.

We need not discuss each and every single variant of such accounts here—the sheer number of proposals prevents this from being a reasonable goal of this study (for good overviews from different theoretical perspectives on sentence processing cf. Crocker et al. 2000; Pecher and Zwaan 2005, Fedorenko et al. 2006). What is important here is the fact that all these treatments assume that processing linguistic structures consumes resources and that it is the intrinsic properties of these structure that determine the magnitude of such resource consumption. So, while these accounts differ in their details of what property consumes how much of the limited resource, they share a perspective that directly associates intrinsic properties and complexity with processing difficulty. As we have seen in our discussions, these assumptions correctly predict some of the corpus data, e.g. the preference for center embedding in written discourse, but have problems explaining other types of data, e.g. the preference for object relatives in spoken language.

The present perspective is closely tied to tradition that is nearly as time-tested as the resource-limitation view and at least goes back to Tom Bever's work in the early seventies (e.g. Bever 1970). Bever proposed that young children interpret complex patterns on the basis of a sentential template, which he termed the *NVN-schema*. The NVN-schema receives its name from a dominant mapping in English from the syntactic sequence **N**OUN **V**ERB **N**OUN to the semantic structure AGENT ACTION PATIENT. Even though the original formulation of the

NVN-schema hypothesis described the regularity in terms of mapping rules, its very name already suggests that it is easily translated into a construction grammar treatment (cf. § 2.1). Notice that this pattern is essentially the English solution to the task of expressing what Langacker calls the canonical event model (CEM). Instances of the NVN-schema are often also instances of the CEM and are thus likely to be interpreted as bounded, forceful events in which an agent acts on a patient to induce a state (cf. § 3.3.2). The NVN-schema hypothesis predicts that subject relatives are easier because they can be interpreted in a way analogous to simple transitive sentences are interpreted. Consider the examples in (122) to (124).

(122) *John kicked Peter* (simple transitive declarative)

SYN:	$\mathbf{N}_1$	**V**	$\mathbf{N}_2$
SEM:	AGT	ACT	PAT

(123) *John* [*who kicked Peter*] (transitive subject RC)

SYN:	$\mathbf{N}_1$	**V**	$\mathbf{N}_2$
SEM:	AGT	ACT	PAT

(124) *John* [*who Peter kicked*] (object RC)

SYN:	$\mathbf{N}_1$	$\mathbf{N}_2$	**V**
SEM:	PAT	AGT	ACT

The serialization of arguments is identical in (122) and (123), but it is different in (124). So, while the comprehension of transitive subject relatives is facilitated by the highly frequent simple transitive declarative pattern and its canonical interpretation, object relatives exhibit a different structure (NNV), whose comprehension is not facilitated by a similar pattern of comparable frequency. They are supported merely by certain topicalization constructions, but as English is a SVO language such NNV-patterns are a lot less frequent. Townsend and Bever (2001:180) also discuss how the NVN-schema can account for the difficulty of certain center embedded *-ed* participial RCs (e.g. *The horse* $_{RC}$*[raced past the barn] fell*).

Numerous studies in the acquisition literature have produced results consistent with that hypothesis (cf. Slobin and Bever 1982, Townsend and Bever 2001, Diessel 2004, Diessel and Tomasello 2005). While arguments from language acquisition to adult processing generally have to proceed with caution, we have good reasons to engage in such argumentation here on account of the exemplar based view on language representation and processing assumed (cf. § 2.3). In this view the NVN-schema is simply induced from the properties of the ambient language (apparently in early stages of the learning process), and there is no reason to assume that the so constituted schema disintegrates as the language user experiences more linguistic input. The association of monotransitive syntax and CEM semantics will continue to be reinforced over time. As on-line linguistic processing is viewed in exemplar-based theories as a categorization process relative to learned and stored exemplars (or exemplar clusters), we may transfer the idea behind the NVN-schema hypothesis directly from the context of acquisition to adult processing.

Recent years have seen a revival of experience based views not only in the domain of linguistic theorizing (cf. § 2.2), but also in treatments in cognitive psychology. (Tabor et al. 1997, Tabor and Tanenhaus 2001, MacDonald and Christiansen 2002, Reali and Christiansen 2007a , 2007b, Gennari and MacDonald 2008, Wells et al. 2008). We have mentioned earlier that for more than twenty years research into sentence comprehension has pursued two distinct paths of inquiry with virtually no overlap nor attempt of unification: on the one hand psycholinguists have asked how speakers cope with locally ambiguous structures and on the other they have tried to explain how language users process complex but supposedly unambiguous structures. The ambiguity strand was dominated by questions pertaining to the architecture of the processing system—whether it best described in terms of a serial processor that draws on different sources at different stages of processing or whether it is better described as a parallel device that uses all informational sources in parallel. In contrast, the complexity research has focused on memory demand (as we have just seen). The first account to challenge the alleged independency of these two strands of research can be found in Traxler et al. (2002). Traxler and colleagues hypothesized that the complexity effects observed in the context of object relative clauses could be due to a local indeterminacy. Strictly speaking cases like (125) and (126) are locally syntactically ambiguous at *that*:

(125) This is the man that John hates __ .

(126) This is the man that __ hates John.

At the time the relativizer is perceived, comprehenders can tell already that they are going to be presented a relative clause. However, they cannot know what type of RC they are about to hear, i.e. they cannot know yet what grammatical role the nominal *the man* will play within that RC. It may be this indeterminacy that causes the observed comprehension difficulties. The idea to understand the subject/object-RC asymmetry as an ambiguity problem has been picked up in Gennari and MacDonald (2008), who interpret it from the perspective of constraint satisfaction mechanisms. In constraint-based accounts of sentence processing (e.g. MacDonald 1994, 1999), processing difficulties arise when there is competition between alternative structures/interpretations. So just as non-finite *-ed* participial RCs are difficult because at the point the *–ed* form is perceived, because the structure is locally syntactically ambiguous between a main verb and a reduced relative clause reading, finite RCs exhibiting a *that*-relativizer are difficult because they are ambiguous between a subject and some non-subject extraction. This view is fully compatible with the here presented view that language processing just is categorization difficulty. In expectation driven accounts—broadly construed—processing difficulty can be modeled as a function of the frequency (and plausibility) of partially activated patterns.

The description of the parsing mechanism here is general enough to be compatible with a large number of accounts, even with some of those that postulate a multi-stage process. What is required however are a certain sensibility to frequency information and a commitment to an anticipatory character of language processing. In the picture sketched here, a relative clause construction is predicted to be easy to process when it is a) an elaboration of a highly entrenched schema and b) dissimilar from other—functionally distinct—exemplar clusters. Their high levels of entrenchment allow the corresponding structures to become active very quickly and their dissimilarity from potentially competing patterns should minimize competition effects.

An important part of this thesis was the proposal of a quantitative corpus-linguistic methodology that allows us to detect schematic constructions in a statistically sound way. This search for complex associative relationship is is viewed as a natural extension of the huge amount of work done in the identification of collocations (Manning and Schütze 1999, Evert 2004 for overviews) and other bi-grams such as collostructions (Stefanowitsch and Gries 2004) or colligations (cf. Firth 1957, Hoey 1998). The logic behind these notions is exactly the same and so are the limitations of the measures that have been proposed to express the strength of association between the members of the pair (cf. Wiechmann 2008b). From a constructionist perspective it certainly is highly attractive to have at one's disposal a set of statistical procedure that can assess the association strengths of *n*-grams with n > 2. The association rule mining technique (*k*-optimal pattern discovery) and (hierarchical) configural frequency are promising candidates for filling this gap. Schmidtke (under review) has already provided applied the here developed procedure to the area of language acquisition and was able to show some interesting developmental pathways of "*going to* V" versus "*gonna* V" constructions by way of applying a CFA technique to data from the CHILDES databank. Future research may apply these techniques to other domains.

Speaking of methodologies, I would like to address a final, more general issue that I consider very important, namely the relationship of methodology to theory—which has been implicitly raised in many prior passages. As I see it this relationship is largely misconceived in large parts of the linguistic community. While a lot of ink has been spilt to argue for and against the exclusive use of introspection (cf., e.g. Gibbs 2006)—which certainly is an important step in the right direction—there is a lot less sensibility for the impact that measurement operations and type of statistical models may have on theory building. In my view, it is important to acknowledge that not only does our understanding of the empirical domain affects our methodological choices that we make to further understand that domain, but our scientific practices also heavily influence our understanding of the empirical domain. To the extent that the latter is true, we are well advised to assume a strong empiricist view and embrace operationalism. It is dangerous to try and test a theoretical claim that involves a particular theoretical construct (say prototypicality, association strength, entrenchment, …), then choose (more or less arbitrarily) a particular measure to express this construct, set up an

empirical design and then reason directly from the outcome of the experiment to the theoretical plane (►*construct X plays/does not play a role for phenomenon Y*). As different measurement operations may very well yield very different results, we need to think about why procedure X should be used to express the target construct (rather than procedure Y). The statistical procedures employed here were carefully selected among a large (and ever-growing) set of possible tools. The pattern detection procedures—particularly the (h)CFA—were employed in consideration of pressing theoretical concerns, namely to faithfully represent the nature of the knowledge as viewed in construction grammars. The statistical concept of a $type_{CFA}$ is the best expression of the theoretical construct of a $construction_{CxG}$ that I am aware of.

The relationship between theory and methodology certainly is a very close one and their dissociation is a serious mistake. From a strong empiricist view on scientific practice, there is in fact no difference between a theoretical construct and its procedure of measurement. To say it with the words of Percy Bridgman:

> "*In general, we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations*." (Bridgman 1927:5)

The development of the right methodological tools is thus viewed to be not only important to be able to address certain problems but in fact necessary in the development and maturation of a scientific discipline and the present study has tried to contribute to the further development of linguistics as an empirical science.

5 References

Software

R 2.5.0. R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. See also http://www.r-project.org/.

Hcfa 3.2 – A script for R for Windows. Stefan Gries 2007.

Available upon request from the author's website

http://www.linguistics.ucsb.edu/faculty/gries.

Magnum Opus 3.0 (Demo version; Windows)

http://www.giwebb.com/.

Works cited

Aarts, F.

1994 Relative Who and Whom: Prescriptive Rules and Linguistic Reality. *American Speech*, Vol. 69(1):71-79.

Abbot-Smith, K., and Tomasello, M.

2006 Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, 23, 275-290.

Abdi, H.

2007 Binomial distribution: binomial and sign tests. In Salkind, N.J. (ed.), *Encyclopedia of Measurement and Statistics*. Thousand Oaks (CA): Sage. 87-89.

Abney, S.P.

1987 *The English Noun Phrase in its Sentential Aspects*. PhD Dissertation, MIT.

Agrawal, R., Imielinski,T., and Swami, A.

1993 Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD conference on Management of Data*, Washington, D.C.

Akmajian, A.

1970 On deriving cleft sentences from pseudo-cleft sentences. *Linguistic Inquiry,* 1:149-168.

Alexiadou, A., A. Meinunger, C. Wilder, and P. Law, (eds.)

2000 The Syntax of Relative Clauses, *Linguistik Aktuell* 32, John Benjamins, Amsterdam.

Anderson, J.A.

1993 *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

1995 *An Introduction to Neural Networks*. Cambridge, MA/Bradford Books.

Anderson, J. R.*,* and *Bower,* G. H.

Arfarli, T.A.

1994 A promotion analysis of restrictive realtive clauses. *The Linguistic Review*, 11. 81-100.

Ariel, M.

1990 *Accessing noun phrase antecedents*. London: Routledge.

2001 Accessibility Theory: An Overview. In Sanders T., Schilperoord, J., and Spooren, W. (eds.) *Text representation: Linguistic and psycholinguistic aspects*, Amsterdam: Benjamins

Altmann, G.T.M.

1998. Ambiguity in Sentence Processing. *Trends in Cognitive Sciences*, 4, 146-152.

Arnold, D.

2004 Non-restrictive relative clauses in construction based HPSG. In Müller, S. (ed.) *Proceeding of the HPSG04 conference*, Stanford: CSLI publications.

Arnold, D., and Lindarski, E.

A Data-Oriented Parsing Model for HPSG. In Søgaard, A. and Haugereid, P.,(eds.), *2nd International Workshop on Typed Feature Structure Grammars (TFSG'07)*. Tartu, Estonia.

Arnold, J., Wasow, T., Asudeh, A., and Alrenga, P.

2004 Avoiding Attachment Ambiguities: the role of Constituent Ordering. *Journal of Memory and Language,* 55.1, 55-70.

Aylett, M., and Turk, A.

2004 The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47 (1), 31-56.

Bach, E. and Cooper, R.

1978 The NP-S analysis of relative clauses and compositional semantics. *Linguistics and Philosophy,* 2. 145-50.

Baddeley, A.D.

1997 *Human memory: Theory and Practice* (Revised Edition). Hove: Psychology Press.

2000 The episodic buffer: a new component of working memory? *Trends in Cognitive Science*, 4, 417-423.

2007 *Working memory, thought and action*. Oxford: Oxford University Press.

Baddeley, A.D. and Hitch, G.J.

1974 Working memory. In Gower, G.A. (ed.) *The psychology of learning and motivation: Advances in research and theory*, Vol. 8, New York: Academic Press. 47-89.

Baird, R. and Koslick, J.D.

1974. Recall of grammatical relations within clause-containing sentences. *Journal of Psycholinguistic Research*, 3: 165-171.

Barlow, H.

2001 The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*, 24, 602-607.

Barsalou, L. W.

1987 The instability of graded structure in concepts. In U. Neisser (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*. New York: Cambridge University Press. 101-140.

1992 Frames, concepts, and conceptual fields. In Kittay, E. and Lehrer, A. (eds.), *Frames, fields, and contrasts: New essays in semantic and lexical organization*. Hillsdale, NJ: Lawrence Erlbaum Associates. 21-74.

Batali, J.

2002 The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In Briscoe, T. (ed.), editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: CUP.

Bates, E., and Devescovi, A.

1989 A cross-linguistic approach to sentence production. In MacWhinney, and Bates, E. (eds.), *The crosslinguistic study of sentence processing*. New York: Cambridge University Press, 225-256.

Bates, E., and Elman, J.

1996 Learning rediscovered. *Science*, 274: 1849-1850.

Bates, E., and MacWhinney

1989 Competition, variation, and language learning. In B.MacWhinney (Ed.), Mechanisms of language acquisition. Hillsdale, NJ: Lawrence Erlbaum.

Bender, E. and Flickinger, D

1999. Peripheral constructions and core phenomena: Agreement in tag questions. In Webelhuth, G., Koenig, J.P., and Kathol, A. (eds.), *Lexical and Constructional Aspects of Linguistic Explanation*. Stanford: CSLI. 199-214.

Bergen, B. and Chang, N.

2005 Embodied Construction Grammar in Simulation-Based Language Understanding. In Ostman, J.O. and Fried, M. (eds.), *Construction Grammars: Cognitive grounding and theoretical extensions*. 147–190

Bever, T.

1970 The cognitive basis for linguistic structures. In J. Hayes (Ed.),. Cognition and the development of language. New York: Wiley.

1974 The ascent of the spurious, or there's a lot we don't know about mirros. In Cohen, D. (ed.), *Explaining linguistic phenomena*: Washington: Hemisphere. 173-200.

Bianchi, V.

1999 *Consequences of antisymmetry: Headed Relative Clauses*. Berlin: Mouton De Gruyter.

2002a Headed relative Clauses in Generative Syntax. Part I.*Glot International* 6.7, 197-204.

2002b Headed relative clauses in Generative syntax. Part II: *Glot International* 6.8, 235-247.

Blaubergs, M.S. and Braine, M.D.S.

1974 Short-term memory limitations on decoding self-embedded sentences. *Journal of Experimental Psychology*, 102, No.4, 745-748.

Blumenthal, A.

1966. Observations with self-embedded sentences. *Psychonomic Science,* 6, 453-4.

Blumenthal, A., and Boakes, R.

1967 Prompted recall of sentences. *Journal of Verbal Learning and Verbal Behavior*, 6, 674-6.

Bock, J. K.

1986 Syntactic persistence in language production. *Cognitive Psychology*, 18, 355-387.

Bock, J. K., and Loebell, H.

1990 Framing sentences.*Cognition*, 35, 1-39.

Bock, J.K and Warren, R.K.

1985 Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21, 47-67

Bod, R

1998 *Beyond Grammar: An Experience-Based Theory of Language*. CSLI Publications /CUP.

2001 Using Natural Language Processing Techniques for Musical Parsing. *Proceedings ACH/ALLC'2001*, New York, NY.

2002 Memory-Based Models of Melodic Analysis: Challenging the Gestalt Principles. Journal of New Music Research, 31(1): 27-37.

2006 Towards a General Model of Applying Science. *International Studies in the Philosophy of Science* 20(1), 5-25.

2009 From Exemplar to Grammar: A Probabilistic Analogy-based Model of Language Learning. *Cognitive Science*, Vol. 33(4).

Bod, R and Kaplan, R.

1998 A Probabilistic Corpus-Driven Model for Lexical Functional Analysis. *Proceedings COLING-ACL'98*.

Bod, R., Hay, J., and Jennedy, S. (eds.)

2003 *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Bod,R.,Scha, R., and Sima'an, K. (eds.)

2003 *Data-Oriented Parsing*. CSLI Publications, University of Chicago Press.

Boersma, P.

2005 Phonology without markedness constraints. ICLaVE 3. http://www.fon.hum.uva.nl/paul/presentations/ICLaVE3.pdf.

Bolinger, D.

1968 *Aspects of language*. New York: Harcourt, Brace, and World.

1972 *That's that*. Mouton: The Hague.

Boole, G.

1854 *An Investigation of the Laws of Thought, on which are Founded the Mathematical*

Theories of Logic and Probabilities, Walton and Maberly: London; Repr.

Dover: New York.

Brandt, S., Diessel,H., and Tomsello, M.

2008 The Acquisition of German Relative Clauses: A Case Study. *Journal of Child Language,* 35. 325-348.

Brants, T., and Crocker, M.

2000 Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th conference on Computational linguistics*, 1: 111-117.

Bridgman, P.W.

1927 *The Logic of Modern Physics*. New York: MacMillan.

Brinton*, L.J. and* Traugott, E.

2006 *Lexicalization and language change*. Cambridge: CUP.

Brown, R.

1973 *A first language: the early stages*. Cambridge, MA: HUP.

Büring, D.

1995 On the Base Position of Embedded Clauses in German. *Linguistische Berichte,* 159. 370-380.

Burnard, L.

2007 *Reference Guide for the British National Corpus (XML Editio*n). Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services. http://www.natcorp.ox.ac.uk/XMLedition/URG/

Bybee, J.

1998 A functionalist approach to grammar and its evolution. *Evolution of Communication,* 2. 249-278.

1999 Use impacts morphological representation. *Behavioral and Brain Sciences,* 22. 1016-1017.

2001 *Phonology and language use*. Cambridge: CUP.

2006 From usage to grammar: the mind's response to repetition. *Language,* 82(4). 711-733.

2007 *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J. and Scheibman, J.

1999. The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics*, 37-4. 575-596.

Casti, J.L.

1996 *Five golden rules: great theories of 20th-century mathematics - and why they matter*. New York: Wiley-Interscience.

Chafe, W.L.

1994 *Discourse, consciousness and displacement of conscious experience in speaking and writing*. Chicago: The University of Chcago Press.

Chandler, S.

1993 Are rules and modules really necessary for explaining language? *Journal of Psycholinguistic Research*, 22:593-606.

Chomsky, N.

1965 *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

1970 Remarks on nominalization. In: Jacobs, R.A., Rosenbaum, P.S. (eds.). *Readings in English transformational grammar*. Waltenham/Mass.: Ginn and Co.

1980 *Rules and Representations*. New York, Columbia University Press.

1981 *Lectures on Government and Binding*. Dordrecht: Foris.

1986 *Knowledge of Language. It's Nature, Origin, and Use*. New York: Praeger.

1995 The Minimalist Program. Cambridge, MA: MIT Press.

Chomsky, N, and Miller, G.A.

1963 Introduction to the formal analysis of natural languages. In R. Luce, D., Bush, R.R., and Galanter, E. (eds.), *Handbook of mathematical psychology*, vol. 2, New York: Wiley. 269-321.

Christiansen,M., and Chater, N.

1999 Connectionist natural language processing: The state of the art. *Cognitive Science*, 23: 417-437.

Christiansen, M. and Kirby, S. (eds.)

2003 *Language Evolution*, New York: Oxford University Press.

Collins, A.M. and Quillian, M.R.

1969 Retrieval time from semantic memory. Journal of verbal learning and verbal behavior, 8 (2): 240-248.

Crain, S.

1991 Language acquisition in the absence of experience. In *The Behavioral and Brain Sciences*, 4, 597-650.

Crawley, M.

2007 *The R book*. New York: John Wiley.

Crocker, M., Pickering, M., and Clifton, C.

2000 *Architectures and Mechanisms of Language Processing*. Cambridge: CUP.

Croft, W.

2001 Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford: OUP.

2008 Evolutionary linguistics. in Durham, W.H., Brenneis, D. and Ellison, P.T. (eds.). *Annual Review of Anthropology*, 37, 219-34. Palo Alto, CA.

Cruse, D.A., and Croft, W.

2004 *Cognitive Linguistics*. Cambridge: CUP.

Cuetos, F. and Mitchell. D.

1988 Cross-linguistic differences in parsing:Restrictions on the use of the late closure strategy in Spanish. *Cognition*, 3: 73-105

Cuetos, F., Mitchell, D.C., and Corley, M.

1996 Parsing in Different Languages. In M. Carreiras, J. García-Albea and N. Sabastían-Gallés (eds.), Language processing in Spanish, Mahwah, NJ: Erlbaum. 145-187.

Cysouw, M.

2005 Quantitative methods in typology. In: Altmann, G, Köhler, R., and Piotrowski, R. (eds.). *Quantitative Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, 554-578.

Daelemans, W.

1998 Towards an Exemplar-Based Computational Model for Cognitive Grammar. In: Van Der Auwera et al. (eds.), *English as a Human Language*. Munchen: LINCOM, 73-82.

1998b Abstraction is Harmful in Language Learning. In *Proceedings of NeMLaP*, pages 1-2, Sydney, Jannuary.

1999 Memory-Based Language Processing. Introduction to the Special Issue. In: *Journal of Experimental and Theoretical AI* (JETAI), 11:3.

2002 A comparison of analogical modeling of language to memory-based language processing.' In: R. Skousen, D. Lonsdale and D. Parkinson (eds.). *Analogical Modeling*. Amsterdam: John Benjamins, 157-179.

Daelemans, W., Berck, P., and Gillis, S.

1997 Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica*, 31, 57-75.

Daelemans, W., van den Bosch, A.

1992 Generalization Performance of Backpropagation Learning on a Syllabification Task. In Drossaers, M.F.J. and Nijholt, A. (eds.) *Connectionism and Natural Language Processing. Proceedings Third Twente Workshop on Language Technology*, 27-38.

2005 *Memory-Based Language Processing*. Cambridge: CUP.

Dahl, Ö.

2004 The *growth and maintenance of linguistic complexity*. Studies in Language Companion Series. Amsterdam/Philadelphia: John Benjamins.

Davidse, K.

2000 A constructional approach to clefts. *Linguistics*, 38(6), 1101–1131.

Davis, J.T.

1995 Center-embedding and Self-embedding in Human Language Processing. Unpublished Master Thesis. MIT, Dept. of Brain and Cognitive Sciences.

Delespaul, P. A. E. G.

1995. *Assessing schizophrenia in daily life. The experience sampling method*. Maastricht: Universitaire Pers Maastricht.

Dennett, D.

1981 True Believers: the Intentional Strategy and Why it Works. In Heath, A. F. (ed.), *Scientific Explanation*. New York: OUP.

Deppermann, A. and Elstermann, M.

2008 Lexikalische Bedeutung oder Konstruktionsbedeutungen? Eine Untersuchung am Beispiel von Konstruktionen mit verstehen. In: Stefanowitsch, A., and Fischer, K. (eds.): *Konstruktionsgrammatik II: Von der Konstruktion zur Anwendung*. Tübingen: Stauffenburg, S. 103-133.

Diessel, H.

2004 *The Acquisition of Complex Sentences*. [Cambridge Studies in Linguistics 105].

Cambridge: CUP.

2007 Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25: 108-127.

2009 On the role of frequency and similarity in the acquisition of subject and non-subject relative clauses. In Talmy Givón and Masayoshi Shibatani (eds.), Syntactic Complexity, 251-276. Amsterdam: John Benjamins.

Diessel, H. and Tomasello, M.

2005 A new look at the acquisition of relative clauses. *Language,* 81: 1-25.

Diewald, G.

2008 Die Funktion "idiomatischer" Konstruktionen bei Grammatikalisierungs prozessen – illustriert am Beispiel der Modalpartikel *ruhig*. In Fischer, K. and Stefanowitsch, A. *Konstruktionsgrammatik II – Von der Konstruktion zur Grammatik*. Tübingen: Stauffenburg. 33-59.

Dixon, R.M.W.

1997 *The rise and fall of languages*. Cambridge: CUP.

Doherty, C.

1993 *Clauses without that: the case for bare sentential complementation in English*. PhD dissertation, Santa Cruz: University of California.

DuBois, J.W.

1980 Beyond definiteness: the trace of identity in discourse. In Chafe, w.L. (ed.) *The pera stories: cognitive, cultural, and linguistic aspects of narrative production*.Norwood, NJ: Ablex: 203-274.

Dryer, M.

2006 Functionalism and the Theory - Metalanguage Confusion. In Wiebe,G., Libben,G., Priestly,T.,Smyth,R., and Wang,S *(eds.) Phonology, Morphology, and the Empirical Imperative: Papers in Honour of Bruce Derwing*, pp. 27-59.Taipei: The Crane Publishing Company

Elder, J.H. and Goldberg, R.M.

1998 The statistics of natural image contours. In *IEEE Workshop on Perceptual Organisation in Computer Vision*.

Ellis, N.

1995 The Study of Second Language Acquisition. Oxford: OUP.

Ellis, S.R. and Hitchcock, R.J.

1986 The emergence of Zipf's law: Spontaneous encoding optimization by users of a command language. *EEE Transactions on Systems, Man and Cybernetics* 16(3): 423-427.EEE Press Piscataway, NJ, USA.

Elman, J.

1990 Finding structure in time. *Cognitive Science*, 14, 179-211.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., and Plunkett, K.

1996 *Rethinking Innateness: A Connectionist Perspective on Development.* Cambridge, MA : MIT Press.

Elsness, J.

1984 That or Zero? A Look at the Choice of Object Clause Connective in a Corpus of American English. *English Studies*, 65.

Erteschik-Shir, N.

2007 *The syntax/discourse interface: Information Structure*. OUP, Oxford.

Everitt, B.S.

1993 Cluster Analysis. New York: John Wiley and Sons.

Evert, S.

2004 The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Eye, A. von

1990 *Introduction to configural frequency analysis. The search for types and anti-types in cross-classification*. Cambride: CUP.

Fauconnier, G.

1994 *Mental Spaces*. New York: CUP.

1997 Mappings in Thought and Language. Cambridge: CUP.

Ferreira, F., Clifton, C.

1986 The independence of syntactic prosessing. *Journal of Memory and Language*, 27, 429-446.

Ferreira, V. S. and G.S. Dell

2000 Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production. *Cognitive Psychology,* 40, 296-340.

Feldman, J.A.

2006 *From Molecule to Metaphor: A Neural Theory of Language*, 1 ed., Bradford Books, Cambridge, MA: MIT Press

Feldman, J. A., and Ballard, D. H.

1982 Connectionist Models and their Properties. *Cognitive Science,* 6: 205-254.

Fedorenko, E., Gibson, E. and Rohde, D.

2006 The Nature of Working Memory Capacity in Sentence Comprehension: Evidence Against Domain-Specific Working Memory Resources. *Journal of Memory and Language,* 54*,* 541-553.

Fidelholtz, J. L.

1975 Word frequency and vowel reduction in English. *CLS*, 11, 200-213.

Fillmore, C.J., P. Kay, and M.C. O'Connor

1988 Regularity and idiomaticity in grammatical constructions: the case of let alone, *Language*, 64, 501-538.

Firth, J.R.

1957 A synopsis of linguistic theory 1930–55, Studies in linguistic analysis, The Philological Society. 1–32.

Fitz, H., and Chang, F.

2008 The role of the input in a connectionist account of the accessibility hierarchy in development. *Proceedings of the 32nd Boston University Conference on Language Development*. 120-131.

Flynn, S and Foley, C.

2004 The Cumulative-Enhancement Model for Language Acquisition: Comparing Adults' and Children's Patterns of Development in First, Second and Third Language Acquisition of Relative Clauses. *International Journal of Multilingualism*, 1:1, 3-17.

Fodor, J.A.

1983 *Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass.: MIT Press.

Fodor, J. A., Bever, T., and Garrett, M.

1974 *The Psychology of Language*.New York: McGraw-Hill.

Fodor, J. and Garrett, T.

1967 Some syntactic determinants of sentential complexity. *Perception and Psychophysics*,

2, 289-96.

Fodor, J.A., and Pylyshyn

1988 Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28: 3-71.

Ford, M.

1983 A method of obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, 22, 203-218.

Fox, B.A.

1987 The Noun Phrase Accessibility Hierarchy reinterpreted: Subject primacy or the Absolutive Hypothesis? *Language* 63.856–870.

Fox, B. A. and Thompson, S. A.

1990 A Discourse Explanation of the Grammar of Relative Clauses in English Conversation. *Language,* 66: 297-316.

Frauenfelder, U.H., Segui, J. and Mehler, J.

1980 Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior*, 19, 2.

Frazier, L.

1987 Sentence processing: A tutorial review, In Coltheart, M., *Attention and Performance XII: The Psychology of Reading*, Lawrence Erlbaum Associates.

Frazier, L., and Flores d'Arcais, G.B.

1989 Filler-driven parsing: a study of gap filling in Dutch. *Journal of Memory and Language*, 28, 331-344.

Frazier, L. and Fodor, J.D.

1978 The sausage machine: A new two-stage parsing model. *Cognition,* 6, 291-325.

Frege, G.

1892 Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, NF 100, 25-50.

Friendly, M.

1994 Mosaic-displays for multi-way contingency tables. Journal of American Statistical Association, 89: 190-200.

1999 Extending mosaic displays: Marginal, conditional, and partial view of categorical data. Journal of Computational and Statistical Graphics, 8 373-395.

Gazdar, G., Klein, E., Pullum, G. K., and Sag, I.

1985 Generalized Phrase Structure Grammar, Basil Blackwell, Oxford.

Geng, J. J., and Behrmann, M.

2005 *Spatial probability as an attentional cue in visual search*. *Perception and Psychophysics*, 67, 1252-1268

Gennari, S. P., and MacDonald, M. C.

2008 Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, 58, 161-187.

Gentner, D., Holyoak, K.J., and Kojkinov, B.N.

2001 *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.

Gernsbacher, M.A.

1990 *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gibbs, R. W.

2006 Just why should cognitive linguists care about empirical evidence, much less want to go to the trouble of gathering it? In Gonzalez-Marquez, M.,Mittelberg, I., Coulson, S., and Spivey, M.J. (eds), Empirical Methods in Cognitive Linguistics. Amsterdam: John Benjamins.

Gibson, E.

1991 *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. PhD thesis, Carnegie Mellon University.

1998 Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1-76.

2000 The Dependency Locality Theory: A distance-based theory of linguistic complexity. In Marantz, A., Miyashita, Y., and O'Neil, W. (eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, 95-126, Cambridge: MIT Press

Gibson, E., Desmet, T., Grodner, D., Eatson, D. and Ko, K.

2005. Reading relative clauses in English. *Cognitive Linguistics,* 16: 313-353.

Gibson, E., Pearlmutter, N., Canseco Gonzalez, E., and Hickok, G.

1996 Recency preference in the human sentence processing mechanism. *Cognition*, 59, 23-59.

Gibson, E., and Thomas, J.

1999 Memory Limits and Structural Forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14, 225-248.

Givon, T.

1983 *Topic continuity in discourse: a quantitatve cross-linguistc study*. Amsterdam: John Benjamins.

Godfrey, J., Holliman, E., and McDaniel, J.

1992 SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of ICASSP-92*, 517-520.

Goldberg, A.

1995 *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

2003 Constructions: A New Theoretical Approach to Language. *Trends in*

Cognitive Science 7. 5:219-224.

2006 *Constructions at work: The Nature of Generalisation in Language*. Oxford: OUP.

Goldinger, S.D.

1996 Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology:Learning, Memory, and Cognition*, 22, 1166-1183.

Goldinger, S.D., Luce, P.A., Pisoni, D.B., and Marcario, J.K.

1992 Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1210-1238.

Goldstein, B.

2001 Sensation and Perception, 6th ed. London: Wadsworth.

Gordon, P.C., Hendrick, R., and Johnson, M.

2001 Memory interference during language processing. *Journal of experimental psychology: learning, memory and cognition*, 27, 1411-1423.

2004 Effects of noun phrase type on sentence complexity. *Journal of memory and language*, 51, 97-114.

Gower, J. C.

1985 Measures of similarity, dissimilarity, and distance. In: Kotz, S. and Johnson, N. L. (eds,). *Encyclopedia of Statistical Sciences*, Vol. 5. Wiley, New York. 397-405.

Granger, S.

1983 *The BE+past participle construction in spoken English with special emphasis on the passive*. Amsterdam: Elsevier Science.

Grice, H. P.

1975 Logic and conversation. In Cole, P. and Morgan, J.L., (eds.). *Syntax and semantics: Speech acts*. Volume 3. New York: Academic. 41-58.

Gries, S.T.

2008 *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoek and Ruprecht.

Grodner, D. and Gibson, E.

2005 Consequences of the serial nature of linguistic input. *Cognitive Science*, 29, 261-291.

Grodner, D., Gibson, E., and Tunstall, S.

2002. Syntactic Complexity in Ambiguity Resolution. *Journal of Memory and Language*, 46, 267-295.

Gundel, J.

1977 Where do cleft sentences come from? *Language,* 53:53–59.

Gundel, J., Hedberg, H., and Zacarski, R.

1993 Referring expressions in discourse. *Language*, 69, 274-307.

Haiman, John.

1983 Iconic and economic motivation. *Language*, 59: 781-819.

1985 *Natural Syntax: Iconicity and Erosion*. Cambridge: CUP.

Hale, J.

2001 A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, volume 2, 159-166.

Halliday, M.A.K and Matthiessen, M.I.M.

2004 Subject, actor, theme in An introduction to functional grammar. Hodder Arnold, London, England.

Harris, C.

1994 Backpropagation representations for the rule-analogy continuum. In J. Barnden, and K. Holyoak, (Eds.), Analogical Connections. Norwood, N.J: Ablex.

1998 Psycholinguistic studies of entrenchment. In J. Koenig (Ed.), Conceptual Structures,

Language and Discourse, Vol 2. Berkeley, CA: CSLI.

Hawkins, J. A.

1994 *A Performance Theory of Order and Constituency*. Cambridge, UK. Cambridge University Press.

2004 *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford

Hay, J.

2001 Lexical Frequency in Morphology: Is Everything Relative? *Linguistics* , 39, 1041-1070.

Hay, J. and Bresnan, J.

2006 Spoken Syntax: The Phonetics of giving a hand in New Zealand English. *The Linguistic Review,* 23.

Hebb, D.

1949 The organization of behaviour: A neuropsychological theory. New York: Wiley.

Helmholtz, H.von.

1925 Physiological Optics. Volume III. *The Theory of the Perceptions of vision* (Translated from 3rd German Edition, 1910). Washington: Optical Society of America.

Hintikka, J.

1962 Knowledge and Belief. Ithaca, NY: Cornell University Press.

Hoekstra, E.

1992 On the parametrization of functional projections in CP.*North-eastern Linguistic Society*, 12. 191-204.

Hoey, M.

1998 Introducing Applied Linguistics: 25 Years on. *The 31st BAAL Annual Meeting: Language and Literacies*. University of Manchester.

Hoffmann, T.

2005 Variable vs. categorical Effects: Preposition pied piping and stranding in British English relative clauses. *Journal of English Linguistics,* 33(3): 257-297.

Holmes, V.M.

1973 Order of main and subordinate clauses in sentence perception. *Journal of Verbal Learning and Verbal Behavior*, 12, 285-293.

Holmes, V., and O'Regan, J.K.

1981 Effects of syntactic structure on eye fixations during reading. *Journal of Verbal Learning and Verbal Behavior*, 20, 417-430.

Hooper, Joan B.

1976 Word frequency in lexical diffusion and the source of morphophonological change. In William Christie (ed.) *Current progress in historical linguistics*. Amsterdam: North Holland. 95-105.

Hopper, P.J.

2001 Grammatical Constructions and their Discourse Origins: Prototype or Family Resemblance? In Pütz, M. and Niemeier, S. (eds.), *Applied Cognitive Linguistics: Theory, Acquisition, and Language Pedagogy*. Berlin: Mouton/De Gruyter. 109-130.

Hopper, P.J. and Traugott, E.

2003 *Grammaticalization*. Cambridge: CUP.

Hsiao, F. and Gibson, E.

2003 Processing relative clauses in Chinese. *Cognition*, 90, 3-27.

Huddleston, R.

1988 *Introduction to the Grammar of English*, Cambridge: CUP.

Huddleston, R, and Pullum, G.K.

2002 *The Cambridge Grammar of the English Language*. Cambridge: CUP.

Hudson, R.

1990 *English Word Grammar*. Oxford: Basil Blackwell.

Hurford,J.R.

2007 *Language in the light of evolution*. New York: OUP.

Hymes, D.

1974 *Foundations in Sociolinguistics: An Ethnographic Approach*. Philadelphia: University of Pennsylvania Press.

Jackendoff, R.

1975 Morphological and semantic regularities in the lexicon. *Language,* 51. 639-671.

1977 *X-bar syntax: A study of phrase structure*. Cambridge, MA: MIT Press.

Jaeger, T.F., Levy, R., Wasow, T. and Orr, D.M

2005 The Absence of "that" is Predictable if a Relative Clause is Predictable. Talk

presented at *AMLaP*, Ghent, Belgium.

Jespersen, O.

1917 *Negation in English and other languages*. Copenhagen: Host.

Johnson, K.

2007 Decisions and Mechanisms in Exemplar-based Phonology. In Sole, M.J., Beddor, P. and Ohala, M. (eds) *Experimental Approaches to Phonology. In Honor of John Ohala*. Oxford University Press. 25-40.

Johnson, T. C.

1983 *Phonological free variation, word frequency, and lexical diffusion*. University of Washington PhD thesis.

Jurafsky, D.

1996 A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20. 137-194.

Just, M. A., and Carpenter, P. A.

1992 A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 98, 122-149.

Kay, P., and Fillmore, C.J.

1999 Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. *Language*, 75(1).

Kayne, R.S.

1994 *The antisymmetry in syntax. Cambridge*, MA: MIT Press.

Keenan, E. L. and Comrie, B.

1977 Noun phrase accessibility and universal grammar'. *Linguistic Inquiry,* 8. 63–99.

Kidd, E., Brandt, S., Lieven, E.V., and Tomasello, M.

2007 Object relatives made easy: A crosslinguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes*, 22. 860-897.

King, J., and Just, M.A.

1991 Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 580-602.

King, J., and Kutas, M.

1995 Who did What and When? Using word- and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience,* 7(3), 376-395.

Kirby, S.

1997 Competing motivation and emergence: Explaining implicational hierarchies. *Linguistic Typology,* 1(1), 5-31.

2007 The evolution of language. In Dunbar, R. and Barrett, L., (eds.), *Oxford Handbook of Evolutionary Psychology*. New York: OUP. 669-681.

Konieczny, L., Hemforth, B., Scheepers, C. and Strube, G.

1997 The role of lexical heads in parsing: evidence from German. *Language and Cognitive Processes*, 12, 307-348.

Kortmann, B., and Szmrecsanyi, B.

Global synopsis - morphological and syntactic variation in English. In: Kortmann, B., Burridge, K., Mesthrie, R., and Schneider, E. (eds.), *A Handbook of Varieties of English, Vol. 2: Morphology and Syntax*. Berlin/New York: Mouton de Gruyter, 1122-1182.

Koster, J.

1978 Why subject sentences don't exist. In Keyser, S.J. (ed.), *Recent transformational studies in European languages*. Cambridge, MA: MIT Press.

Krauth, J.

1993 *Einführung in die Konfigurationsfrequenzanalyse*, Beltz-Verlag, Weinheim

Krauth J., and Lienert G. A.

1973/1995 Die Konfigurationsfrequenzanalyse (KFA) und ihre Anwendung in Psychologie und Medizin Beltz Psychologie Verlagsunion.

Kripke, S.

1963 Semantical analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9:67-96

Kucera, H. and Francis, W. H.

1967 *Computational analysis of present-day American English*. Providence, RI: Brown University

Lakoff, G.

1987 *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

Lambrecht, K.

1988 There Was a Farmer Had a Dog: Syntactic Amalgams Revisited.*Proceedings of BLS* 14:319-339.

1994 *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents*. Cambridge: CUP.

2001 A Framework for the Analysis of Cleft Constructions. *Linguistics,* 29, 463-516.

Langacker, R.

1987 *Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

1990 *Concept, Image, Symbol: The cognitive basis of grammar*. Berlin/New York. Mouton/Walter De Gruyter.

1990b The Rule Controversy: A Cognitive Grammar Perspective. *CRL Newsletter* 4.3.4-15.

1999 *Grammar and Conceptualization*. Berlin/New York. Mouton/Walter De Gruyter.

2007 *Ten Lectures on Cognitive Grammar* (New Titles in New Format in Cognitive Linguistics from Beijing)

2008 *Cognitive Grammar: A Basic Introduction*. New York: Oxford University Press

Lapata, M., Keller, F. and Schulte im Walde, S.

2001 Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research*, 30(4):419-435.

Larkin, W. and Burns, D.

1977 Sentence comprehension and memory for embedded structure. *Memory and Cognition*, 5, 17-22.

Lautsch, E, and vonWeber, S.

1995 *Methoden und Anwendungen der Konfigurationsfrequenzanalyse*. Beltz-Verlag, Weinheim.

Lehmann, C.

1984 *Der Relativsatz. Typologie seiner Struktur, Theorie seiner Funktionen, Kompendium seiner Grammatik*. Narr, Tübingen.

Levin, B.

1993 English *Verb Classes and Alternations: A Preliminary Investigation*.Chicago and London: The University of Chicago Press.

Levinson, S.

2000 *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.

Levy, R. and Jaeger, T.F.

2007 Speakers optimize information density through syntactic reduction. In B. Schlökopf, Platt, J., and Hoffman, T. (eds.), Advances in neural information processing systems (NIPS) 19, Cambridge, MA: MIT Press. 849-856.

Lewis, D.

1969 *Convention: A Philosophical Study*. Harvard: HUP.

Lewis, R.L.

1995 *A theory of grammatical but unacceptable embeddings*. Technical report. Princeton University.

Limber, J.

1976 Syntax and sentence interpretation. In Wales, R.J. and Walker, E. (eds.), *New Approaches to language mechanisms*. Amsterdam: North Holland.

Lyons, C.

1999 *Definiteness*. Cambridge: CUP.

MacDonald, M. C.

1994 Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, 9, 157-201.

1999 Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney. (Ed.), *The Emergence of Language*. Mahwah, NJ: Erlbaum. 177-196.

MacDonald, M. C., and Christiansen, M. H.

2002 Reassessing working memory: A comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109, 35-54.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S.

1994 The lexical nature of syntactic ambiguity resolution. Psychological Review, 101, 676-703.

MacDonald, M. C., and Seidenberg, M. S.

2006 Constraint satisfaction accounts of lexical and sentence comprehension. In M. J.

Traxler., and M. A. Gernsbacher. (eds.), *Handbook of Psycholinguistics*, 2nd Edition). London: Elsevier Inc. 581-611.

MacWhinney, B.

1977 Basic syntactic processes. In Kuczaj, S. (ed.), *Language development: Volume 1, syntax and semantics*. Erlbaum, Hillsdale, NJ. 73-136.

semantics pp. 73-136. Erlbaum, Hillsdale, NJ

MacWhinney, B., and Pleh, C.

1988 The processing of restrictive relative clauses in Hungarian. *Cognition*, 29, 95-141.

Mak, W.M., Vonk, W., and Schriefers, H.J.

2002 The influence of animacy on relative clause processing. *Journal of Memory and Language*, 47 (1), 50-68.

2006 Animacy in processing relative clauses: The hikers that rocks crush. *Journal of Memory and Language*, 54(4), 466-490.

Manning, C. and Schütze, H.

1999 *Foundations of Statistical Natural Language Processing*. Cambridge, MA:MIT Press.

Marascuilo, L.A.

1970 Extensions of the significance test for one-parameter signal detection hypotheses, *Psychometrika*, 35,237-243.

Marcus, G. F.

2001 *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S.

1995 German inflection: The exception that proves the rule. *Cognitive Psychology,* 29, 186-256.

Marks, L.E.

1968 Scaling of grammaticalness of self-embedded English sentences. *Journal of Verbal Learning and Verbal Behavior*, 7, 965-967.

Marr, D.

1982 *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Martin, S.E.

2004 *A reference grammar of Japanese*. Honolulu: University of Hawai'i Press.

Matsumoto, Y.

1997 Noun-Modifying Constructions in Japanese: A Frame Semantic Approach. *Studies in Language Companion Series* 35. John Benjamins.

2007 Integrating frames: Complex noun phrase constructions in Japanese.In *Aspects of Linguistics: In Honor of Noriko Akatsuka (Gengogakuno Syosoo: Akatsuka Noriko Kyoozyu Kinen Ronbunsyuu)*, (eds.) Kuno, S., Makino, S., and Strauss, S. Tokyo: Kurosio Publishers. 131-154.

McCawley, J.D.

1982 Parentheticals and discontinuous constituent structure. *Linguistic Inquiry*, 13. 91-106.

McClelland, J.L., and Elman, J.L.

1986 The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

McClelland, J. L., and Rumelhart, D. E.

1981 An interactive activation model of context effects in letter perception, Part 1: An account of basic findings. *Psychological Review*, 88, 375-405

1989 *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Excercises*. London: The MIT Press.

McLeod, P., Plunkett, K., and Rolls, E.T.

1998 *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: OUP.

Merin, A.

1997 Information, relevance and social decision making. In Moss, L., Ginzburg, J., and De Rijke, M. (eds.) *Logic, Language and Computation*, 2. Stanford: CSLI publications.

Micheal, G.

2007 A significance test of interaction in 2 × K designs with proportions. *Tutorials in Quantitative Methods for Psychology*, 3(1), 1-7.

Mill, J.S.

1862 *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Oxford: Parker and Son.

Miller, G. A.

1956 The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.

2003 The cognitive revolution: A historical perspective. Trends in Cognitive Sciences, 7(3), 141-144.

Miller, G.A., and Isard, S. 1964. Free recall of elf-embedded English sentences. *Information and Control* 7: 292-303.

Müller, S.

2006 Phrasal or Lexical Constructions? *Language,* 82(4), 850-883.

Narayanan, S. and Jurafsky, D.

2001 A Bayesian Model Predicts Parse Preferences and Reading Times in Sentence Comprehension, Neural Information Processing Systems.

Neath, I. and Surprenant, A. M.

2003 *Human Memory* (2nd ed.). Belmont, CA: Wadsworth.

Nelson, G.

1996 The Design of the Corpus'. In: Greenbaum, S. (ed.) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press, 27-35.

Newmeyer, F.J.

1999 *Language form and language function*. Cambridge, MA: MIT Press.

Nilsson, Nils

1998 *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann Publishers.

Nosofsky, R. M.

1986 Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, Vol. 115, No. 1, pp. 39-57.

Ong, E.J., Micilotta, A., Bowden,R., and Hilton, A. Viewpoint invariant exemplar-based 3D human tracking. *Computer Vision and Image Understanding* 104(2-3): 178-189.

Pecher, D. and Zwaan, R.A.

2005 *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*. Cambridge: CUP.

Perli, H.G.

1985 *Testverfahren in der Konfigurationsfrequenzanalyse*. Erlangen: Palm und Enke.

Phillips, B. S.

1984 Word frequency and the actuation of sound change. *Language*, 45, 9-25.

1999 The mental lexicon: evidence from lexical diffusion. *Brain and Language*, 68, 104-109.

Pickering, M.J. and Branigan, H.P.

1999 Syntactic priming in language production. *Trends in Cognitive Science*, 3, 136-141.

Pierrehumbert, J. B.

2001 Exemplar dynamics: Word frequency, lenition and contrast. In Bybee, J. anf Hopper, P. (eds.), Frequency and the emergence of linguistic structure. John Benjamins. 137-157.

Pinker, S.

1974 Formal models of language learning. Cognition, 7, 217-83.

1989 Learnability and Cognition: The acquisition of Argument Structure. Cambridge, Mass.: MIT Press.

Pinker, S. and Prince, A.

1988 Rules and connectionsin human language. Trends in Neurosciences, 11, 195-202.

Pisoni, D.B.

1996 Word identification in noise.Language and Cognitive Processes, 1996, 11 (6), 681-687.

Plunkett, K. and Marchman, V. A.

1991 U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. Cognition, 38, 43-102.

Pothos, E. M., and *Juola*, P.

2007 Characterizing linguistic structure with mutual information. *British Journal of Psychology*, 98, 291-304.

Prat-Sala, M., and Branigan, H. P.

1999 Discourse constraints on syntactic processing in language production: a cross-linguistic study in English and Spanish. *Journal of Memory and Language*, 1-15.

Prince, E.

1981 Toward a taxonomy of given/new information. In Cole, P. (ed.), *Radical Pragmatics*. Academic Press. New York, p. 223-255.

Pu, M.

2007 The distribution of relative clauses in Chinese discourse. *Discourse Processes*, 43,

1, 25-53.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J.

1985 *A Comprehensive Grammar of the English Language*. Longman.

Race, D. S., and MacDonald, M. C.

2003 The use of "that" in the production and comprehension of object relative clauses. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*.

Rao, R.P.N., Ohlshausen, B.A., and Lewicki, M.S.

2002 *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press.

Reali, F. and Christiansen, M.H.

2007a Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 53, 1-23.

2007b Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology*, 60, 161-170

Redington, M., Chater, N., and Finch, S.

1993 Destributional information and the acquisition of linguistc categories: A statistical approach. In *Proceedings of the 15th annual meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc. 848-853.

Recanati, F.

1993 *Direct Reference: From Language to Thought*. Oxford: Blackwell.

Rizzi, L.

1997 The fine structure of the left periphery, in Haegeman, L.(ed.) *Elements of Grammar*, 281-337. Dordrecht: Kluwer.

Rodriguez, A.E., Vadera,S., Sucar, L.E.

2000 A Probabilistic Exemplar-Based Model for Case-Based Reasoning. *MICAI 2000*: 40-51.

Roeper, P.

2003 Multiple Grammars, Feature Attraction, Pied-Piping, and the Question: Is AGR inside TP? In Müller, N. (ed.) *(In)vulnerable Domains in Multilinguialism*. Amsterdam: John Benjamins, pp. 335-360.

Roland, D., Elman, J., Ferreira, V.S.

2006 Why is that? Structural prediction and ambiguity resolution in a very large corpus of

English sentences. *Cognition,* 98, 245-272.

Roland, D., Dick,F., and Elman, J.

2007 Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language,* 57(3). 348-379.

Romero-Trillo, J.

2006 Discourse markers. In: Encyclopedia of Language and Linguistics. Oxford: Elsevier, pp. 639-642.

Rosch, E.

1973 Natural categories, *Cognitive Psychology,* 4: 328-50.

1978 Principles of categorization. In Rosch, E., and Lloyd, B.B. (eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.

Rosenblatt, F.

1958 The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386-408.

Ross, John R.

1986 *Infinite syntax!*. Norwood, NJ: ABLEX.

Rumelhart, D. E., McClelland, J. L., and The PDP Research Group

1986 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. London: The MIT Press.

Sag, I.

1997 English relative clause constructions. *Journal of Linguistics*, 30. 587-620.

Saitou, N., and Nei M.

1987 The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4:406-425.

Sanford, A.J. and Garrod, S.C.

1981 *Understanding written language*. Chichester: John Wiley and sons.

Schachter, P.

1973 Focus and relativization, *Language*, 49, 19-46.

Schlesinger, I.M.

1968 *Sentence structure and the reading process*. The Hague: Mouton.

Schmidtke, K.

to appear *A Typology of Purpose Clauses*. [Typological Studies in Language] Amsterdam/Philadelphia: John Benjamins.

under review Going-to-V and gonna-V in child language: A quantitative approach to constructional development. *Cognitive Linguistics*.

Schuchardt, H.

1885 *Über die Lautgesetze: gegen die Junggrammatiker*. Berlin: Oppenheim.

Searle, J.

1969 *Speech acts: An essay in the philosophy of language*. Cambridge, England: Cambridge University.

Seidenberg, M.S.

1997 Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275: 1599-1603.

Selfridge, O.G., Sutton, R.S., and Anderson, C.W.

1988 Selected bibliography on connectionism. In Lee, Y.C. (ed) *Evolution, Learning and Cognition*, Singapore: World Scientific. 391–403.

Sheldon, A.

1974 The Role of Parallel Function in the Acquisition of Relative Clauses in English. *Journal of Verbal Learning and Verbal Behavior*, 13, 272-281.

Skousen, R.

1989 *Analogical modeling of language*. Dordrecht: Kluwer Academic Press.

Skousen, R., Lonsdale, D. and Parkinson, D.B. (eds)

2002 *Analogical Modeling. An exemplar-based approach to language*. Amsterdam: John Benjamins.

Slobin, D.

1973 Cognitive prerequisites for the development of grammar. In Ferguson, C.A. and Slobin, D. (eds.), *Studies of Child Language*. New York: Holt, Rinehart and Winston, 175-208.

Slobin, D. and Bever, T.G.

1982. Children Use Canonical Sentence Schemas: A Cross-linguistic Study of Word Order and Inflections. *Cognition*, 12(3), 229-265.

Smith, C.

1964 Determiners and relative clauses in a generative grammar of English. *Language,* 40. 37-52.

Smith, E. E., and Medin, D. L.

1981 *Categories and concepts*. Cambridge, MA: HUP.

Sperber, D. and Wilson, D.

1995 *Relevance*. Oxford: Blackwell.

Stabler, E.P.

1994 *The finite connectvity of linguistic structure*. Technical report, UCLA.

Steels, L. and De Beule, J.

2006 A (very) Brief Introduction to Fluid Construction Grammar. *Third International Workshop on Scalable Natural Language Understanding (ScaNaLU 2006)*, New York City.

Stefanowitsch, A.

2003 A construction-based approach to indirect speech acts. In Panther, K.-U. and Thornburg, L.L. (eds.) *Metonymy and Pragmatic Inferencing*. 105-126.

2006 Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory,* 2(1), 61-77.

Stefanowitsch, A. and Gries, S.T.

2003 Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics 8.2:209-43.

Stockwell, R.P., Schachter, P., and Hall Partee, B.

1973 *The major syntactic structures of English*. New York: Holt: Rinehart and Winston.

Stolz, T.

2006 (Wort-)Iteration: (k)eine universelle Konstruktion. In Fischer, K. and Stefanowitsch, A. *Konstruktionsgrammatik. Von der Anwendung zur Theorie.*Tübingen: Stauffenburg, 105-132.

Stolz, W.S.

1967 A study of the ability to decode grammatically novel sentences. *Journal of Verbal Learning and Verbal Behavior*, 6, 867-873.

Sutton, J.

1998 *Philosophy and memory traces: Descartes to connectionism*. Cambridge: CUP.

Swinney, D.

1979 Lexical access during sentence comprehension: (Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18, 645-659.

Tabor, W., Juliano, C., and Tanenhaus, M.

1997 Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, 12, 211–271.

Tabor, W., and Tanenhaus, M. K.

2001 Dynamical systems for sentence processing. In Christiansen, M, and Chater, N. (eds.), *Connectionist psycholinguistics: Capturing the empirical data*. Westport, CT: Ablex

Tanenhaus, M. K., Leiman, J. M., and Seidenberg, M. S.

1979 Evidence for multiple stages in the processing of ambiguous words in syntactic context. *Journal of Verbal Learning and Verbal Behavior*, 18, 427-441.

Tavakolian, S.

1981 The conjoined clause analysis of relative clauses. In Tavakolian, S. (ed.),

Language Acquisition and Linguistic Theory. Cambridge:MIT Press, 167-187.

Temperley, D.

2003 Ambiguity avoidance in English relative clauses. *Language* 79(3), 464-484.

Tomasello, M.

1992 *First Verbs - A Case Study of Early Grammatical Development*. Cambridge:CUP.

Tottie, G.

1995 The man 0 I love: An analysis of factors favouring zero relatives in written British and American English. In Melchers, G., and Warren,B. (eds.), *Studies in Anglistics*. Stockholm:Almqvist and Wiksell. 201-215.

Townsend, D.J., and Bever, T.G.

2001 *Sentence Comprehension: The Integration of Habits and Rules*. MIT Press.

Traxler, M. J., Morris, R. K., and Seely, R. E.

2002 Processing subject and object relative clauses: evidence from eye movements. *Journal of Memory and Language*,47, 69-90.

Trueswell, J. C. and Tanenhaus, M. K.

1994 Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In

Clifton, C., Frazier, L. and Rayner, K. (eds.) *Perspectives in Sentence Processing*. Lawrence Erlbaum Assoc.: Hillsdale, NJ, pp. 155-179.

Trueswell, J.C., Tanenhaus, M.K., Garnsey, S.M.

1994 Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285-318.

Turchin, P.

2003 *Complex Population Dynamics: a Theoretical/Empirical Synthesis*. Princeton University Press.

Tversky, A.

1977 Features of similarity. *Psychological Review*, 84, 327-352.

Van den Bosch, A.

1999 Instance-family abstraction in memory-based language learning. In: Bratko, I. and Dzeroski, S. (eds.), *Machine Learning: Proceedings of the Sixteenth International Conference, ICML'99*, Bled, Slovenia.

Van Gelder, T. J.

1995 What might cognition be, if not computation? *Journal of Philosophy*, 91, 345-381.

von Eye, A., and Peña, E.

2004 Configural Frequency Analysis: the Search for Extreme Cells. *Journal of Applied Statistics*, 31, 981-997.

von Eye und Rovine, M. J.

1988 A comparison of significance tests for configural frequency analysis. *EDV in Medizin und Biologie*, 19, 6-13

Wanner, E., and Maratsos, M.

1978 An ATN approach to comprehension. In Halle, M., Bresnan, J., and Miller, G.A. (eds.), *Linguistic Theory and Psychological Reality*, chapter 3, Cambridge, MA: MIT Press, 119–161.

Warren, T. and Gibson, E.

2005 Effects of NP type in reading cleft sentences in English. *Language and Cognitive Processes*, 20 (6), 751-767.

Wasow, T., Jaeger, F., Orr, D.

to appear Lexical Variation in Relativizer Frequency. In Simon, H. and Wiese, H.

Proceedings of the workshop on exceptions. Springer.

Webb, G. I.

1995 OPUS: An Efficient Admissible Algorithm For Unordered Search. *Journal of Artificial Intelligence Research,* 3, 431-465.

2000 Efficient Search for Association Rules. In Ramakrishnan, R. and Stolfo, S. (eds.), *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2000), Boston, MA. New York. 99-107.

2006 Discovering Significant Rules. In Ungar, L., Craven, M., Gunopulos, D., and Eliassi-Rad, T. (eds.), *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2006) Philadelphia, PA. New York. 434 - 443.

Webb, G. I., and Zhang, S.

2005 k-Optimal-Rule-Discovery. *Data Mining and Knowledge Discovery,* 10,(1). 39-79.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., and MacDonald, M. C.

2009 Experience and sentence comprehension: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58, 250-271.

Wiechmann, D.

2007 Weighing discourse-pragmatic and processing related factors governing the omission of optional relativizers in English non-subject relative clauses. paper presented at the *International Cognitive Linguistics Conference* 10 (ICLC), Krakow, July 2007.

2008a Sense-contingent lexical preferences and early parsing decisions: Corpus-evidence from local NP/S-ambiguities. *Cognitive Linguistics*, 19(3), 439-455.

2008b On the Computation of Collostruction Strength- Testing measures of associations as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory,* 4-2, 253-290.

Zipf, G.K.

1935 *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

1949 *Human Behavior and the Principle of Least-Effort*. Addison-Wesley.

Zhu, X, and Davidson, I. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. Hershey, New York.

Zwicky, A.D., and Zwicky, A.M.

1986 The Thing Is, Some That's Aren't There at All. *American Speech,* 61 (2), 182-183