# THE FUNCTION OF SPANISH AND ENGLISH RELATIVE CLAUSES IN DISCOURSE AND THEIR SEGMENTATION IN CENTERING THEORY

by

Loreley Marie Wiesemann
Master of Arts, Simon Fraser University, 2005
Bachelor of Arts, Simon Fraser University, 2002

DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In the
Department of Linguistics



SIMON FRASER UNIVERSITY

Fall 2009

# APPROVAL

**Name:** **Loreley Marie Wiesemann**

**Degree:** **Doctor of Philosophy**

**Title of Thesis:** **The function of Spanish and English relative clauses in discourse and their segmentation in Centering Theory.**

**Examining Committee:**

**Chair:** **Panayiotis Pappas**
Associate Professor, Department of Linguistics

_____

**María Teresa (Maite) Taboada**
Senior Supervisor
Associate Professor, Department of Linguistics

_____

**John Dean Mellow**
Supervisor
Associate Professor, Department of Linguistics

_____

**Nancy Hedberg**
Supervisor
Associate Professor, Department of Linguistics

_____

**Réjean Canac-Marquis**
Internal Examiner
Associate Professor, Department of French

_____

**Jeanette K. Gundel**
External Examiner
Professor, Program in Linguistics
College of Liberal Arts, University of Minnesota

**Date Defended/Approved:** December 9, 2009

# Declaration of
# Partial Copyright Licence

# ABSTRACT

This study explores the processing of English and Spanish relative clauses (RCs) in discourse. The main goal is to understand how RCs contribute to the textuality of a text and, on the basis of this understanding, to propose the most adequate method for their segmentation in Centering Theory. Centering Theory is a theory of discourse structure that models textual cohesion from one "utterance" to the next. The definition of "utterance" is thus instrumental to the application of the Centering algorithm. It is also a key step for any theory of discourse structure. To this point, there is no consensus on what the basic unit of analysis of discourse should be, though the sentence and the clause tend to be the most widely accepted proposals. An analysis of complex clauses reveals that the choice between these two segmentation categories is not always straightforward. In particular, RCs present a challenge for the discourse analyst: While they are finite clauses, they are either embedded in or dependent on another clause.

In order to address this challenge, this study investigates the processing of 200 RCs selected from English and Spanish texts belonging to four different genres. It evaluates five different approaches to their segmentation following Systemic Functional Linguistics (SFL). The evaluation takes into consideration different functional properties of RCs that are associated with their restrictiveness. The adequacy of the different segmentation approaches is measured in two ways: (a) by assessing the degree with which the focus of attention is maintained from an utterance to the next, following Constraint 1 and Rule 2 of Centering Theory; and (b) by identifying the frequency of subsequent mentions of RC entities in the unfolding discourse. The results of a factorial mixed-design ANOVA show that the segmentation approach that identified independent clauses and/or finite clauses in paratactic relations as the unit of analysis had the highest scores in all measures. Based on these findings, we are able to specify the notion of "utterance" in Centering Theory at the same time as we move towards a more systematic approach to the segmentation of discourse.

**Keywords:** Discourse structure, segmentation, relative clauses, Centering Theory, Systemic Functional Grammar

# ACKNOWLEDGEMENTS

Without the support and guidance of very talented researchers, the research work reported here would have not been completed. I would like to express my gratitude to all who have contributed to this project in its different stages of development.

I extend my deepest gratitude:

- To the members of my supervisory committee: Dr. Maite Taboada, Dr. Dean Mellow and Dr. Nancy Hedberg, who shared their knowledge of language and passion for linguistic research so generously. It was a privilege to work with them, and I thank them for their steady support, advice and encouragement.
- To my examiners: Dr. Jeanette Gundel and Dr. Réjean Canac-Marquis, for the valuable questions, comments and suggestions I received in the examination.
- To Oliver Hartmann, who guided me through the world of statistics, and to Mayo Kudo and Yuri Romero Cortés, who helped me ensure the reliability of the coding categories.
- To the graduate students in the Linguistics Department (past and present), especially Anne, Dennis, Lorna and Susan, for their fellowship and friendship.

- To the departmental staff: Carol Jackson, Rita Parmar and Grace Wattanga, who were always ready to help in any way they could.

On a personal note, I would like to thank my mother and brothers for their love and support. I want to remember my father and my grandmother, who passed away in the months leading to my defense and were not able to see me complete my degree. I owe special thanks to my friends, in Canada and in Germany, who have been there for me when I needed them the most.

Last, but certainly not least, I want to express my deepest gratitude to my husband, Henning Wiesemann, for his unwavering faith in me. Being married to an academic is no easy task, and I am extremely thankful for his patience and understanding.

This dissertation was partially funded by a SSHRC Canada Graduate Scholarship (CGS) Doctoral Scholarship (Award No. 767-2006-2048). Simon Fraser University supported my doctoral studies through Graduate Fellowships and a President's PhD Research Stipend. I would like to acknowledge the financial support I received through research assistantships from Dr. Maite Taboada, Dr. Peter Muntigl and Prof. Adam Horvath.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF NOTATIONS AND ABBREVIATIONS

1. Italics: Examples from all languages are italicized when discussed in text.
2. Single quotation marks: Glosses of non-English words are in single quotation marks.
3. Brackets: Relative clauses are in square brackets [ ].

4. Abbreviations:

| | |
|---|---|
| ANOVA: | analysis of variance |
| Cb: | backward-looking center |
| Cf: | set of forward-looking centers |
| Cl: | classifying relative clause |
| Cp: | preferred center |
| Ct: | Centering transition |
| DoC: | Degree of Cohesion |
| DP: | discourse purpose |
| DSP: | discourse segment purpose |
| Id: | identifying relative clause |
| Narr: | narrative relative clause |
| NP: | noun phrase |
| NPAH: | Noun Phrase Accessibility Hierarchy |
| NRRC: | non-restrictive relative clause |
| Post-RC: | transition to the utterance following the relative clause |
| Pre-RC: | transition to the utterance containing (or originally containing) the relative clause |
| RC: | relative clause |
| Rel: | relevance relative clause |
| RRC: | restrictive relative clause |
| RST: | Rhetorical Structure Theory |
| SFL: | Systemic Functional Linguistics |
| Subj: | subjectivity relative clause |

# CHAPTER 1: INTRODUCTION

## 1.1 Research statement

This study explores the processing of English and Spanish relative clauses (RCs) in discourse. The main goal is to understand how RCs contribute to the textuality (or texture) of a text and, on the basis of this understanding, to propose the most adequate method for their segmentation in Centering Theory (Grosz, Joshi, & Weinstein, 1995). Centering Theory is a theory of discourse structure that models textual cohesion from one 'utterance' to the next. The definition of "utterance" is thus instrumental to the application of the Centering algorithm. It is also a key step for any theory of discourse structure.

The segmentation of discourse into smaller units of analysis has been the focus of much research in studies of discourse structure (e.g., Chafe, 1988, 1992; Matsumoto, 2003; Mosegaard Hansen, 1998; Thompson & Couper-Kuhlen, 2005). As possible units of analysis, researchers have considered sentences, clauses, turns, tone groups or intonation units, utterances, propositions, speech acts, and communicative acts. The search for a minimal unit of discourse structure has also been pursued in the field of Computational Linguistics (Kameyama, 1998; Miltsakaki, 2002, 2003, 2005; Poesio, et al., 2000; Poesio, Stevenson, Di Eugenio, & Hitzeman, 2004a, 2004b; Suri & McCoy, 1994; Taboada & Hadic Zabala, 2008). Consensus on what the basic unit of analysis of discourse should be is yet to be reached. The most widely accepted proposals are

the sentence (Grosz, et al., 1995; Miltsakaki, 2002, 2003, 2005) and the finite clause (Kameyama, 1998; Suri & McCoy, 1994; Taboada & Hadic Zabala, 2008). An analysis of complex clauses reveals that the choice between these two segmentation categories is not always straightforward. In particular, RCs present a challenge for the discourse analyst. On one hand, RCs are finite clauses consisting of a subject and a predicate with a tensed verb and should be treated as finite clauses. On the other hand, RCs modify nouns and noun phrases that are constituents of another clause, a higher clause. In this sense, they belong with the main clause that contains them. In this study, we specifically examine whether RCs such as *that he likes* in Example (1) and *which was a little bit greasy* in Example (2) are processed in conjunction with the main clause that contains them, in which case they are treated as a unit, or whether these RCs are processed independently from the main clause that contains them and thus constitute a separate discourse unit.

**1**

That's the brand of jeans [$_{RC}$ that he likes]. (Kingsbury, Strassel, McLemore, & McIntyre, 1997)

**2**

I had a Reuben [$_{RC}$ which was a little bit greasy] (Kingsbury, et al., 1997)

The nature of our question demands that we look at a corpus of RCs that have been found in naturally occurring discourse to see the different ways in which the different types of RCs contribute to textual cohesion. In order to assess this contribution, we measure the degree of cohesion between an utterance containing a RC and its immediate co-text, as modelled by Centering Theory. Thus, in this study we investigate the processing of 200 RCs in English and

Spanish texts selected from four different genres: casual conversation, broadcast news, online blogs, and newspaper articles. We identify the discourse properties of the RCs in the corpus and evaluate five different approaches to their segmentation. Following Systemic Functional Linguistics (SFL) (Halliday & Matthiessen, 2004), the segmentation approaches distinguish among three types of clauses: embedded clauses (restrictive RCs), clauses in a hypotactic relation (non-restrictive RCs), and clauses in a paratactic relation. The evaluation takes into consideration different functional properties of RCs that are associated with their restrictiveness: Whether restrictive RCs identify or classify a referent; whether non-restrictive RCs continue the narrative, provide relevant background information or express an opinion.

The adequacy of the different segmentation approaches is measured in two ways. The first measure relates to the focus of attention. We measure the degree with which the focus of attention is maintained from an utterance to the next, following Constraint 1 and Rule 2 of Centering Theory (Grosz, et al., 1995). Constraint 1 states that all utterances in a discourse segment (except for the first utterance) must have a backward-looking center (Cb), that is, that each utterance in a discourse must contain an entity that links it to the previous utterance. Rule 2 states that given two utterances, a transition that maintains the focus of attention (CONTINUE or RETAIN) will be preferred to one that shifts the focus of attention (SMOOTH SHIFT or ROUGH SHIFT). These two properties capture the fact that discourse tends to be about one thing, a "topic". The second measure we use to evaluate our segmentation approaches is theory-external. Following Miltsakaki (2003), we examine the contribution of entities in RCs to textual cohesion by

identifying the frequency of subsequent mentions of those entities in the unfolding discourse. The segmentation approach that most faithfully adheres to Constraint 1 and Rule 2 of Centering and shows a higher proportion of subsequent mentions is identified as the preferred segmentation approach.

As we shall see in Chapters 6 and 7, the results of a factorial mixed-design ANOVA identified Paratactic-Clause Centering as the preferred segmentation approach. This is to say, segmenting the discourse into independent clauses and clauses in paratactic relations results in the best model of discourse cohesion according to Centering Theory.

## 1.2 Motivation

The motivations for this study are twofold. First, there is the need to understand the role of RCs in discourse processing, as noted, for example, in Kameyama (1998), Miltsakaki (2003, 2005), Poesio et al. (2000) and Poesio et al. (2004a, 2004b). Kameyama (1998, p. 108) proposed to segment RCs with their main clauses, but noted that this is an area in which further research is needed. Even though Miltsakaki (2003) was devoted to the understanding of the topicality of entities in subordinate clauses (including RCs), studies with larger corpora, in different languages, and including different types of RCs are needed to provide a comprehensive analysis of the contribution of RCs to topic continuation in discourse.

The second motivation behind this study is the need for a systematic approach to the segmentation of discourse, in particular, as noted in Taboada and Hadic Zabala (2008), for studies that involve corpus coding and analysis. In this

study, the main concern is the definition of utterance (or unit of analysis) in Centering Theory, following work by Kameyama (1998), Miltsakaki (2002, 2003, 2005), Poesio et al. (2000) and Poesio et al. (2004a, 2004b). The definition of this unit of analysis is of particular interest to Centering Theory, since it models cohesion (a type of connectivity) from one utterance to the next; the computation of the Centering algorithm relies on our definition of the notion of utterance. But defining the utterance is also a key step for any theory of discourse structure, so the applications of this research go beyond Centering Theory.

## 1.3 Research questions

As mentioned in the research statement, this study seeks to establish the best approach to RC segmentation based on the contribution of RCs to textual cohesion. Specifically, we ask the following research questions:

Q1: What is the contribution of RCs to textual cohesion, which is understood as the textual connectivity that results from co-text dependent nominal interpretation?

Q1a: If RCs are treated as separate utterances, is the model of discourse that results from a Centering analysis more or less cohesive? Do we get more or less violations of Constraint 1 and Rule 2 of Centering? Does the Centering algorithm give us a model of discourse that is more cohesive (or less cohesive) for some types of RCs than other types of RCs?

Q1b: Are entities in RCs mentioned in subsequent discourse? Are subsequent mentions co-referential with the antecedent of the RC or with an

entity inside the RC? Are there differences among the different types of RCs, so that entities in some types of RCs are more likely to be subsequently mentioned than entities in other types of RCs?

Q2: Given the answers to Q1a and Q1b, what is the approach to discourse segmentation that best captures the functional properties of RCs?

## 1.4 Contribution

We identify three areas in which our study contributes to linguistic research. First, we believe our analysis of different types of RCs in naturally occurring language will lead to a deeper understanding of the role of RCs in the processing of discourse. Second, the specification of the notion of utterance in Centering Theory is crucial for the application of the Centering algorithm. Our specification is based on empirical studies and is evaluated with both theory-internal and theory-external tests. Finally, proposing an empirically tested unit of analysis at the utterance (sentence/clause) level contributes to the general segmentation of discourse, and as a result, to a general understanding of discourse structure.

## 1.5 Theoretical assumptions

Given the emphasis we place on understanding the role of RCs in discourse cohesion, it is clear that we adopt a functionalist theory of language. We understand language as language in use and we believe, echoing Dwight Bolinger (1977), that linguistic meaning extends beyond semantic interpretation.

> Linguistic meaning covers a great deal more than reports of events in the real world. It expresses, sometimes in very obvious ways, other times in ways that are hard to ferret out, such things as what

> is the central part of the message against the peripheral part, what our attitudes are toward the person we are speaking to, how we feel about the reliability of our message, how we situate ourselves in the events we report, and many other things that make our messages not merely a recital of facts but a complex of facts and comments about facts and situations. (Bolinger, 1977, p. 4)

We believe language function to be a determinant factor in the linguistic choices we make, and therefore draw from M.A.K. Halliday's work in Systemic Functional Linguistics (SFL) to formalize our view of language. The centrality of language function is a main tenet of SFL. The Hallidayan notion of language function, however, extends beyond contextualized meaning and refers to the way in which language construes reality. The central functions of language allow speakers to create a version of reality, to enact social relationships and to determine the way in which they represent their version of reality and they enact their social relationships. These three functions of language are identified as the ideational, interpersonal and textual metafunctions in SFL (Halliday & Matthiessen, 2004, pp. 29-30). The ideational metafunction allows us to express one part of the message as central and another part as peripheral; the interpersonal metafunction allows us to express our attitude towards our listener and the degree of confidence in our message; the textual metafunction allows us to express our view of reality and our attitudes and feelings in text, that is, in a cohesive and coherent manner.

Paired with this notion of language function as ideational, interpersonal and textual meaning is a view of language as a semiotic system, that is, a system of meaning-making resources (Halliday & Matthiessen, 2004, pp. 4-5). The meaning-making resources of language are the ones employed to realize the

metafunctions of language, to make sense of our experience and to act out are social relationships (Halliday & Matthiessen, 2004, p. 29). They are stratified in three inter-related semiotic systems: (a) (discourse) semantics, (b) lexicogrammar, and (c) phonology. The first two are content planes. Discourse semantics is concerned with meaning resources at the textual level and is the stratum in which our personal experience and our social relationships are transformed into meanings. Lexicogrammar is concerned with the meaning resources at the clause level that transform discourse-semantic meanings into wording. Finally, phonology is the expression plane in which the discourse-semantic and lexicogrammatical choices we make are realized phonetically or graphically (Halliday & Matthiessen, 2004, pp. 24-5). A crucial point in this conceptualization of language is that linguistic meaning is found in all parts of the language system. In other words, SFL views language as a network of resources. We have three different systems corresponding to three different content and expression planes. Any choices we make in any of these three systems are meaningful, so that no two forms of expression can have the same meaning, nor can two different meanings be expressed by one form.

While we adopt the traditional notion of linguistic function as contextualized meaning to discuss the properties of RCs, we believe that the choice of a RC over any other type of clause is in itself meaningful, and is therefore accounted for in a clause taxonomy that distinguishes between RCs based on their function.

## 1.6 Outline of the thesis

In Chapter 2, we provide a brief discussion of discourse and discourse structure and motivate our choice of Grosz and Sidner's (1986) model of global and local discourse structure for this study. In Chapter 3, we focus on units of analysis. While we identify both units of global discourse and units of local discourse, our focus is on the latter. Drawing from SFL's system of TAXIS, we propose to reformulate existing approaches to segmentation in terms of paratactic and hypotactic clauses. Chapter 4 is devoted to RCs. There, we briefly review the structural and functional properties of Spanish and English RCs and end with a discussion of our proposed RC taxonomy. In Chapter 5, we present the methodology for our study. We motivate the choice of corpora and we describe our coding procedure. In Chapter 6, we present the results of our study and we follow them with a discussion in Chapter 7. In Chapter 8, we conclude the dissertation with a review of major findings, limitations of the study and directions for further research.

# CHAPTER 2: DISCOURSE STRUCTURE

The main goal of the study is to propose an approach to the segmentation of RCs in discourse that reflects their contribution to cohesion. Justifying the choice of cohesion as the guiding principle for RC segmentation necessitates a brief introduction to the properties of discourse and the levels of discourse structure.

Crucial to the definition of discourse is the distinction between functionalist and formalist approaches. Schiffrin (1994) pointed out that in the formalist view, discourse is defined as language above the sentence and the goal of discourse analysis is to identify the structural units of discourse and the patterns in which they can occur. The functionalist approach, on the other hand, is said to focus on language use and so it emphasizes the function language serves in human affairs. For functionalists, the analysis of discourse cannot rely solely on the linguistic system and must take into account the contexts of situation and culture in which certain meanings are expressed. The adoption of one point of view to the exclusion of the other has negative implications for the study of discourse, as it would leave discourse devoid of function or devoid of structure. Schiffrin proposed to unify formalist and functional perspectives and view discourse "as a collection of inherently contextualized utterances of language use" (p. 39), thus merging structure and function. This integration of structure and function in the definition of discourse is important, as it also points to the interrelatedness of structure and function in any aspect of discourse analysis.

Schiffrin's definition of discourse, however, is vague with respect to what contextualized means. In order to specify how context manifests itself in the utterances of a discourse, we draw from Renkema's (1993, 2004) discussion of De Beaugrande's (1981) seven criteria for textuality (Renkema, 2004, pp. 34-37)[1]: cohesion, coherence, intentionality, acceptability, informativeness, situationality and intertextuality. As Renkema (2004, p. 51) points out, not all of these criteria are considered to be equally important in the study of discourse. In fact, intentionality, cohesion and coherence are more often associated with the study of discourse structure and will therefore be explored in more detailed here.[2] Cohesion and coherence constitute different types of connectivity in a text. While cohesion refers to the connectivity that is obtained "when the interpretation of a textual element is dependent on another element in the text" (Renkema, 1993, p. 35), the term coherence has been used both to refer to the connectivity that results form the interpretation of a textual element with an element outside the text (Renkema, 1993, p. 35) and the connectivity that is manifested in discourse relations between clauses and sentences (Renkema, 2004, p.51). The first aspect of coherence concerns nominal interpretation that relies on the background knowledge or the world knowledge of the participants, or in the context of situation (Renkema, 1993, p. 35). The second aspect of coherence captures semantic and pragmatic relationships between clauses and sentences in a text (Renkema, 2004, p. 51). The third criterion for textuality, intentionality, refers to

[1] The reference to De Beaugrande (1981) is found in Renkema (2004, p. 49).

[2] Our discussion of coherence as context-dependent interpretation can be seen to include the criterion of situationality and intertextuatlity, given that context of situation and background and world knowledge may be used for nominal interpretation. The criteria of informativeness (i.e., that the message be informative) and acceptability (i.e., that the message be acceptable for the intended audience) seem to be related to Grice's (1975) maxims of quantity and relation.

the conscious intention speakers have when they utter a message, their purpose or goal (Renkema, 1993, p. 36).

Given that the concepts of coherence and cohesion are widely used in the field of discourse studies, it is important to clarify how De Beaugrande's and Renkema's notions of coherence and cohesion fit with other conceptualizations of coherence and cohesion found in the literature. The most general characterization of coherence refers to Halliday and Hasan's (1976) notion of texture, that is, the properties of a text that allow us to identify it as a unit (p. 2). Texture is dependent on the semantic system of cohesion, that is, the "relations of meaning that exist within the text, and that define it as a text" (p. 4). The individual connections between the different parts of a text are labelled cohesive ties (p. 3). Cohesion is seen as part of the system of language, and is said to be expressed in part through grammatical resources, in part through lexical resources (p. 5). Reference, substitution and ellipsis establish grammatical cohesive ties; reiteration (repetitions, synonyms, superordinates and general words) and collocation establish lexical cohesive ties. The cohesive ties established by conjunction have both a grammatical and a lexical component (p. 6). What all types of cohesive ties have in common is that in order to interpret one element in the text, the listener or reader must relate it to another element in the text. This is indeed the definition of cohesion: It is a meaning relation "in which ONE ELEMENT IS INTERPRETED BY REFERENCE TO ANOTHER" (p. 11). When we talk about interpreting elements in a text, we are indirectly making reference to a potential reader or listener. Halliday and Hasan (1976, p. 20) distinguished between internal and external aspects of cohesion:

> The internal and the external aspects of 'texture' are not wholly separable, and the reader, or listener, does not separate them when responding unconsciously to a passage of speech or writing. But when the linguist seeks to make explicit the basis on which these judgments are formed, he is bound to make observations of two rather different kinds. The one concerns relations within the language, patterns of meaning realized by grammar and vocabulary; the other concerns the relations BETWEEN the language and the relevant features of the speaker's and hearer's (or writer's and reader's) material, social and ideological environment.

While both of these aspects are seen to fall within the domain of linguistics, Halliday and Hasan's cohesive ties focus on the internal, that is, linguistic properties of the text. The study of the external factors that affect textual interpretation are said to fall within the domain of Register Theory (p. 21), a theory, as we shall see in Chapter 5 (section 5.2), that examines the influence of context of situation on a text.

A different notion of coherence distinguishes between coherence at the microstructure level of discourse and coherence at the macrostructure level of discourse (Kintsch & Van Dijk, 1978, pp. 365-6). At the microstructure level, coherence is related to sentences and propositions and is achieved through referential coherence: There are co-referential expressions in sequences of propositions. Co-referential relations between sentences do not alone achieve coherence at the macrostructure level. The notion of discourse topic is needed to bring the propositions in a text together to form a meaningful whole. The linking together of propositions to form a coherent text is said to be mediated by world knowledge, specifying the actions, both linguistic and non-linguistic, that are conventionally associated with a social act.

Bringing together these three conceptualizations of coherence and cohesion, we identify a general concept of coherence that corresponds to Halliday and Hasan's notion of texture and De Beaugrande's notion of textuality. This general concept of coherence relates to the properties of a set of sentences that make them a text. In this sense, a text is coherent when it can be identified as a text. A text can be identified as such on the basis of internal and external factors. Internal factors are co-text dependent. This is what Halliday and Hasan (1976), and Renkema (1993) (drawing from De Beaugrande, 1981) labelled cohesion. As mentioned before, cohesion extends beyond reference and includes substitution, ellipsis and lexical cohesion. There is, however, one aspect of Halliday and Hasan's cohesion that we will not include under co-text dependent interpretation, namely conjunction. The relationship between clauses and sentences in a text, called conjunction by Halliday and Hasan, and coherence by Renkema (2004), is also known as coherence relations (Mann & Thompson, 1988). Coherence relations (e.g., conjunction ties) are different from the other cohesive ties: The relationship between two clauses in a text is inferred in relation to the speaker's or writer's goals, and we can therefore not speak of one element in the text being interpreted in reference to another element in the text. In order to interpret coherence relations, we move away from internal factors in the direction of external factors, and begin to consider the speaker's intentions and desired effects. The overall or primary goal of the speaker is captured in DeBeaugrande's notion of intentionality, which corresponds to Kintsch and Van Dijk's notion of discourse topic. The final aspect of textuality we consider is also external to the text and relates (a) to the world knowledge that allows us to link

together propositions to form a coherent text by specifying the actions that are conventionally associated with an event, and (b) to the particular context of situation in which this event takes place. This type of coherence we have called context-dependent interpretation.

In sum, there is a general notion of coherence that refers to a text that has texture, that is, a text that can be identified as a text[3]. The narrow sense of coherence refers to the world and situational knowledge needed to interpret a specific text. Coherence relations concern the relations that can be inferred between two clauses or set of clauses in a text. Finally, we use cohesion to refer the meaning relation that allows us to interpret one element in a text in reference to another element in the text.

Having specified the components of textuality, we can now provide a definition for discourse. A discourse is a collection of contextualized utterances, that is, a collection of utterances that are cohesive, coherent and intentional (in the sense of purposeful). It may consist of one or more utterances that a speaker utters in order to achieve a goal. These utterances are interpreted in relation to other elements in the same text, its co-text, and elements outside the text, its context. We illustrate these concepts with an example.

Example (3) below is a fragment of a transcript from a telephone conversation taken from the English CallHome Corpus collected by the Linguistics Data Consortium (Kingsbury, et al., 1997). In this fragment, B is

[3] The term text is used in the sense of product and it may refer to either written or spoken products. We are analyzing the final product and not the process, as we do not have access to the contextual information needed to analyze text as it is (was) being produced.

telling A about her trip to Taipei. If we listened to the conversation (or read the entire transcript), we would intuitively recognize this sequence of utterances as a discourse (or a text). If we just concentrate on the first four lines, we will also be able to see how our definition of discourse helps us identify this sequence of utterances as a discourse. In line 1, B tells A that she wants to tell her something about her stop over in Chicago. The sequence of utterances in Example (3) is thus purposeful: The speaker's goal is to narrate something that happened to her. If we look closely at line 2, we see that in order to interpret some elements in this utterance, we need to link them with other elements in the preceding text: The pronominal adverb *there* is linked to *Chicago* in the first utterance. This is what we call cohesion, or co-text dependent nominal interpretation. In order to interpret an element like *Chicago*, we rely on background and world knowledge. In uttering *Chicago*, the speaker assumes that the hearer knows that *Chicago* is a city in the United States of America, a usual stop in long-haul flights. In this case, an element of the text is linked to our background or world knowledge; we can interpret it because we have this knowledge. This is what we call coherence in the narrow sense, or context-dependent nominal interpretation. Line 2 consists of two clauses: *My sister Agnes was right there to meet me* and *so that was very nice*. The second clause is an evaluation of the proposition expressed in the first clause. There is an evaluation relation between these two clauses. This type of connectivity is what we call relational coherence.

**3**

ENG5208 (Conversation)

1 B: I want to tell you that when I got to Chicago

2 B: My sister Agnes was right there to meet me so that was very nice

| | | |
|---|---|---|
| 3 | A: | oh that's nice |
| 4 | B: | and then she waited until I got on the plane |
| 5 | B: | and then after I got on the plane um |
| 6 | B: | there was a mixup because |
| 7 | B: | the people in Chica- I didn't realize this until I got to San Francisco |
| 8 | B: | that the people in Chicago they took the wrong ticket from me |
| 9 | A: | oh |
| 10 | B: | they took the ticket that says from San Francisco to Taipei ((laugh)) and so when they |
| 11 | A: | oh and then what happened? |
| 12 | B: | when they got to San Francisco |
| 13 | B: | the lady said I could really be charged for a second ticket and I said oh no ((laugh)) |
| 14 | A: | oh |
| 15 | B: | but she didn't she didn't charge me she she called Chicago and she recognized the mistake they made |
| 16 | B: | well that caused great confusion for me because when I got on the plane in Chicago |
| 17 | B: | I had the wrong seat number somebody else had that seat number that's because I was ((laugh)) dealing with a different ticket. |
| 18 | A: | oh my goodness |
| 19 | B: | so anyhow it all worked out but it was |
| 20 | A: | well I'm glad |
| 21 | B: | a little confusion |
| 22 | B: | but while I was on after I got on the plane in Chicago and they were they were still boarding it was very very crowded it was a seven forty seven |
| 23 | B: | All of a sudden they called my name |
| 24 | B: | to come to the uh to the uh front of the plane so I finally got through all the crowd and got there |
| 25 | A: | oh |
| 26 | B: | and they told me to got out to the desk and here my nephew Tommy C K |
| 27 | B: | the one that has cancer |
| 28 | A: | mhm yes |
| 29 | B: | he came to see me and he he didn't get there on time and so my sister Agnes |
| 30 | B: | she went up to the man and asked if if they could call me off the plane ((laugh)) |
| 31 | A: | oh my good- |
| 32 | B: | ((breath)) so I got off the you know I got on on time to see him so that was very nice yeah ((laugh)) |

If we were asked to say what this discourse is about, we would probably come up with more than one label: B's flight to Taipei, B's relatives greeting her in Chicago, and the mix-up with B's ticket. In fact, these labels can be assigned to different sequences of utterances. From line 1 to 4 and line 22 to 32, B is telling A

that her sister and nephew went to the airport to greet her. From line 5 to line 21, B is telling A that there was a mix-up with her boarding pass. And both groups of utterances develop the larger topic: B's flight to Taipei. This grouping of utterances reveals that discourse is structured. Johnstone (2002) argued that the fact that discourse is structured is what enables us to break a text into smaller units, which themselves are composed of other units. Looking at Example (3), we can say that the sequence of utterances depicting the mix-up with B's boarding pass is a digression that is preceded and followed (i.e., enclosed) by the sequence of utterances narrating B's relatives visit at the airport[4]. Recognizing the units in a discourse allows us to see the relations between them and as a result, the overall structure of the discourse.

In this study, we adopt Grosz and Sidner's (1986) computational model as our model of discourse structure. We have chosen to follow Grosz and Sidner's model of discourse structure because it makes a distinction between global and local discourse structure. At the global level, we model the unfolding of discourse in terms of discourse segments and the intentions (purposes) associated with them. At the local level, we model the unfolding of discourse from one utterance to the next, focusing on entity realization, that is, the people and objects that are introduced to the discourse. We believe that this distinction between immediate or local structure and global structure allows us to model the three criteria of textuality we see as defining discourse: coherence, cohesion and intentionality.

[4] In Conversation Analysis (CA), this type of digression is termed side sequence. It is worth noting, that this same example could be analyzed in terms of turns within CA. Such analysis would focus on the organization of conversation in turns, which are believed to be responsible for the way in which the interaction unfolds (Sacks, Schegloff, & Jefferson, 1974). This type of analysis will not be pursued here, since it only applies to one of the genres being examined.

The criteria of intentionality and coherence can be accounted for at the global level whereas cohesion can be accounted for at the local level. Since we want to assess the contribution of RCs to textual cohesion, this study focuses on the modelling of cohesion at the local level. Centering Theory (Grosz, et al., 1995), which was developed as the local component of Grosz and Sidner's theory of discourse structure, is thus the ideal candidate to model cohesion in our data.

## 2.1. Global discourse structure

The distinction between global and local structure concerns the type of context needed for utterance interpretation. In global focusing, both the overall discourse and the situational context affect utterance interpretation. In local focusing, it is the immediate linguistic co-text, that is, the preceding utterance, what determines the interpretation of the following utterance (Grosz & Sidner, 1998). In this section, we review Grosz and Sidner's (1986) model of global structure and argue that this account of global structure allows us to model one of the criteria of textuality we identified in our definition of discourse: intentionality.

In their model of global structure, Grosz and Sidner (1986) identified three basic components of discourse structure: (a) linguistic structure, (b) intentional structure, and (c) attentional state. These three components model different aspects of discourse structure: Linguistic structure deals with the linguistic properties of utterances, intentional structure with the intentions associated with those utterances and attentional state with the salient semantic entities, relations and intentions at any point in the flow of discourse (Grosz & Sidner, 1986, p. 177). The first component, linguistic structure, consists itself of two components:

discourse segments and the relations between discourse segments. We shall see in the next chapter that utterances can be grouped together into larger units, which are called discourse segments. The structure of these discourse segments can interact with the utterances in the discourse in two ways: The actual linguistic expressions in an utterance may reveal information about the structure of the discourse, or the structure of the discourse may constrain the interpretation of the linguistic expression in an utterance (p. 177).

The second component, intentional structure, is concerned with the intentions and goals participants have which have led them to engage in discourse. Grosz and Sidner (1986, p. 178) distinguished between discourse purpose (DP) and discourse segment purpose (DSP). The DP is the most fundamental purpose of the discourse, whereas the DSP is the intention that underlies each discourse segment and contributes to the realization of the fundamental purpose of the discourse, the DP. A DSP can have one of two structural relations to other DSPs in the discourse, depending on how the satisfaction of one DSP contributes to the satisfaction of another. If the satisfaction of DSP1 partly contributes to the satisfaction of DSP2, then DSP2 dominates DSP1. This is a dominance relation. If DSP1 must be satisfied before DSP2, then DSP1 satisfaction-precedes DSP2. The relation between the DSPs is one of satisfaction-precedence (p. 179).

The final component of Grosz and Sidner's discourse theory is attentional state, which is defined as "an abstraction of the participant's focus as their discourse unfolds" (p. 179). The attentional state records objects and relations

that are salient in the discourse. The attentional state changes as the discourse unfolds, as some objects may become more salient and others fade away. Grosz and Sidner's model captures the changes in the attentional state with the process of focusing (p. 179). In this process, each discourse segment is associated with a focus space that contains the salient entities for that discourse segment (p. 179). As the discourse unfolds, the focus space for each discourse segment is placed on the focus space stack. There, two types of operations can take place between focus spaces: pushing and popping, depending on the relations among DSPs. Grosz and Sidner (1986, p. 180) explained that "[a] push occurs when the DSP for a new segment contributes to the DSP for the immediately preceding segment." That is, the focus space for the new DSP is pushed on the focus space stack. The alternative operation is popping: Focus spaces are popped (removed) from the stack before the focus space for the new DSP can be inserted. This happens "[w]hen the DSP contributes to some intention higher in the dominance hierarchy" (Grosz & Sidner, 1986, p. 180).

A new look at our discourse example will help us bring these concepts together. In (4) below, we reproduce lines 5 to 21 from Example (3), which we had identified as a digression, an embedded segment within a larger segment.

**4**

ENG5208 (Conversation)

| | | |
|---|---|---|
| 5 | B: | and then after I got on the plane um |
| 6 | B: | there was a mixup because |
| 7 | B: | the people in Chica- I didn't realize this until I got to San Francisco |
| 8 | B: | that the people in Chicago they took the wrong ticket from me |
| 9 | A: | oh |
| 10 | B: | they took the ticket that says from San Francisco to Taipei ((laugh)) and so when they |

11 A: oh and then what happened?
12 B: when they got to San Francisco
13 B: the lady said I could really be charged for a second ticket and I said oh no ((laugh))
14 A: oh
15 B: but she didn't she didn't charge me she she called Chicago and she recognized the mistake they made
16 B: well that caused great confusion for me because when I got on the plane in Chicago
17 B: I had the wrong seat number somebody else had that seat number that's because I was ((laugh)) dealing with a different ticket.
18 A: oh my goodness
19 B: so anyhow it all worked out but it was
20 A: well I'm glad
21 B: a little confusion

The utterances in lines 5 to 21 constitute the linguistic structure of this discourse. These utterances correspond to one discourse segment and have a discourse purpose (DP) associated with it: B is telling A about the boarding pass mix-up. Within this segment, there are smaller discourse segments with their own discourse segment purpose (DSPs): Lines 5 to 10 explain what the mix-up was; lines 11 to 15 what the possible consequence may have been, and so forth. Each discourse segment is associated with a focus space. The focus space for the discourse segment in lines 5 to 10 contains the most salient entities in that segment: B, PEOPLE IN CHICAGO, TICKET. Once this discourse segment has been processed, we move on to the next. Since the DSP of the segment in lines 11 to 15 contributes to the DSP of the segment in lines 5 to 10, we have a push, that is, the focus space associated with the segment in lines 11 to 15 is stacked on top of the focus space associated with the segment in lines 5 to 10. This means that the set of salient entities in the previous focus space is still accessible when processing the current segment.

In formulating the DP and the DSPs in Example (4), we model the way in which the discourse realizes the participants' intentions. When we spell out the relations of domination or satisfaction-precedence between two different DSPs, we model the way in which smaller segments of discourse contribute to the realization of those intentions. Identifying the discourse and discourse segment purposes and the relationships among them thus enables us to provide a theoretical model of intention in discourse.

In addition to intentionality, the textuality or texture of a discourse manifests itself in connectivity: in coherence relations and in cohesive links. In order to model the coherence relations in discourse, we need to set aside Grosz and Sidner's (1986) model of global discourse structure in favour of a theory that provides a more descriptive account of how utterances are linked together in text. While Grosz and Sidner's model specifies two types of relationships between DSPs, the focus is on the realization of purposes and not on connectivity between utterances. For the latter, Rhetorical Structure Theory (RST) provides a more suitable linguistic model. RST (Mann & Thompson, 1988) is a theory of text structure that describes the relations that hold between spans of text, that is, between clauses or groups of clauses in a text. Relations are identified on functional criteria: The main component in a rhetorical relation is the effect, which is a plausibility judgment the analyst makes about the goals the author may have had in producing the text. Some goals are more central than others. RST captures this asymmetry by making a distinction between Nucleus and Satellite elements in a relation: The text span in the rhetorical relation that realizes the more central goal of the writer is the Nucleus, whereas the text span

that realizes a supplementary or ancillary goal is the Satellite. In this way, an RST analysis provides not only a functional account of the relations that link spans of texts, but also a representation of the hierarchical structure of the text in terms of nuclear and satellite spans.

Given that the focus of our study is on the contribution of RCs to textual cohesion and not to coherence, an RST analysis will not be further pursued. Our goal is to provide a linguistic model of discourse cohesion that will allow us to assess the role RCs play in co-text-dependent connectivity. Grosz and Sidner (1986, p. 191) argued that this type of connectivity is best accounted for at the local level of discourse structure, with Centering Theory.

## 2.2 Local discourse structure: Centering Theory

Centering Theory (Grosz, et al., 1995; Walker, Joshi, & Prince, 1998) is a theory of local discourse structure that is concerned with the relationship between participants' attention and the choice of referential expression. It provides a mechanism to determine the degree of local cohesion[5] between two given utterances on the basis of the entities realized in those utterances: A discourse will be perceived as more coherent (in the sense of texture and textuality, that is, more connected, woven together) if the same entities are kept from one utterance to the next; a discourse will be perceived as less coherent if the attention shifts to new entities from one utterance to the next.

[5] Grosz et al. (1995, p. 204) use the term local coherence to refer to the coherence among the utterance in a segment. Since the major focus of the theory is on the interaction between local coherence and choice of referring expression (p. 204), we can equate the term local coherence to Kintsch and Van Dijk's (1978) referential coherence, which as we saw earlier, corresponds to our notion of cohesion.

Examples (5) and (6), from Walker et al. (1998, p. 1), illustrate these differences in perceived coherence.

**5**

(a) **Jeff** helped Dick wash the car.

(b) **He** washed the windows as Dick waxed the car.

(c) **He** soaped a pane.

**6**

(a) **Jeff** helped Dick wash the car.

(b) **He** washed the windows as Dick waxed the car.

(c) He buffed the hood.

The discourse in (6) is perceived as less coherent than the discourse in (5). In (5), the utterances (a), (b) and (c) are about **Jeff**, so the topic is maintained in all three utterances. In (6) on the other hand, while the utterances (a) and (b) are about **Jeff**, utterance (c) is about Dick. The topic has been shifted. The choice of referring expression, the pronoun he, appears to be in conflict with the attentional state of its referent. Centering Theory allows us to explain these perceived differences in coherence by modelling the cohesive ties between utterances. In order to explain the difference in perceived coherence between (5) and (6) in terms of their local cohesion, we must introduce Centering Theory.

Grosz et al. (1995) explained that Centering Theory originated from two strands of work: Grosz and Sidner's research on the inferences needed for anaphora interpretation and Joshi, Kuhn and Weinstein's work on the inferences needed to integrate utterance meaning into discourse meaning. One of the basic assumptions of Centering is that different referring expressions place different

inference loads on the listener: There is the amount of inference needed to resolve anaphoric relations and there is the amount of inference needed to integrate the meaning of the utterance into the discourse meaning (Grosz, et al., 1995). The burden, however, does not rest solely on the listener: It is assumed that speakers tend to minimize the processing burden on their listeners by producing utterances that are optimally relevant (Ballantyne, 2004). In our discussion of global structure, we mentioned that the participant's attentional state is updated as the discourse unfolds from one utterance to the next. This updating of the attentional state is related to the amount of inference needed to process a referring expression: The inference process is said to be easier when the referring expression is associated with a salient entity (Beaver, 2000). Centering models this inferencing process in terms of transitions. Thus the salience of an entity in an utterance as well as its relation to the preceding discourse are key components of Centering Theory.

The starting point for Centering Theory are the semantic entities found in the discourse model of each utterance in the discourse segment. These are the centers. They are semantic objects, discourse constructs, that link the utterances that contain them (Grosz, et al., 1995, p. 208). For each utterance, the list of centers is ranked in terms of discourse salience in what is called the Cf-set, that is, the set of forward-looking centers. Walker et al. (1998) present a basic (almost language-neutral, English-based) Cf-template which follows the ranking proposed in Brennan, Walker Friedmann and Pollard (1987) and lists entities following grammatical role, as shown in (7).

7

Subject > Object(s) > Other

This ranking is said to be partial, meaning that entities in subject position are the most salient, but making no specific claims about the salience of direct objects with respect to indirect objects or other subcategorizations of the main verb. In other words, this ranking highlights the salience of subjects in discourse.

The ranking provided in (7) is the ranking usually adopted in the analysis of English. However, not all languages code salience in the same way. Centering Theory allows for language diversity in the ranking of entities in the Cf-list. In other words, Cf-rankings are language-specific (Walker, Iida, & Cote, 1994). For instance, Kuno's notion of empathy, which was introduced in Centering by Kameyama (1985) and refers to speaker identification with an event, has been shown to be so crucial in the ranking of entities in languages such as Japanese (Walker, et al., 1994)[6], Turkish (Turan, 1996), Italian (Di Eugenio, 1996, 1998), and Spanish (Taboada, 2002, 2008) that it overrides grammatical function. In fact, the adequacy of grammatical function to capture salience has further been questioned, in particular in relation to languages with free word order (as opposed to languages with fixed word order like English). Strube and Hahn (1996, 1999) proposed functional information structure (IS) as the ranking criterion for the Cf in German. According to the 1999 version of their ranking, entities that are discourse-old or hearer-old are ranked higher than entities that are discourse-new or hearer-new.

[6] In addition, Walker et al. (1994) argued that grammatical topics are more salient than subjects and objects in Japanese discourse.

Researchers have explored the adequacy of these different Cf-ranking criteria in languages as diverse as English, Yapese and Korean. Their results showed that different languages favour certain approaches over others. For example, Ballantyne (2004) examined the adequacy of grammatical role, linear order and Gundel, Hedberg and Zacharski's (1993) Givenness Hierarchy as ranking criteria in Yapese (an Oceanic language spoken on the islands of Yap). The results of the Centering analysis revealed that Gundel et al.'s Givenness Hierarchy resulted in a greater number of low-cost transitions. Poesio et al. (2004a) tested grammatical function, linear order and Strube and Hahn's information structure as ranking criteria in English. Strube and Hahn's approach was the one that yielded the highest proportion of cohesive transitions. Roh and Lee (2006) found surface word order, as opposed to grammatical function, simplified information status and type of pro-form, to be the best suited ranking criterion for Korean. The results of these three studies point to different ranking criteria, and thus lend support to the claim that the ranking of entities in the Cf-list is language-specific (Cf-rankings for English and Spanish are provided in section 5.4.1.2).

Once the entities that are realized in an utterance are ranked in the Cf-list, Centering Theory selects the highest element in the list as the preferred center (Cp) (Grosz, et al., 1995). The Cp is the most salient entity in the utterance and is said to make a prediction with respect to the next utterance: The Cp of the current utterance is likely (but not necessarily) the Cb, the backward-looking center, of the next utterance. The Cb is a special center since it links two utterances: The highest element in the Cf of the previous utterance that is

realized in the Cf of the current utterance becomes the Cb of the current utterance. Even though the Cp is the most salient entity in the utterance, Grosz, Joshi and Weinstein (1983) argue that the Cb is the most central entity in the utterance, as it is the one that holds the discourse together.

The Cb and the Cp can realize the same entity or different ones. The different configurations of the Cb and the Cp yield four possible transitions types between two utterances, which are provided in Table 1 (adapted from Walker, et al., 1998, p. 6). The vertical axis compares the realization of the Cp and the Cb in the current utterance. When they realize the same entity, the transition can be either a CONTINUE or a SMOOTH SHIFT. If they realize different entities, the choice falls between a RETAIN and a ROUGH SHIFT. The horizontal axis compares the realization of the Cb in the current utterance with the Cb of the previous utterance. Should they be the same, then two types of transition are possible: CONTINUE and RETAIN. If they differ, the transition will be a SHIFT.

**Table 1. Centering transitions**

| | **$Cb(U_{i-1}) = Cb(U_i)$**<br>**$Cb(U_{i-1}) = o$** | **$Cb(U_{i-1}) \neq Cb(U_i)$** |
|---|---|---|
| **$Cb(U_i) = Cp(U_i)$** | CONTINUE | SMOOTH SHIFT |
| **$Cb(U_i) \neq Cp(U_i)$** | RETAIN | ROUGH SHIFT |

Kibble (1999) classified these Centering transitions according to the principles of cohesion and salience[7]. In his classification, the principle of cohesion accounts for the realization of the previous and current Cb, whereas the principle of salience accounts for the realization of the current Cb and Cp. This

[7] Kibble uses the terms of cohesion and salience in a narrow sense to refer to the realizations of the Cp and the Cb.

means that CONTINUE transitions adhere to both principles, whereas RETAIN and SMOOTH SHIFT transitions favour one principle over the other: RETAIN favours cohesion and SMOOTH SHIFT salience.

The transitions in Table 1 capture most of the realizations of the Cp and the Cb, but they fail to account for the relationship between the Cp of the previous utterance and the Cb of the current utterance. Strube and Hahn (1996, 1999) complemented traditional Centering transitions with CHEAP and EXPENSIVE transitions. CHEAP transitions are those in which the Cp of the previous utterances predicts the Cb of the current utterance. EXPENSIVE transitions are those in which the prediction does not hold. So, for example, the transition between two utterances is a CHEAP-CONTINUE when the previous utterance and the current utterance have the same Cb, the Cp and the Cb of the current utterance refer to the same entity and the Cp of the previous utterance and the Cb of the current utterance coincide. The transition is an EXPENSIVE-CONTINUE when the last condition is not met. Even though the addition of CHEAP and EXPENSIVE transitions to the Centering repertoire seems theoretically sound given the relation between the previous Cp and the current Cb, research on natural data has not been able to show a strong link between cohesion and CHEAP transitions, and thus it remains an area for further research.

In addition to centers and transitions, Centering Theory posits a series of constraints and rules, which are provided in (8) and (9) below. The first constraint is known as Constraint 1 of Centering and has two interpretations. The strong version states that each utterance must have precisely one Cb; the weak

version states that each utterance must have at most one Cb. The second constraint concerns the population of the list of forward-looking centers and requires that the entities listed in the Cf of an utterance be realized in that utterance. The last constraint defines the Cb and at the same time imposes a locality condition on it: The Cb is the highest element of the previous utterance realized in the current utterance; it must be listed in the previous Cf-list.

**8 Constraints**

For each utterance $U_i$ in a discourse segment D consisting of Utterances $U_1$, ...., $U_m$:

1. There is precisely one backward-looking center Cb ($U_i$, D).

2. Every element of the forward centers list, Cf ($U_i$, D), must be realized in $U_i$.

3. The center, Cb ($U_i$,D), is the highest-ranked element of Cf ($U_{i-1}$, D) that is realized in $U_i$. (Walker, et al., 1998, p. 3)

There is a second condition imposed on the Cb besides the locality stipulation: Rule 1 of Centering demands that the Cb be realized as a pronoun if any other entity in the same utterance is realized as a pronoun (or a zero pronoun, in languages that allow those, e.g., Spanish). Gordon, Grosz and Gilliom (1993) found psycholinguistic evidence to support this preference for the Cb to be realized as a pronoun in a series of reading time experiments. Longer reading times were observed when the Cb was realized as a name or a definite description compared to a pronoun. Further research showed that this preference for pronominal form is associated with grammatical role: The Cb tends to be realized as the grammatical subject, which tends to be pronominal (Gordon & Chan, 1995; Gordon & Hendrick, 1997).

**9 Rules**

For each utterance $U_i$ in a discourse segment D consisting of Utterances $U_1$, ...., $U_m$:

1. If some element of Cf ($U_{i-1}$, D) is realized as a pronoun in $U_i$, then so is Cb ($U_i$, D).

2. Transition states are ordered. The CONTINUE transition is preferred to the RETAIN transition, which is preferred to the SMOOTH SHIFT transition, which is preferred to the ROUGH-SHIFT transition. (Walker, et al., 1998, p. 4)

The final rule, Rule 2, provides the preferred ranking of transitions in Centering. This order of preference is based on the inferential load placed upon the listener, and is said to reflect attentional state changes in the listener (Grosz, et al., 1995). Psycholinguistic research, however, has only shown weak evidence in support of this rule. In particular, Gordon et al. (1993) found that when the shifts were unprepared, CONTINUE transitions were read faster than shifts transitions. When the shifts were prepared (by following a RETAIN transition), SHIFTS transitions were read faster than CONTINUE transitions. The weak evidence for the ranking of transitions may explain why Rule 2 has not been entirely validated in corpus studies of natural language (e.g., Byron & Stent, 1998; Poesio, et al., 2004a; Taboada & Hadic Zabala, 2008; Taboada & Wiesemann, in press). Kibble (Kibble, 2001) and Kibble and Power (2004) proposed to incorporate the principles of cheapness and continuity to the principles of salience and cohesion in the ranking of Centering transitions. The principle of cheapness captures the prediction that the Cp of the previous utterance tends to become the Cb of the following utterance. The principle of continuity states that the previous and the current utterance should at least have one entity in common (which can be seen as a reformulation of Rule 1 of Centering). Kibble and Power (2004) ranked

cheapness ahead of salience and cohesion, and these two ahead of continuity. This ranking would result in the preferred order of transitions provided in (10) and supported in Table 2. At this point, there is no strong evidence for this ranking, except for the findings of Taboada and Wiesemann (in press).

**10**

CONTINUE > RETAIN, SMOOTH SHIFT > ROUGH SHIFT > NOCB

**Table 2. Ranking of Centering transitions**

| | **CHEAPNESS** | **SALIENCE** | **COHESION** | **CONTINUITY** |
|---|---|---|---|---|
| CONTINUE | ? | ✓ | ✓ | ✓ |
| RETAIN | ? | X | ✓ | ✓ |
| SMOOTH | ? | ✓ | X | ✓ |
| ROUGH | ? | X | X | ✓ |
| NOCB | X | X | X | X |

Now that the Centering algorithm has been introduced, it is possible to account for the differences in perceived coherence noticed for Examples (5) and (6) above. As shown in Table 3, the Cb of utterance (5b) is maintained in utterance (5c). Since the Cb and the Cp of (5c) coincide, we have a CONTINUE transition. In (6c) the current Cb is not the same as the previous Cb. This constitutes a shift in the focus of attention. Because the Cb and Cp of (6c) coincide, we have a SMOOTH SHIFT. According to Rule 2 of Centering, CONTINUE transitions are preferred to any other type of transition, including SMOOTH SHIFTS. It follows then, that the discourse in (5) is more coherent (and therefore easier to process) than the discourse in (6).

**Table 3. Centering transitions for Examples (5) and (6)**

| Utterance | Cf-list | Cp | Cb | Ct |
|---|---|---|---|---|
| (**5**a) **Jeff** helped Dick wash the car. | JEFF>DICK>CAR | JEFF | 0 | NOCB |
| (**5**b) **He** washed the windows as Dick waxed the car. | JEFF>WINDOWS>DICK>CAR | JEFF | JEFF | CONTINUE |
| (**5**c) **He** soaped a pane. | JEFF>PANE | JEFF | JEFF | CONTINUE |
| (**6**a) **Jeff** helped Dick wash the car. | JEFF>DICK>CAR | JEFF | 0 | NOCB |
| (**6**b) **He** washed the windows as Dick waxed the car. | JEFF>WINDOWS>DICK>CAR | JEFF | JEFF | CONTINUE |
| (**6**c) He buffed the hood. | DICK>HOOD | DICK | DICK | SMOOTH |

These particular examples provide a strong argument in favour of Centering Theory over other models of discourse processing. Walker et al. (1998) argued that even though other semantic or inferential theories of discourse may be able to resolve the pronominal reference in Examples (5c) and (6c), only Centering can account for the differences in perceived coherence.

Additional validation for our choice of Centering Theory for this study comes from the fact that Centering has been widely applied in a variety of fields and to languages so diverse as the ones listed in Table 4.

**Table 4. Languages to which Centering has been applied**

| Language | Source |
|---|---|
| English | (Brennan, Walker Friedman, & Pollard, 1987; Chambers & Smyth, 1998; Miltsakaki, 2002; Poesio, et al., 2000; Poesio, et al., 2004a, 2004b; Strube & Hahn, 1999) |
| Cakchiquel Mayan | (Hedberg, in press; Hedberg & Dueck, 1999) |
| Chinese | (Yeh & Chen, 2004) |
| Dutch | (Maes, 1997) |
| French | (Frossard, Cardebat, & Nespoulous, 2001) |

| | |
|---|---|
| German | (Rambow, 1993; Strube & Hahn, 1996, 1999) |
| Greek | (Dimitriadis, 1995, 1996; Miltsakaki, 2001, 2002, 2003, 2005) |
| Hindi | (Prasad & Strube, 2000) |
| Italian | (Di Eugenio, 1996) (Di Eugenio, 1998) |
| Japanese | (Fais, 2004; Fais & Yamura-Takei, 2003; Iida, 1998; Kameyama, 1985; Okumura & Tamura, 1996; Walker, et al., 1994) |
| Korean | (Roh & Lee, 2006) |
| Polish | (Stys & Zemke, 1995) |
| Spanish | (Taboada, 2002, 2008; Taboada & Hadic Zabala, 2008; Taboada & Wiesemann, in press) |
| Turkish | (Hoffman, 1996; Turan, 1996) |
| Yapese | (Ballantyne, 2004) |
| Yiddish | (Prince, 1998) |

Particularly noteworthy is the role of Centering in anaphora resolution, where a considerable body of research has found a relationship between attentional state and form of referring expressions in a variety of languages (e.g., Di Eugenio, 1996, 1998; Dimitriadis, 1995, 1996; Frossard, et al., 2001; Hedberg, in press; Hedberg & Dueck, 1999; Iida, 1998; Maes, 1997; Miltsakaki, 2001; Taboada, 2002, 2008; Turan, 1996). Both Cf-ranking and Rule 2 of Centering have been adopted as criteria for pronoun resolution either in their original formulations (e.g., Brennan, et al., 1987; Di Eugenio, 1996, 1998; Dimitriadis, 1995, 1996; Miltsakaki, 2002) or in modified versions (Baldwin, 1995; Kehler, 1993; Kim, Cho, & Seo, 1999; Prasad & Strube, 2000; Strube & Hahn, 1999; Tetreault, 1999, 2001).

Other applications of Centering Theory fall within automatic language annotation and generation and include areas such as automatic discourse

representation (Barzilay & Lapata, 2005; Forbes & Miltsakaki, 2002), automatic text generation (Beaver, 2000; Karamanis, Mellish, Poesio, & Oberlander, 2009), pronoun generation (Yuksel & Bozsahin, 2002), sentence compression (Clarke & Lapata, 2007), machine translation (Hoffman, 1996; Stys & Zemke, 1995), sentence ordering (Karamanis, 2006; Karamanis, Poesio, Mellish, & Oberlander, 2004) and text planning (Kibble, 1999; Kibble & Power, 1999, 2004).

Centering Theory, however, is not without critics. Among the major criticisms raised against the theory we find (a) those concerning Centering's ability to model discourse structure, (b) those challenging the locality of the Cb and (c) those concerning the implementation of the Centering algorithm.

Asher (2004, p. 259) questioned the validity of Centering as a theory of discourse structure given that it is unable to provide an adequate model of contextual and background knowledge (see also Fais & Yamura-Takei, 2003). This point is only valid if Centering Theory is taken to be an all-encompassing theory of discourse. In this study, we followed Grosz and Sidner's model of discourse structure because it provided a layered view of discourse structure. In this multi-level approach to the analysis of discourse, Centering Theory is responsible for modelling co-text dependent connectivity, or cohesion. It is in combination with other theoretical approaches that Centering can provide a complete model of discourse structure. For instance, Cristea, Ide and Romary's (1998) Veins Theory (VT) combines elements of Centering Theory and Rhetorical Structure Theory (RST) in order to account for both local and global discourse structure. In their analysis, they identified the coherence relations between text

spans as well as the list of semantic entities realized in those spans. Their data showed a preference for nuclear spans of text to be linked referentially to other nuclear spans in the discourse, rather than to satellites, which is evidence of an interaction between hierarchical structure and cohesion[8].

The second issue of contest is the locality constraint on the Cb, which does not allow Centering to account for long-distance pronominal resolution (e.g., Dimitriadis, 1996; Hitzeman & Poesio, 1998; Iida, 1998; Kruijff-Korbayova & Hajicova, 1997). Suri and McCoy (1994) proposed an alternative algorithm, the RAFT/RAPR (Revised Algorithms for Focus Tracking and Revised Algorithms for Pronoun Resolution), which lacks the locality stipulation we find in Centering Theory and thus allows for the resolution of pronouns whose antecedents are not in the immediately preceding utterance.

Finally, the criticisms concerning the implementation of the Centering algorithm relate to the inclusion of information on structural parallelism in the ranking of entities in the Cf-list and to the specification of the unit of analysis. Kehler (1997) and Chambers and Smyth (1998) argued that a parallel effect, whereby a pronoun co-refers with an antecedent that has the same grammatical function, has been observed in pronoun interpretation and is not fully incorporated in the Centering algorithm. Okumura and Tamura (1996) and Callaway and Lester (2002) observed that the underspecification of the notion of utterance leads to difficulties in the implementation of the Centering algorithm.

[8] It is important to note, however, that Karamanis (2007) attempted to improve the performance of the Centering algorithm by including rhetorical relations in his analysis and found that this inclusion did not have a significant impact on the proportion of NoCb transitions.

The criticisms outlined here point to areas in Centering that necessitate further research. This study in particular, addresses the shortcoming Okumura and Tamura (1996) and Callaway and Lester (2002) make reference to. As we shall see in Chapter 3, the specification of the notion of utterance has received considerable attention in the field, and it is only to be hoped that research will follow suit in other deficit areas as well.

## 2.3 Summary

To sum up, in this chapter, we introduced a view of discourse in which intentionality, coherence and cohesion are the main building blocks. We argued that purpose and connectivity are the features of textuality that allow us to recognize a sequence of utterances as a text, as a collection of contextualized utterances of language use. We showed that these different aspects of textuality can be modelled theoretically if one adopts a theory of discourse that distinguishes between a global and a local level of structure. Grosz and Sidner's (1986) model, in particular its intentional structure component, enables us to provide a theoretical account of intentionality in discourse. Centering Theory allows us to model local cohesion in discourse. Modelling discourse structure, however, requires that we identify units that realize speakers' intentions and that stand in relations of coherence and cohesion with other units in the discourse. What those units of discourse structure may be is the topic we address next.

# CHAPTER 3: DISCOURSE UNITS

In the previous chapter, two levels of discourse structure were identified: global structure and local structure. Each level of discourse structure has its own minimal structural unit. At the global level, Grosz and Sidner (1986) identified the discourse segment as the structural unit. At the local level, the original definition proposed within Centering Theory (Grosz, et al., 1995; Walker, et al., 1998) establishes the utterance as the unit of local structure. In this chapter, we briefly discuss Grosz and Sidner's discourse segment. Following work within Systemic Functional Linguistics (SFL), we propose an alternative unit of global discourse structure, stages within a genre, which, as we will see, also captures two of the criteria for textuality we identified in our characterization of discourse: intentionality and coherence (in the sense of context-dependent interpretation). We then move on to explore the definition of utterance. The vagueness of the concept has led researchers to postulate different linguistic structures such as the sentence or the clause as the unit of analysis. Here we review two possible units of analysis that have been explored in the Centering literature, and then we propose to adopt SFL's system of TAXIS to standardize the different proposals to be tested in our study.

## 3.1 Global structure segmentation

### 3.1.1 The segment

We first mentioned the segment as the unit of global discourse structure in our discussion of Grosz and Sidner's (1986) model of discourse structure. In section 2.1, we saw that discourse segments comprise sequences of utterances, which are said to have specific functions in that segment in much the same way as each segment is said to have particular functions in the overall discourse. In our discussion of Example (3), we identified sequences of utterances belonging to two different segments. The utterances in lines 5 to 21 were grouped in a discourse segment that had the purpose of narrating the boarding pass mix-up. This discourse segment was embedded within the discourse segment that contained lines 1 to 4 and 22 to 32, which had the purpose of narrating the meeting with the relatives at the Chicago airport. Both of these segments, we said, contributed to another purpose, that of narrating what the trip to Taipei had been like.

We can say then, that the aggregation of utterances into a discourse segment is related to their intentional structure, that is, their purpose. Grosz and Sidner argued that discourses are naturally divided into segments, and reported on several studies that have been able to identify this segmental structure in a variety of texts, from task-oriented dialogues to Watergate transcripts, informal debates and therapeutic discourse (see Grosz & Sidner, 1986, p. 177). Although at the time their study was published, no psycholinguistic evidence for the existence of discourse segments was available, later studies have been able to provide this evidence and move one step forward, in the direction of automated discourse

segmentation. One such study is Passonneau and Litman (1997), in which the authors devised a method for identifying discourse segments. Passonneau and Litman's (1997) study focused on both human segmentation as well as automatic segmentation of texts. In the first stage of their study, they asked seven naïve subjects to segment 20 spoken narratives using the notion of communicative intention as the segmentation criterion (p. 109). Agreement on segmentation was found to be highly significant (p. 116). This can be then seen as support for Grosz and Sidner's (1986) claim regarding the natural aggregation of utterances into discourse segments.

### 3.1.2 Stages

There are, however, different ways in which we can segment discourse at the global level. One of those ways takes into account the type of communicative event – or genre. Within SFL, genre describes the context of culture and refers to both (a) the way in which "people use language to achieve culturally appropriate goals" (Eggins, 1994, p. 25) and (b) "the overall purpose or function of the interaction" (p. 26). We can see genre, then, as an activity that has a social function and a communicative function and that is expressed through language. How these activities take place and what type of language is used is contextually determined, by the context of culture and also by the context of situation, or register. In fact, Ventola (1987, p. 61) argued that the social acts realized through genre "are established and maintained within a society." So, for example, if we take the genre of university lectures in North America, we can say that it is realized in stages of instructor talk followed by student participation. It is our

culture that defines a university lecture as the sequence of these stages and it is our culture that maintains it through the preservation of this staging in all of its realizations. We can have variation in the context of situation, the register, when, for example, instead of professors as instructors, we allow PhD students to teach courses. But even then, the degree of variation in itself is determined by the genre and thus our culture: A child is not allowed to teach a course, so context variation is culture-dependent.

The way in which social acts unfold is structured, that is, it takes place in stages, and these stages are also culturally dependent. We can recognize the genre of a text by its stages, by the steps we go through as the discourse unfolds. For example, Hadic Zabala (2007), in her description of the genre of online personal ads, identified the following six stages: (a) Opening (salutation), (b) Self-Identification/Self-Description, (c) Action (Description of Desired Partner, Envisaged Relationship), (d) Note of Restriction, (e) Solicit Response and (f) Closing (salutation). Example (11), taken from the database of Hadic Zabala (2007), illustrates the stages of Self-Identification/Self Description, Action (Description of Desired Partner), Solicit Response and Closing (Salutation).

**11**

Fun, outgoing and looking!
Age: 22; Vancouver, BC
In my own words

**Self-Identification/Self-Description**

1 I've lived and gone to school in Vancouver,
2 but still think of myself as a small town girl at heart.
3 I've graduated from UBC
4 and now work full time
5 and really get to enjoy Vancouver.

6 In my spare time, I love to swim, explore the city, hang out with my friends and spend sunny evenings outside somewhere on a patio or by the water.
7 Now that the weather has become more summer like,
8 I'm looking forward to some camping trips, rollarblading and beach days!
9 But then, watching a good movie snuggled up on the couch is great too!
10 My friends and family are very important to me,

**Action: Description of Desired Partner**

11 and I am looking for someone with the same values.

**Solicit Response**

12 If your a fun, outgoing, like to have a good time, no nonsense guy,
13 I'd love to hear from you.

**Closing: Salutation**

14 Cheers,
15 -K

Stages, such as these illustrated in Example (11) are known as the schematic structure of a genre (Eggins, 1994, p. 36). Each stage is said to contribute to a part of the overall goal (p. 36). For instance, the Self-Identification/Self-Description stage has the function of listing those qualities possessed by the writer that are likely to attract a partner, whereas the Solicit Response stage has the function of motivating the reader to contact the writer. As we see, the individual contribution of a stage to the overall goal is given by the functional labels we attribute to each stage (Eggins, 1994, p. 37). Labelling can be done following two criteria: formal criteria or functional criteria.

> **Formal** criteria: we could divide the text into stages/parts according to the **form** of the different constituents. This approach emphasizes sameness, as we divide the text so that each unit/stage is a constituent of the same type.
>
> **Functional** criteria: we could divide the genre into stages/parts according to the **function** of the different constituents. This approach emphasizes difference, as we divide the text according to the different functions of each stage. (Eggins, 1994, p. 37)

Once the stages have been identified and labelled, it is possible to state the schematic structure of a text by listing its stages in the order in which they occur (p. 40). If this schematic structure is available from many texts of a text type, a genre, it is possible to generalize the schematic structure of the genre from that of the different texts, identifying which stages are defining and therefore obligatory and which are optional. The set of all obligatory and all optional stages constitute the generic structure potential of a genre (Hasan, 1985 cited in (Eggins, 1994). Returning to our previous example, Hadic Zabala (2007) proposed the following schematic structure for online personal ads:

**12**

(Opening)^[SI/SD,Action]$^{n}$ ^(Note of Restriction)^Solicit Response^(Closing)^(Solicit Response)[9]

The stages of Opening, Note of Restriction and Closing were found to be optional, whereas Self-Identification/Self-Description, Action and Solicit Response were found to be obligatory.

The identification of the stages within the genre and, as a consequence, of the genre itself is based on the purpose or goal of each stage and that of the overall text. In this way, the stage as a unit of global discourse embodies one of the three defining properties of discourse, the criterion of intentionality. The criterion of coherence, inasmuch as it refers to the contextual information needed for textual interpretation, is also accounted for by generic structure, given that both the context of culture and the context of situation determine any particular realization of a genre.

[9] Parentheses indicate optionality; the caret symbol indicates linear order; n superscript indicates that the stage (or sequence of stages) could be recursive.

In our analysis of RC function, we adopt a generic analysis of texts and we identify the stages in which RCs occur on the basis of functional criteria. Identifying the global structure of the texts in which RCs occur is important for our study, given the different levels of accessibility that we may find in entities across segment boundaries. Grosz and Sidner (1998) argue that a better understanding of the interaction between local and global structure is needed in order to assess the accessibility of centers across discourse segment boundaries, in particular, whether the forward-looking centers and the backward looking centers of the utterances that precede the new discourse segments are still accessible to the utterances that initiate the new discourse segments. As a consequence, we restrict our data to segment-internal RCs and thus exclude segment boundaries as a possible cause for disruptions in textual cohesion. In other words, following our discussion of stages, the RCs analyzed in this study are stage-internal.

Before moving on to the discussion of the utterance as the unit of local discourse structure, it is important to note that our choice of stages as the unit of global structure is not without problems. Eggins and Slade (1997, p. 269) noted that a significant portion of their casual conversation data could not be accounted for with a generic analysis. About 50% of casual conversation consisted of chat, which are said to be non-generically structured segments. This limitation applies only to our conversation data. If RCs are found in non-generically structured segments, following Eggins and Slade (1997), these segments will be labelled chat, when no other functional label can be assigned. Our coding of generic

structure is described in Chapter 5 (section 5.4.1.1). In the next section, we examine the different approaches to the segmentation of local structure.

## 3.2 Local structure segmentation

It has been previously indicated (e.g., Poesio, et al., 2004a; Poesio, et al., 2004b; Taboada & Hadic Zabala, 2008) that the definition of the notion of utterance is central to the application of the Centering algorithm, given that textual cohesion is measured from one utterance to the next. In fact, the definition of the local unit of analysis is relevant for most types of discourse annotation, and has consequently been the focus of much research. Among several possibilities, the sentence and the clause have been proposed for written texts, the turn and the intonation unit for spoken data. Given that one of the central goals of this study is to establish a systematic approach to text segmentation that could be used for both written and spoken data, both the turn and the intonation unit are excluded from the list of possible candidates. In recent research on segmentation in Centering Theory (Kameyama, 1998; Miltsakaki, 2002, 2003, 2005; Poesio, et al., 2000; Poesio, et al., 2004a, 2004b; Taboada & Hadic Zabala, 2008), the clause and the sentence have received the most attention. The next subsections review this previous research and incorporate findings from other researchers involved in discourse annotation.

### 3.2.1 Sentence-based Centering

Miltsakaki (2002, 2003, 2005) proposed the sentence as the unit of analysis at the local level of discourse. Miltsakaki's (2003) dissertation investigated the

relationship between attention management and discourse structure, focusing on the topicality of entities in subordinate clauses. She argued that entities in subordinate clauses are less topical and accessible than entities in main clauses, first, because they are not favoured as antecedents for subject pronouns in main clauses and second, because they are not likely to be mentioned in subsequent discourse in pronominal form. As a consequence, the sentence, consisting of the matrix clause and its subordinating clauses, should be the unit of analysis in Centering Theory[10].

The first distinction in topicality between entities in subordinate clauses vs. entities in matrix clauses concerns the likelihood of entities in subordinate clauses to be antecedents for subject pronouns in subsequent main clauses. Evidence for this distinction was obtained from a series of sentence completion experiments in Greek and English in which participants were requested to complete the second clause of either a main-main clause pair or a main-subordinate clause pair. The second clause in the clause pair contained only a connective (in the main-main pair) or a subordinator (in the main-subordinate pair) and a pronoun that had at least two possible antecedents in the first member of the clause pair. Examples are provided in (13) (from Miltsakaki, 2003, p. 74).

[10] Under the rubric of subordinate clauses, Miltsakaki collapses adverbial clauses and RCs. It is then important to note that Miltsakaki's use of the term subordinate subsumes both subordinate (i.e. hypotactic) clauses and embedded clauses, a distinction that we will return to in our discussion of RCs.

**13**

(a) The groom hit the best man. However, he…

(b) The beggar pushed the gentleman although he …

(c) The boxer kicked the referee. Then, he…

(d) The policeman shot the burglar when he…

The results of two experiments in English showed a strong effect for clause type: "when the second clause was subordinate, the subject pronoun showed a much weaker tendency to refer to the subject of the preceding main clause. Reference to the subject of the preceding main clause was strongly preferred when the subject pronoun appeared in a main clause" (p. 79)[11]. In other words, Miltsakaki found that the pronominal subject of a main clause tends to resolve to the subject of the preceding main clause. This preference was not as strong for subordinate clauses. These results are taken to show that structural focusing (i.e., structural salience) is responsible for pronominal interpretation in matrix clauses but not in subordinate clauses, and they are further interpreted as evidence against treating subordinate clauses as units of analysis.

It is important to note, however, that not all pronouns in main-main sequences are resolved through structural focusing. For instance, Miltsakaki noted that the connective *so* might be subject to semantic focusing instead (p. 194). In other words, the preference for pronouns to refer to the subject of the

[11] The Greek experiment was slightly different. It was a naturalness judgment questionnaire of main-main and main-subordinate pairs, which differed only in the realization of the subject of the second clause. The subject of the second clause co-referred with the object of the first clause and was either a strong pronoun or a weak pronoun. Miltsakaki's results showed that strong pronouns are preferred in the main-main condition, whereas weak pronouns are preferred in the main-subordinate condition (pp. 91 – 93).

preceding main clause may be overridden by the semantic properties of connectives.

Miltsakaki draws further support for this distinction in topicality between main clauses and subordinate clauses from similar findings observed by Cooreman and Sanford (1996) in their work on adverbial clauses. Cooreman and Sanford (1996) examined the effect of main and subordinate clauses linked by connectives such as *before*/*after*, *when*/*while* and *since*/*because* on discourse processing. Based on previous studies that have found main and subordinate clauses to be processed differently at the sentence level, they hypothesized a main clause effect on discourse processing, that is, main clauses are more accessible in subsequent discourse. Indeed, the results of a sentence continuation study showed a main clause effect: The main clause referent was preferred over the subordinate clause referent, albeit in different proportions for the different connectives. The results of a timed reading experiment that asked participants to read a target clause that was either a continuation of a main clause or of a subordinate clause were less homogenous. A main clause effect (i.e., faster reading times) was observed for the connectives *after* and *before*, an interaction between main clause effect and order for the connectives *when* and *while*, and no main clause effect or order effect for the connectives *since* and *because*. In sum, the results of Cooreman and Sanford's sentence completion study show that entities in main clauses are more accessible than entities in subordinate clauses. More importantly, the results of their reading experiment reveal that the semantics of the connectives are an important cue in discourse processing, so

that not only the type of clause, but also the type of connective and the type of task affect discourse processing.

The second distinction in topicality between entities in main clauses vs. entities in subordinate clauses concerns the likelihood of entities in subordinate clauses to be mentioned in subsequent discourse. In order to assess this likelihood, Miltsakaki (2003) performed a reference test on 300 English RCs and 200 Greek RCs. The reference test consisted of establishing whether or not the entity denoted by the head noun and other entities in the RC were subsequently mentioned in the discourse, and if so, with which type of referring expression (p. 107).

The results for the English relatives showed that the likelihood of subsequent mention is related to the type of relative clause (type of relativizer): Subsequent mention of the entity denoted by the head noun was highest for *who*-relatives (47%) whereas subsequent mention of other entities in the RC was highest for *which*-relatives (39%). The distribution of the subsequent mentions by type of referring expressions resulted in numbers too low to be conclusive (see Tables 5.2, 5.3 and 5.5, on pp. 114, 116, and 124, for details). The functional properties of RCs (i.e., restrictiveness) were also found to affect subsequent mention[12]: Entities functioning as head noun in non-restrictive RCs were more

[12] Miltsakaki (2003, p. 110) adopts the traditional distinction between restrictive and non-restrictive RCs: "Restricting relative clauses are necessary to identify the referent of the head noun. … Non-restricting relative clauses provide additional information about the referent of the head noun." Miltsakaki follows McCawley's (1981) criterion to distinguish between restrictive and non-restrictive RCs: if the RC is restrictive, the antecedent of *one* in (a) includes the RC; if the RC is non-restrictive, the antecedent of *one* in (b) does not include the RC (pp. 109 – 10):

a. Tom has two cats that once belonged to Fred, and Sam has one.

likely to be mentioned in subsequent discourse than entities functioning as head noun in restrictive RCs (Miltsakaki, 2003, p. 126). Based on these observations, Miltsakaki tentatively concluded that the topical status of entities in RCs is lower than the topical status of entities in main clauses, given that "the results of corpus studies showed that entities evoked in relative clauses are not subsequently referenced with a pronoun unless a) they are already pronominalized in the relative clause or b) the highest ranked entity in the main clause is also pronominalized" (p. 158).

Having established a distinction in the topicality of entities in subordinate clauses vs. entities in main clauses, Miltsakaki (2003, 2005) investigated whether non-restrictive relative clauses (NRRCs) should be treated as separate utterances in Centering Theory, that is, as topic update units. The analysis consisted of subjecting 200 NRRCs (100 English NRRCs taken from the Wall Street Journal corpus and 100 Greek NRRCs taken from online newspaper articles) to two different conditions, the complex sentence condition and the single clause condition, and evaluating their performance with the Centering algorithm. In the complex sentence condition, the NRRC belonged to the same topic update unit as the main clause; in the single clause condition, the NRRC constituted an independent unit from the main clause.

Miltsakaki computed the transitions to the unit following the NRRC and observed that in English the single clause condition yielded more cohesive Centering transitions in 13 utterances (13%) and less cohesive Centering

---

b. Tom has two violins, which once belonged to Heifetz, and Sam has one.

transitions in 46 utterances (46%) (Miltsakaki, 2003, p. 130). More cohesive means that the transition obtained was a more preferred transition in Rule 2, which ranks transitions from CONTINUE to ROUGH SHIFT in terms of the inference load placed upon the reader. For the remaining 41 utterances, there was no effect. Similar findings were obtained for Greek (p. 150), where the single clause condition resulted in more cohesive Centering transitions in 8 utterances (8%) and in less cohesive transitions in 44 utterances (44%). Miltsakaki interpreted these findings as support for the complex sentence hypothesis, namely, that NRRCs do not constitute topic update units on their own, but belong to the topic update unit of their main clause (p. 131).

In sum, the findings of Miltsakaki (2003, 2005) and Cooreman and Sanford (1996) discussed in this section show a difference in degrees of topicality between entities in subordinate clauses and entities in matrix clauses. Entities in subordinate clauses are not as likely to be antecedents for pronominal subjects or to be mentioned in subsequent discourse as entities in main clauses are. But the results of Miltsakaki and Cooreman and Sanford also revealed that properties of the discourse, the type of clause and the semantics of connectors may revert the effect of structural focusing. In this study, while we acknowledge that mention in a RC does not grant an entity topic status, we do not exclude the possibility that mention in a RC serves topic continuity, in which case, RCs would contribute to topic management.

It is important to note, however, that Miltsakaki (2002, 2003, 2005) were not the only studies that have found the Centering algorithm to better model

textual cohesion when the sentence is the unit of analysis. Poesio et al. (2000) and Poesio et al. (2004a, 2004b) also attempted to specify some of the underspecified notions in Centering Theory, including the notion of utterance. These studies tested different realizations of utterance, including the sentence, the finite clause and the non-finite clause and evaluated their performance with respect to Rule 1 and Constraint 1 of Centering. Recall that Rule 1 states that if any entity of the previous utterance is realized in the current utterance as a pronoun, then the Cb is realized as a pronoun. Constraint 1 states that each utterance should have precisely one Cb. Poesio et al. (2000) and Poesio et al. (2004a, 2004b) found that sentences performed better than finite clauses with respect to Constraint 1. Adopting the sentence as the unit of analysis, however, led to more violations of Rule 1. Although this could be seen as support for sentence-based theory, Poesio et al. (2004a, p. 31) are reluctant to make such a claim:

> Even though identifying utterances with sentences leads to much better results for Strong C1, we will not simply abandon the hypothesis that utterances may coincide with finite clauses. This is in part for theoretical reasons, such as the fact that in other theories of discourse where 'units' are assumed, such as RST, these units are generally finite clauses.

While compatibility with other theoretical frameworks is a strong motivation for rejecting the sentence as the unit of analysis, there are also theory-internal complications with sentence-based Centering. In particular, if subordinate clauses are processed with their matrix clause, anaphoric references in the subordinate clause cannot be resolved with the Centering algorithm and must be resolved with a different mechanism. A clause-based approach would

overcome this issue and would provide a more economical approach to the resolution of pronominal reference.

### 3.2.2 Clause-based Centering

Arguing against the sentence as the unit of analysis in Centering we find Kameyama (1998). Kameyama extended the Centering algorithm to account for intrasentential pronouns, that is, pronouns that occur in the same sentence as their antecedent. In order to account for intrasentential anaphora, Kameyama (1998) favoured a clause-based approach to Centering, rather than a sentential approach (pp. 97, 103). There are computational and linguistic reasons militating against a sentence-based approach. From a computational point of view, Kameyama argued that the computational load needed to process sentences that consist of multiple clauses becomes manageable when these sentences are segmented into clauses and processed one at a time. From a linguistic point of view, she argued that a clause-based segmentation of utterances in Centering would best reflect the grammaticization of parallel and monadic tendencies that are observed in complex sentences (p. 99). Parallel tendency refers to a preference for structural parallelism: "Two adjacent utterances in discourse seek maximal parallelism" (p. 96)[13]. Monadic tendency refers to the tendency of discourse "to be about one thing at a time" (p. 90). Structural parallelism is said

[13] This parallelism preference is said to interact with attentional preference: When attentional preference is determinate (i.e. when there is only one maximally salient entity), attentional preference overrides parallelism preference; when attentional preference is indeterminate (i.e., the set of maximally salient entities is larger than one), there is a weak preference for parallelism (p. 96). Kameyama notes that attentional indeterminacy has not received proper study in Centering research (p. 95) and proposes a reformulation of the Centering algorithm that incorporates attentional determinacy and structural parallelism (pp. 92 -6). See Kameyama (1998) for details.

to be grammaticized in ellipsis phenomena, whereas monadicity is said to be found in control phenomena, topicalization, left- and right-dislocation and clefting (pp. 99 – 100).

Her clause-based approach to Centering segmentation is captured in her Intrasentential Centering Hypothesis (ICH), provided in (14) below (from Kameyama, 1998, p. 100).

**14**

Intrasentential Centering Hypothesis (ICH): A complex sentence is broken up into a set of center-updating units corresponding to the 'utterances' in intersentential Centering.

Kameyama (1998) further distinguished between sequential intrasentential Centering and hierarchical intrasentential Centering (p. 101). Sequential intrasentential Centering results in a flat sequence, that is, "there is always a single centering state, and the output of a complex sentence is the output of the last subsentential unit" (p. 101). Hierarchical intrasentential Centering results in a tree structure. This means that there will be "multiple [C]entering states simultaneously active at different depths of embedding" (p. 101).

Kameyama's clause-based intrasentential Centering makes a distinction between the types of clauses that require sequential segmentation and those that require hierarchical segmentation (pp. 103 – 9). Within the types of clauses that require hierarchical segmentation, Kameyama further distinguished between clauses whose entities are accessible to the superordinate utterance and clauses whose entities are inaccessible (p. 108). Tensed conjuncts and tensed adjuncts

constitute sequential Centering structures (pp.104 - 5). Tenseless conjuncts and tenseless adjuncts do not constitute sequential Centering structures; they constitute one unit with their superordinate clause (p. 105). Reported speech, tensed nonreport complements and RCs require hierarchical segmentation and constitute embedded Centering segments (pp. 107 – 8). They differ in the accessibility of their centers for subsequent discourse: Whereas the entities in tensed nonreport complements and RCs are accessible to higher-level Centering (p. 108), the entities in reported speech utterances are not (p. 107). Tenseless nonreport complements do not constitute embedded Centering units and are processed with their superordinate clause (p. 108). Examples are provided in Table 5 below (from Kameyama, 1998, pp. 103-9).

Support for the distinction between hierarchical and sequential processing comes from Suri and McCoy's (1994) methodology for complex sentence segmentation. Suri and McCoy (1994) observed that native-speakers of English tend to resolve pronominal subjects of matrix clauses with subjects of matrix clauses, and the pronominal subject of subordinate clauses with the subject of the superordinate matrix clause. In other words, in sentences of the type *SX because SY*, the subject of *SX* tends to be resolved with the subject to the previous sentence; the subject of *SY* tends to be resolved with the subject of *SX*; and the subject of the following utterance tends to be resolved with the subject of *SX*. For the subject of the following utterance, then, the entities in the subordinate clause appear not to be salient enough. These native speaker interpretations were later supported with a corpus study of 81 text sequences containing *SX because SY*

sentences (extracted from the Brown corpus and from works of 20th century literature) (Suri, McCoy, & DeCristofaro, 1999).

**Table 5. Clause-based Intrasentential Centering**

| | |
|---|---|
| **Sequential Intrasentential Centering**[a] | |
| Tensed conjunct | $U_1$ Her mother was a Greer<br>$U_2$ and her father's family came from the Orkney Isles. (p. 104) |
| Tensed adjunct | $U_1$ Although she's still a teenager who looks like a baby,<br>$U_2$ she is getting married. (p. 105) |
| Tenseless conjunct | $U_1$ I wanted [to grab her by the arm and beg her [to wait, to consider, to know for certain]] (p. 105) |
| Tenseless adjunct | $U_1$ [In the fullness of her vocal splendor], however, she could sing the famous scene magnificently. (p. 105) |
| **Hierarchical Intrasentential Centering** | |
| Reported speech complement | $U_1$ Sunday he added,<br>(a) 'We can love Eisenhower the man<br>(b) even if we consider him a mediocre president<br>(c) but there is nothing left of the Republican Party without his leadership.'<br>$U_2$ Mitschell said<br>(a) the statement should become a major issue in the primary and the fall campaign. (pp. 107 – 8) |
| Tensed nonreport complement | $U_1$ Her choice of one color means<br>(a) she is simply enjoying the motor act of coloring without having reached the point of selecting suitable colors for different objects. (p. 108) |
| Tenseless nonreport complement | $U_1$ We watched them [set out up the hill hand in had on a rainy day in their yellow raincoats [∅ to finger paint at the grammar school]]. (p. 108) |

[a] $U_1$ or $U_2$ signal center-updating units, or utterances. (a), (b) etc., signal embedded Centering units. Clauses in square brackets do not constitute center-updating units, nor embedded Centering units.

In a recent study, Taboada and Hadic Zabala (2008) tested four different approaches to discourse segmentation, including Miltsakaki's sentence-based

approach, Kameyama's hierarchical clause-based approach, and a clause-based approach based on Poesio et al.'s (2000) work. The methods were evaluated with respect to: (a) the number of empty backward-looking centers (Cb); (b) the proportion of topic – Cb agreements; (c) the proportion of CHEAP vs. EXPENSIVE transitions; and (d) the proportion of antecedents found in the same utterance as the anaphors.

Two of these evaluation methods are theory-internal and were introduced in our discussion on Centering in Section 2.2. The number of empty backward-looking centers refers to Constraint 1 of Centering, which, in its strong version, requires each utterance to have one Cb. The proportion of CHEAP and EXPENSIVE transitions is a reformulation of Rule 2 of Centering, proposed by Kibble (2001) based on Strube and Hahn (1999). According to this reformulation, CHEAP transitions, that is, transitions in which the Cp of the current utterance is predicted to be the Cb of the next utterance, are preferred. The other two characteristics are theory-external: Characteristic (b) evaluates the topichood of the Cb by checking the number of times topic and Cb coincide; characteristic (d) evaluates the applicability of Centering to anaphora resolution, given the different approaches to segmentation.

As with Poesio et al. (2004a, 2004b), the results of Taboada and Hadic Zabala (2008) were non-conclusive. While sentence-based Centering was found to perform better with respect to characteristics (a) and (b) in one of the two languages examined (Spanish), it had the worst performance with respect to characteristic (d). In the end, Taboada and Hadic Zabala (2008) favoured a

clause-based approach to discourse segmentation, in part, because it would allow to combine Centering-based with RST-based analyses (p. 45).

#### 3.2.2.1 Support for the clause

While recent corpus studies have not found strong support for the clause as the unit of discourse analysis, there is some support from two other sources. First, several researchers have reported a tendency for intonation units, that is, spurts of spoken language that have single intonation contour with a clause-final or sentence-final cadence (Chafe, 1988, p. 1), to coincide with a single clause. Chafe (1992, p. 91) observed that studies of spoken corpora clearly point to the intonation unit as the unit of discourse. Moreover, most intonation units were found to have the grammatical form of a clause, which is defined as consisting of a subject and a predicate and expressing the idea of an event or state (p. 91). Chafe (1988) recognized that this correlation between intonation units and clauses is not perfect: Only about 70% of all intonation units were expressed as single clause (p. 3)[14]. Matsumoto (2003, p. 26) referred to this correlation between IUs and clauses as the clause centrality proposal; the clause is seen as "the basic unit for information processing and segmentation in human spoken discourse" (p. 27)[15].

In a recent study, Thompson and Couper-Kuhlen (2005, p. 484), identified the clause, not just as a unit of interaction, but as the locus of

[14] The remaining 30% of intonation units may be expressed in units smaller than a clause, such as prepositional phrases, or in units consisting of multiple clauses, such as combinations of main and adverbial clauses, main and relative clauses and main and verb-complement clauses.

[15] It is interesting to note, however, that the same centrality of the clause was not observed in Japanese, which according to Matsumoto (2003, p. 9, and references therein), "shows a preference for non-clausal, or phrasal IUs that lack verbal predicates."

interaction, meaning that the clause is the format oriented to by speakers when projecting the actions of others and in reacting to these projections. For English, Thompson and Couper-Kuhlen (2005, pp. 489-497) argued that the status of the clause[16] as the locus of interaction is revealed through speakers' orientation to the clause in three practices: next-turn onset, joint utterance completion and turn unit extension. Next speakers tend to start their turn when the previous speaker has completed a clause and not before clause completion (p. 489). Speakers tend to complete each other's turns and these jointly constructed utterances take the form of clauses. Additions are either the last words in a one-clause unit, or another clause in a multi-clausal unit (p. 492). Finally, Thompson and Couper-Kuhlen (2005, p. 495) argued that turn unit extensions orient to the clause as the unit of interaction, meaning that when speakers extend their turns, they do so by adding syntactically dependent material (in the form of recurrent phrases in the language) to a clause that constitutes their current turn (p. 495). Based on this evidence, they concluded that "clauses are interactionally warranted units" (p. 497).

The second source of support for the clause as the unit of discourse analysis comes from studies of language processing. Roberts and Gibson (2002) empirically identified the unit of sentence processing by testing whether sentence memory is a function of clauses or a function of the number of words or new referents. Participants heard sentences consisting of two, three, four and five clauses and were asked to answer a question testing their memory of the subject of the clause or of the main verb of the clause (p. 584). Sentences contained two

[16] Clause was understood as predicate and accompanying phrases.

to five clauses of one of these two types: restrictive RCs and sentential complements. Examples are provided in (15a,b,c), below (from Roberts & Gibson, 2002, p. 585, examples (6) to (8)).

**15**

**(a) Relative Clause (RC)**

The barber lectured the sailor who hit the singer who worked in the jazz club.

**(b) Sentential Complement (SC)**

The violinist insisted that the immigrant doubted that the chef had trained in Paris.

**(c) Double Object (DO)**[17]

The psychologist showed the document to the criminal who sent a gift to the editor who was compiling an anthology.

Roberts and Gibson (2002, p. 587) found that the number of clauses in a sentence had an effect on accuracy in all three sentence types. On the basis of these findings they concluded:

> The results of the experiment suggest that participants recalled the content of the sentences as function of the recency of presentation of the number of clauses in a sentence, and not the recency of presentation of the number of NPs or discourse referents in a sentence. Thus although new discourse structure appears to be an important measure of locality in on-line sentence comprehension (Gibson, 1998), the current results suggest that the clause is a more important storage unit for sentence memory. This finding thus confirms the hypothesis from the literature that the unit of sentence memory is the clause (…). (Roberts & Gibson, 2002, p. 593)

[17] Note that the Double Object sentence type is really a variant of the RC type. In the RC type, the first RC modifies the Direct Object of the main clause, the second RC modifies the Direct Object of the first RC. In the Double Object type, the first RC modifies the Indirect Object of the main clause, the second RC modifies the Indirect Object of the first RC.

Roberts and Gibson's identification of the clause as the unit of sentence processing is problematic in that both types of clauses used in the sentences are embedded clauses in the Hallidayan sense, that is, they are not clauses that are in relation to other clauses, but clauses that are part of (i.e. build) other clauses. It would seem necessary then to include sentences with all three types of clauses (paratactic, hypotactic and embedded clauses, which will be explained in section 3.2.3) in the analysis in order to make this claim. On the other hand, it could be argued that if embedded clauses were empirically found to be the unit of sentence memory, all other clause types, inasmuch as they share the basic predicate-argument structure of embedded clauses, are also units of sentence memory. This limitation notwithstanding, the argument could still be made for single-clause or single-predicate units of sentence memory and against complex units.

Within the field of language processing, more indirect evidence for a clause-based approach may be found. In an early study of RC processing in English, Sheldon (1977) posited that interruptions of canonical structure such as those found in centre-embedded or subject head RCs would be indicative of clause-based sentence processing:

> Notice that the assumption that sentences with subject modifiers contain an interruption rests on a particular interpretation in which the head of the relative clause is a constituent of the main clause. Actually, the whole complex NP is the subject of the sentence. If we take this view, there is no interruption of the subject and predicate. And facts about the structure of the subject relatives are irrelevant to Slobin's principle. The anti-interruption principle, therefore, must be part of a model of sentence processing which assumes the integrity of clauses. In such a model, the clause is a unit that sentences are segmented into. (Sheldon, 1977, p. 307).

Gibson (1998, p. 3 and references therein) argued that the complexity of centre-embedded RCs is such that multiple embeddings render them uninterpretable. There are, however, factors that decrease the difficulty associated with centre-embedded RCs, namely, age of the speaker, properties of the noun phrases and properties of the language. Kidd and Bavin (2002) investigated the processing of centre-embedded and right-branching RCs by English-speaking children aged 3, 4 and 5. The children were asked to act out sentences describing the activities of ill-behaved animal toys that were manipulated for embedding (centre-embedded or right-branching) and focus (subject gap or object gap). They found that centre-embedded RCs were more difficult to process and that the difficulty in processing centre-embedded RCs decreased with age (p. 608). Warren and Gibson (2002) examined the role of referential processing on sentence processing and tested whether the difficulty in processing centre-embedded RCs is affected by the type of NP found in the RC and by the accessibility of this NP in the discourse (following Gundel, Hedberg and Zacharski's (1993) Givenness Hierarchy). Their results showed (a) that centre-embedded RCs with pronouns are easier to process than centre-embedded RCs with lexical NPs, and (b) that the processing of centre-embedded RCs is facilitated when the accessibility status of the RC NP is found in the more given end of the hierarchy. Hoover (1992) showed that differences in processing strategies between Spanish and English speakers account for the difference in complexity between multiply-embedded Spanish RCs and multiply-embedded English RCs. The results of two self-paced reading experiments showed that Spanish speakers were able to comprehend sentences with double centre-

embedding (e.g., *El carnicero que el camarero que el boxeador ayudó mató avisó al gitano*. p. 292) whereas English speakers were not (*The butcher that the waiter that the boxer helped killed warned the Gypsy*. p. 292). English speakers were said to wait till the end of the sentence to assign semantic roles, whereas Spanish speakers were said to assign semantic roles on-line. These mitigating factors notwithstanding, the complexity of centre-embedded RCs is well documented, and in view of Sheldon's argument, a source of support for clause-based sentence processing.

A final and also indirect source of evidence for clause-based processing comes from psycholinguistic studies of sentence ambiguity. Studies on sentences temporarily ambiguous between a main clause and a RC reading have found that in certain contexts (and counter-expectation given the salience of main clauses), the RC reading is preferred. Phillips and Gibson (1997) examined processing preferences in sentences that are ambiguous between a RC and a matrix clause reading, as in Example (16) (p. 326). Given the fragment, *Because Rose praised the recipe I made*, two continuations are possible: (16a) involves a RC reading; (16b) involves a matrix clause reading.

**16**

(a) *Because Rose praised the recipe I made* for her birthday I also made it for her graduation.
(b) *Because Rose praised the recipe I made* it for her birthday.

The results of a self-paced reading experiment showed a preference for the RC reading, when the subordinate clause was a non-temporal clause (e.g., *because, since, although*) and a preference for the matrix clause reading, when

the subordinate clause was a temporal clause (e.g., *when, while*, etc.). This split between temporal and non-temporal clauses was not expected and was tentatively attributed to the effects of a tense-matching constraint or to the presence of pronouns vs. full NPs. Even though more research is needed on this issue, these results motivate an analysis of sentence processing that takes into consideration properties of the discursive context, such as the function of dependent clauses in relation to their dominant clause.

Similarly, Sedivy's (2002) study on temporarily ambiguous sentences highlighted the role of discourse context in sentence processing, which can be manipulated to elicit a RC reading over a main clause reading. In her study, Sedivy (2002) examined the role of discourse context in the processing of sentences that are temporarily ambiguous between a main clause and a reduced RC reading. She manipulated the discourse context and the focus operator so that sentences were presented with or without an explicit contrastive set in the preceding context and with or without the focus operator *only*. The results showed an effect of the focus operator: The presence of the focus operator resulted in shorter reading times for the reduced RC condition. In addition, there was an effect for discourse context: Sentences in the main clause condition were read faster when the context had an explicit contrast set than when it did not; the inverse was true for reduced RCs. And finally, an interaction between discourse context and focus operator was observed: In the absence of the focus operator *only*, sentences in the main clause condition were read faster than sentences in the reduced RC condition. Beyond showing that the effects of the focus operator *only* in sentence processing are context-dependent, these results are indicative of

the importance of discourse information in sentence processing to the extent that it can override the preference for a major clause reading in favour of a minor clause reading.

In sum, there is considerable research pointing to the clause as the local unit of analysis. We found direct support in (a) studies of spoken discourse that have found intonation units to coincide with single clauses; (b) studies in Conversation Analysis that have found the clause as the locus of interaction; and (c) psycholinguistics studies that have argued for the clause as the unit of sentence memory. Indirect evidence was provided by (a) studies in language acquisition and language processing that have found centre-embedded structures, which disrupt canonical clause structure, to be more difficult to process and (b) studies in sentence ambiguity that have shown preferential readings for relative-clause-interpretations over main-clause-interpretations. In the next section, we provide a clause taxonomy that would allow us to test the adequacy of the clause (or alternatively of the clause-complex) as the unit of analysis at the local level of discourse structure.

### 3.2.3 A clause taxonomy

With the ultimate goal of testing the segmentation approaches reviewed in sections 3.2.1 and 3.2.2, we introduce here the clause taxonomy we adopt for this study, which is the clause taxonomy proposed by Systemic Functional Linguistics (e.g., Halliday & Matthiessen, 2004). With this classification of clauses we seek (a) to standardize the different approaches so as to facilitate comparison and (b)

to ensure that all possible approaches to local segmentation are included in the analysis.

The clause plays a major role in SFL: Halliday and Matthiessen (2004, p. 10) see it as "the central processing unit in the lexicogrammar – in the specific sense that it is in the clause that meanings of different kinds are mapped into an integrated grammatical structure." By meanings of different kinds, they mean the ideational (experiential)[18], interpersonal and textual metafunctions introduced in Chapter 1 (section 1.5). The ideational metafunction is expressed in the way in which we linguistically realize processes, participants and circumstances (i.e., verb types and argument structure), which allows us to create a version of reality, a model of our experience. The interpersonal metafunction is expressed in the way in which we linguistically realize giving and demanding (i.e., mood and modality), which allows us to enact our social relationships. And the textual metafunction is expressed in the way in which we organize our message (i.e., theme and rheme), which allows us to represent our model of experience and how we enact our social relationships.

These three types of meaning are mapped onto the clause: The clause is a representation of human experience; the clause is an exchange between speaker and listener; and the clause is a message (Halliday & Matthiessen, 2004, pp. 58-9). The metafunctions along with their corresponding statuses in the clause are summarized in Table 6 (from Halliday & Matthiessen, 2004, p. 61).

[18] The experiential metafunction concerns the expression of ideational meaning within the clause, whereas the logical metafunction concerns the expression of ideational meaning above the clause in the clause complex.

**Table 6. Metafunctions of language**

| Metafunction | Definition (kind of meaning) | Corresponding status in clause |
|---|---|---|
| experiential | construing a model of experience | clause as representation |
| interpersonal | enacting social relationships | clause as exchange |
| textual | creating relevance to context | clause as message |
| logical | constructing logical relations | -- |

The fourth metafunction listed in Table 6, the logical metafunction, has no corresponding status in the clause: It is embodied in the clause complex, that is, clauses linked together by a logico-semantic relation (Halliday & Matthiessen, 2004, p. 361). How clauses are linked together in discourse (and therefore how they can be segmented in text analysis) is what concerns us here and will thus be the focus of our analysis. A full discussion of the different metafunctions and their corresponding statuses in the clause can be found in Halliday and Matthiessen (2004).

Halliday and Matthiessen (2004, p. 373) identify two basic systems (network of choices, paradigmatic oppositions according to (Martin, 1992)) that determine how clauses are linked together to form a clause complex: TAXIS (or the degree of interdependency) and the LOGICO-SEMANTIC RELATION. In explaining TAXIS, Halliday and Matthiessen (2004, p. 373) tell us that "all clauses linked by a logico-semantic relation are interdependent: that is the meaning of relational structure – one unit is interdependent on another unit." They distinguish between two degrees of interdependency: "**Hypotaxis** is the relation between a dependent element and its dominant, the element on which it is dependent.

Contrasting with this is **parataxis**, which is the relation between two like elements of equal status, one initiating and the other continuing" (pp. 374 – 5).

A pair of clauses related by interdependency is called a clause nexus (p. 375). Each clause nexus consists of a primary clause and a secondary clause, as displayed in Table 7 (from Halliday & Matthiessen, 2004, p. 376).

**Table 7. Types of clauses in a clause nexus**

| | **primary** | **secondary** |
|---|---|---|
| **parataxis** | 1 (initiating) | 2 (continuing) |
| **hypotaxis** | α (dominant) | β (dependent) |

Note. Greek notation is used to represent hypotactic structures, whereas numerals are used to represent paratactic structures (Halliday & Matthiessen, 2004, p. 375).

The types of logico-semantic relations that can hold between a primary and a secondary clause are classified in two groups: expansion and projection. Halliday and Matthiessen identify the subtypes of expansion in (a, b, c) and projection (d, e) provided in (17):

**17**

| | |
|---|---|
| (a) Elaborating: 'i.e., for example, viz.' | one clause expands another by elaborating on it (or some portion of it): restating in other words, specifying in greater detail, commenting or exemplifying. |
| (b) Extending: 'and, or' | one clause expands another by extending beyond it: adding some new element, giving an exception to it, or offering an alternative. |
| (c) Enhancing: 'so, yet, then' | one clause expands another by embellishing around it: qualifying it with some circumstantial feature of time, place, cause or condition. |
| (d) Locution: 'says' | one clause is projected through another, which presents it as a locution, a construction of wording. |
| (e) Idea: 'thinks' | one clause is projected through another, which presents it as an idea, a construction of meaning. (Halliday & Matthiessen, 2004, p. 378) |

The intersection of the systems of TAXIS and LOGICO-SEMANTIC RELATION results in the basic set of clause nexuses, as shown in Table 8 (adapted from Halliday & Matthiessen, 2004, p. 380).

**Table 8. TAXIS and LOGICO-SEMANTIC TYPE in clause combining**

| | **paratactic** | **hypotactic** |
|---|---|---|
| (a) Elaboration | 1 John didn't wait;<br>=2 he ran away. | α John ran away,<br>=β which surprised everyone. |
| (b) Extension | 1 John ran away,<br>+2 and Fred stayed behind. | α John ran away,<br>+β whereas Fred stayed behind. |
| (c) Enhancement | 1 John was scared<br>x2 so he ran away. | α John ran away,<br>xβ because he was scared. |
| (d) Locution | 1 John said:<br>"2 "I'm running away" | α John said<br>"β he was running away. |
| (e) Idea | 1 John thought to himself:<br>'2 'I'll run away' | α John thought<br>'β he would run away. |

Note. 1 identifies an initiating clause and 2 a continuing clause in a paratactic relation. α signals a dominant clause and β a dependent clause in a hypotactic relation. = signals a relation of elaboration, + one of extension, x one of enhancement, " one of locution and ' one of idea.

The system of LOGICO-SEMANTIC RELATION is concerned with the type of connectivity found between one clause and another clause, the coherence

relations identified in our definition of discourse. The main focus of our study is on cohesion, or co-text dependent nominal interpretation. For this reason, we will not pursue an analysis of the LOGICO-SEMANTIC relations in our texts. Our focus is on the interdependency between clauses and their consequences for textual cohesion.

In relation to this interdependency between clauses, Halliday and Matthiessen (2004, p. 384) identify three differences between paratactic and hypotactic relations. First of all, since the clauses in a paratactic relation are of equal status, the paratactic relation is in principle logically symmetrical and transitive[19] (although this symmetry can be modified by the type of logico-semantic relation) (p. 384). The relationship between clauses in a hypotactic relation, on the other hand, is neither symmetrical nor transitive[20]. A second difference between clauses in a paratactic vs. a hypotactic relation concerns their finiteness. While clauses in a paratactic relation are always finite, clauses in a hypotactic relation can be either finite or non-finite (p. 386). The final difference is found in their sequencing. According to Halliday and Matthiessen (2004, p. 387), the ordering of clauses in a paratactic relation is that which is represented by the sequence, whereas the ordering of clauses in a hypotactic relation is independent of the sequence and relies on dependence: The dependent clause

[19] "(i) 'salt and pepper' implies 'pepper and salt', so the relationship is symmetrical; (ii) 'salt and pepper', 'pepper and mustard' together imply 'salt and mustard', so the relationship is transitive." (Halliday & Matthiessen, 2004, p. 384)

[20] "(i) 'I breathe when I sleep' does not imply 'I sleep when I breathe'; (ii) 'I fret when I have to drive slowly' and 'I have to drive slowly when it's been raining' together do not imply 'I fret when it's been raining'." (Halliday & Matthiessen, 2004, p. 384)

may follow the dominant clause, precede the dominant clause, be enclosed in the dominant clause, or enclose the dominant clause.

Having introduced SFL's system of TAXIS, it is now possible to reformulate sentence-based and clause-based Centering in terms of Halliday and Matthiessen's (2004) clause taxonomy. Our goal is to identify the unit of analysis at the local level of discourse structure. We call this a center-updating unit, as this unit updates the attentional state of the discourse participants, which keeps track of the entities that are mentioned in the discourse. It is important to note that the relations of parataxis and hypotaxis apply to clauses that combine with other clauses to form clause complexes. Independent clauses (i.e., independent monoclausal units) will continue to be considered center-updating units in our analysis. In (**18**), we provide the first draft of the segmentation approaches to be examined in the study. The list of segmentation approaches is to be revised with the addition of RCs in Chapter 4 and Kameyama's sequential/hierarchical distinction in Chapter 5.

**18**

**Paratactic-Clause Centering**

A center-updating unit is an independent clause or a clause that is in a paratactic relation with another clause. Clauses that are in a hypotactic relation with another clause belong to the center-updating unit of their dominant clause.

**Hypotactic-Clause Centering**

A center-updating unit is an independent clause or a clause that is either in a paratactic relation or a hypotactic relation with another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause.

Even though Halliday and Matthiessen (2004, p. 386) see non-finite clauses as clauses that can be in hypotactic relations (as well embedded clauses in nominal and verbal groups, p. 427), it has been our decision to exclude non-finite clauses as possible units of analysis. This decision was based on properties of non-finite clauses concerning (a) the expression of interpersonal meaning, and (b) subject recoverability.

According to SFL, the finite element of the verbal group is the element that fixes the process described in the verb in relation to the speech event, that is, it relates the process to the 'here and now' (Halliday & Matthiessen, 2004, p. 336). Non-finite verbal groups lack this temporal deictic component (p. 344), and this temporal deictic element, along with modality, is largely responsible for the realization of the interpersonal metafunction, for expressing our feelings and attitudes towards the process and towards our conversational partners. In order to account for this difference between finite and non-finite clauses in

interpersonal meaning in our analysis, the finite/non-finite distinction would have to be incorporated into our segmentation approaches by subdividing the hypotactic approach (as well as the embedded approach, which will be introduced in Chapter 4) into hypotactic finite and hypotactic non-finite, thereby increasing the number of segmentation approaches.

A second difference between finite and non-finite clauses concerns the subject of the clause. The subject of non-finite clauses has to be recovered, and while they tend to co-refer with the subject of the dominant or matrix clause, this may not always be the case, as illustrated by Halliday and Matthiessen (2004, p. 387). The inclusion of non-finite clauses in the analysis would have required a distinction between non-finite clauses with subjects that co-refer with the subject of the main clause vs. non-finite clauses with subjects that co-refer other entities, leading again to an increase in the number of segmentation approaches.

The necessary additions to the segmentation approaches required to include non-finite clauses as candidates for units of analysis in Centering did not seem justified given the findings of Poesio et al. (2004a, 2004b). In particular, Poesio et al. (2004b) found that treating all verbed clauses as units still resulted in a high number of violations of Constraint 1 of Centering, that is, that each utterance must have one Cb. The total number of utterances the proportion of violations was significantly lower than the one observed for finite clauses (47% vs. 49%) (p. 335), but it is important to note that subjects of non-finite clauses were assumed to co-refer with the subject of the main clause (p. 335). In any case,

violations of Constraint 1 of Centering were still higher than the ones obtained when the sentence was taken as the unit of analysis (36.8%) (p. 338).

Following our reformulation of units of analysis in terms of SFL's sytem of TAXIS, Miltsakaki's sentence-based Centering now corresponds to Paratactic-Clause Centering, according to which all clauses that are in a hypotactic (subordinating) relation to a dominant clause belong in the same Centering utterance as said dominant clause. Kameyama's clause-based Centering is in part captured by our Hypotactic-Clause Centering approach, according to which finite hypotactic clauses constitute their own Centering utterances[21]. In this way, the introduction of the system of TAXIS allows us to standardize and systematize our approaches to segmentation.

## 3.3 Summary

In this chapter, we have reviewed different approaches to the segmentation of discourse at both the global and the local level of discourse structure. We have complemented and reformulated existing approaches with categories drawn from work in Systemic Functional Linguistics. We have proposed the generic stage as the unit of global discourse structure and we have reformulated both sentence-based and clause-based Centering in terms of SFL's system of TAXIS. Our goal in this study is to identify a segmentation approach for local discourse structure that best captures the contribution of RCs to textual cohesion. In order to find such an

[21] The remaining aspects of Kameyama's Centering approach will be accounted for in the next chapter, when we discuss embedded clauses.

approach, it is necessary to identify what constitutes a RC and what its discourse properties are. These are the topics we review next.

# CHAPTER 4: RELATIVE CLAUSES

In Chapter 3, we described two approaches to the segmentation of local discourse structure. These approaches differed in their notion of utterance. For sentence-based Centering, the main clause and all of its subordinate clauses constitute the unit of analysis. For clause-based Centering, certain clauses (e.g., conjuncts, finite adjuncts) but not others (e.g., non-finite adjuncts) constitute the unit of analysis. The status of RCs in the segmentation of discourse is problematic: They are finite clauses either embedded within a matrix clause or dependent on a dominant clause. This has led researchers to ponder whether RCs should be processed as independent units or in conjunction with the matrix or dominant clause to which they are attached or linked. Miltsakaki (2002, 2003, 2005) investigated the segmentation of RCs in Greek and English. Her study examined (a) the topical status of entities in restrictive and non-restrictive RCs in English as well as Greek and (b) the contribution of non-restrictive RCs to textual cohesion in both English and Greek. Her study dealt mostly with written language, and the non-restrictive RCs examined were limited to one type, sentence-final. Our study extends Miltsakaki's work in two ways: (a) We include restrictive and non-restrictive RCs from spoken and written texts in English and Spanish, and (b) we expand the classification of RCs beyond restrictive and non-restrictive to include other discourse properties. In this chapter, we briefly review

the structural and functional properties of Spanish and English RCs and end with a discussion of our proposed RC taxonomy.

## 4.1 Structural description

English and Spanish RCs are structurally similar in that in both languages, RCs are postmodifiers, that is, finite clauses that modify a (usually) nominal antecedent (Biber, Johansson, Leech, Conrad, & Finegan, 1999, p. 604; Brucart, 1999, p. 397). In addition to this general property, which Downing (1978, p. 381) identified as a consequence of the type of word order a language has (SVO in both cases), English and Spanish RCs share further similarities. They can differ in the type of gap, the type of relativizer and the properties of the head noun they modify. In the next few sections, we will review each of these properties.

### 4.1.1 RC defining properties

Downing (1978, pp. 377-80) identified the following as defining properties of RCs: (a) that they contain a nominal that is co-referential with a nominal outside the RC; (b) that the RC is an assertion about the relativized NP (or gap); and (c) that the RC modifies the antecedent NP (or head noun). The examples in (19) and (20) [22] show that this definition holds for English and Spanish RCs.

[22] These are my translations of Brucart's examples. I have tried to preserve as much of the Spanish word order as grammatically possible.

**19**

There are plenty of existing owners [$_{RC}$ who are already keen to make the move]. (NEWS) (Biber, et al., 1999, p. 608)

In Example (19), the RC *who are already keen to make the move* contains the nominal *who* which co-refers with the nominal *owners* found outside the RC. The RC makes an assertion about the relativized NP (which is co-referent with *who* and *owners*): They are already keen to make the move. The RC modifies its antecedent, the head noun *owners*.

**20**

La casa tenía dos habitaciones [$_{RC}$ que daban al parque]. (Brucart, 1999, p. 397)

'The house had two bedrooms that faced (to)-the park.'

In (20) the RC *que daban al parque* contains the nominal *que* that co-refers with the nominal *habitaciones* found outside the RC. The RC makes an assertion about the relativized NP (which co-refers with *que* and *habitaciones*): They face the park. The RC modifies its antecedent, the head noun *habitaciones*.

### 4.1.2 RC formation process

RCs in both languages not only share defining properties, they also share formation processes. Downing (1978, p. 388) observed that three independent processes may apply in RC formation: (a) an initial relative particle may be inserted; (b) the relativized NP may be copied in the form of a relative pronoun in clause-initial position; and (c) the relativized NP may be deleted. Of these three processes, the last two hold for English and Spanish RCs as in both languages, the relativized NP is copied in the form of a relative pronoun in clause-initial position

and the relativized NP is deleted. Given these formation processes, we can identify three major components of English and Spanish RCs: the gap (the relativized NP that has been deleted), the relativizer (the relative pronoun copy of the relativized NP that has been deleted) and the head noun (the antecedent of the RC) (Biber, et al., 1999, p. 608; Brucart, 1999, p. 398). These three components are co-referential. Examples (19) and (20) are reproduced below to illustrate the discussion of RC components.

**19**

There are plenty of existing owners [$_{RC}$ who are already keen to make the move]. (NEWS) (Biber, et al., 1999, p. 608)

As mentioned before, the RC *who are already keen to make the move* modifies the head noun *owners*. The RC has a subject gap. In other words, the relativized NP that has been deleted was the subject of the RC. The relative pronoun copy of the (deleted) relativized NP is the relativizer *who*. The head noun *owners* functions as the logical subject of the existential-*there* sentence and is co-referential with the subject gap and with the relativizer *who* in the RC.

**20**

La casa tenía dos habitaciones [$_{RC}$ que daban al parque]. (Brucart, 1999, p. 397)

'The house had two bedrooms that faced (to)-the park.'

In (20), the RC *que daban al parque* modifies the head noun *habitaciones*. The RC has a subject gap. In other words, the relativized NP that has been deleted was the subject of the RC. The relative pronoun copy of the (deleted) relativized NP is the relativizer *que*. The head noun *habitaciones* functions as the

direct object of the matrix clause and is co-referential with the subject gap and with the relativizer *que* in the RC.

### 4.1.3 Type of gap: The Noun Phrase Accessibility Hierarchy

A further area in which English and Spanish RCs show similar properties is in the type of positions within the RC that can be relativized (i.e., type of gap). While RCs in both languages allow for the relativization of subject and non-subject positions (Biber, et al., 1999, pp. 621-2; Brucart, 1999, p. 476; Gervasi, 2000, pp. 26-30; Keenan & Comrie, 1977, p. 79), not all non-subject positions identified in Keenan and Comrie's (1977) Noun Phrase Accessibility Hierarchy (NPAH) may be relativized in Spanish. In fact, while English allows for relativization of all positions in the NPAH, Spanish only allows the first five.

The NPAH is a universal characterization of RCs with respect to the syntactic positions that are relativized. According to the hierarchy, which is reproduced in (21) below, some positions are more easily relativized than others.

**21**

Subject > Direct Object > Indirect Object > Object of Preposition or Oblique > Genitive > Object of Comparative

The ranking in (21) indicates that a RC with a subject gap is more accessible for relativization than a RC with a direct object gap, which in turn is more accessible than a RC with an indirect object gap, and so forth (Keenan & Comrie, 1977, p. 66). The NPAH was proposed as an implicational hierarchy meaning that if a language allows for relativization of any position on the NPAH, it allows for relativization of all higher positions (i.e., positions to the left) (p. 69).

Examples of each position in the hierarchy (or type of gap) are provided in Table 9 below.

**Table 9. Examples of RCs as identified by the NPAH**

| NPAH | Example |
|---|---|
| Subject | I know the girl [$_{RC}$ who lives next door]. |
| Direct Object | I know the girl [$_{RC}$ who(m) you hate]. |
| Indirect Object | I know the girl [$_{RC}$ to whom you wrote a letter]. |
| Object of Preposition | I know the girl [$_{RC}$ for whom you bought the ring]. |
| Genitive | I know the girl [$_{RC}$ whose mail was lost]. |
| Object of Comparative | I know the girl [$_{RC}$ who(m) I'm younger than]. |

In addition to all positions in NPAH, English and Spanish also allow for RCs with adverbial gaps, which are not considered in Keenan and Comrie's (1977) typology. In RCs with adverbial gaps, the gap or relativized noun functions as an adverb (usually of place or time) in the RC. Example (22) and (23) below illustrate RCs with an adverbial gap. In (22), the *area* is the place in which chapels have closed. In (23), *la casa* is the place in which he was born.

**22**

the area [$_{RC}$ where the chapels have closed] (CONV) (Biber, et al., 1999, p. 624)

**23**

La casa [$_{RC}$ en donde nació] es ahora una tienda. (Gervasi, 2000, p. 23)

'The house where (he/she) was born is now a store.'

Some of these gaps, however, may be filled with resumptive pronouns, as shown in Examples (24) and (25). The presence of resumptive pronouns has been linked to difficulty in processing (Biber, et al., 1999, p. 622), which may be associated with their position in the NPAH (Keenan & Comrie, 1977, p. 92).

**24**

There was a case of one girl [$_{RC}$ who back in 1968 she killed two boys when she was eleven]. (CONV) (Biber, et al., 1999, p. 622)

In (24), the RC *who back in 1968 she killed two boys when she was eleven* modifies the head noun *girl*. The relativizer *who* is co-referential with the head noun *girl* as well as with the resumptive pronoun *she* (*she killed*), which fills the subject gap of the RC.

**25**

Hay profesores [$_{RC}$ que provocan ellos mismos la animadversión de sus estudiantes]. (Brucart, 1999, p. 405)

'There-are professors who provoke they themselves the animosity of their students.'

In (25), the RC *que provocan ellos mismos la animadversión de sus estudiantes* modifies the head noun *profesores*. The relativizer *que* is co-referential with the head noun *profesores*, as well as with the resumptive pronoun *ellos* found in the RC. The presence of resumptive pronouns in Spanish has been observed in particular when the relativizer *que* is used. Brucart (1999, pp. 403-4) suggested that the relativizer *que* may be losing its pronominal value and could be seen more as a subordinator, given that it is homophonic with the complementizer *que*.

### 4.1.4 Relativizer

English and Spanish RCs also allow for variation in the form of the relativizer. The relativizer can be a pronoun[23] or an adverb (Biber, et al., 1999, p. 608; Brucart, 1999, p. 397), as shown in Table 10. A crucial difference between the two languages is that Spanish does not allow for zero relativizers (Brucart, 1999, p. 398).

**Table 10. Relativizer types in English and Spanish**

| | Examples | |
|---|---|---|
| **Relativizer** | English | Spanish |
| Relative Pronoun | *who*, *which*, *that*[24], *whom*, *whose*, *zero* | *que* ('that'), *quien* ('who'), *cual* ('which'), *cuanto* ('how much'), *cuyo* ('whose') |
| Relative Adverb | *when*, *where*, *why* | *cuando* ('when'), *como* ('how'), *donde* ('where') *cuan* ('how much') |

### 4.1.5 Type of antecedent

As mentioned before, RCs usually modify nominal antecedents, and these nominal antecedents may have different grammatical roles in the matrix clause. The languages in our study allow the head noun to have the grammatical role of subject as well as the grammatical role of a non-subject, in the form of direct

[23] Brucart's (1999) relative adjectives (e.g., *cual* 'which', *cuanto*, and *cuyo* 'whose') were grouped with relative pronouns. This categorization is comparable with Biber et al.'s classification for English.

[24] The use of the term relative pronoun to refer to *that* follows Biber et al. (1999, p. 608). The status of *that* in syntactic theory is usually that of a complementizer and not a relative pronoun. Even within theoretical approaches to Syntax however, this status is not undisputed, as Sag (1997, pp. 462-4) considers *that* to be a relativizer rather than a complementizer. While other descriptive grammars also label *that* a relative pronoun (Sinclair, 1998, p. 362), the Oxford English Grammar (Greenbaum, 1996, pp. 225-6) adopts a more neutral position and uses the term *relative that*. Thanks to Nancy Hedberg for pointing the controversy about the status of *that*.

object, indirect object, object of a preposition, predicate or logical subject of existential *there*-sentences (Biber, et al., 1999, p. 623)[25].

Not only the grammatical role of the head noun but also its linguistic form can vary. RC antecedents can be explicit or implicit (i.e., non-expressed). Explicit antecedents of RCs can take different structural forms, both in English and Spanish. Both languages allow the relativization of nouns, personal pronouns, pronominal adverbs (such as *allá* and *there*) and sentences or propositions (Biber, et al., 1999, pp. 195, 767; Brucart, 1999, p. 398; Sinclair, 1998, pp. 362-3, 368). Examples are provided below (from Brucart, 1999, p. 398).

**26**

**RCs with personal pronoun heads**

Él, [$_{RC}$ que no está acostumbrado a perder], encajará este revés como una injusticia.
'He, who is not used to lose, will see this setback as an injustice.'

**RCs with pronominal adverb heads**

Iremos allá [$_{RC}$ dónde tú digas].
'(We) will go there where you say.'

**RCs with sentential heads**

Improvisó un discurso brillantísimo, [$_{RC}$ lo cual provocó general admiración].
'(He/she) improvised a brilliant speech, which caused great admiration.'

There are two types of RCs, however, in which the RC antecedent is not expressed as the head of the RC (Brucart, 1999, pp. 446, 449). Semi-free RCs are

[25] While no explicit discussion of grammatical role of the head noun was found in Brucart (1999) or Gervasi (2000), both sources provide examples of RCs with head nouns performing the types of grammarical roles identified for English.

RCs in which the nominal nucleus is elided[26]. Free RCs (or nominal RCs) are RCs that have no explicit antecedents (Biber, et al., 1999, pp. 195, 683). Examples of semi-free and free RCs are provided in (27) for English and (28) for Spanish.

**27**

(a) The one [$_{RC}$ she actually cancelled] was the one to her family because I had already cancelled the group one. (CONV) (Biber, et al., 1999, p. 353)

(b) [$_{RC}$ What baffles me] is how few of them can spell. (NEWS) (Biber, et al., 1999, p. 683)

In (27a) the pronoun *one* takes the place of a countable noun in the position of the RC antecedent. This is a type of cohesive device, *one*-substitution, (Halliday & Hasan, 1976), commonly used in English and not only in the case of semi-free relatives. Given that this is a de-contextualized example, we do not know what *one* refers to throughout the example.

In (27b), *what baffles me* is a RC in which the antecedent is not expressed. Biber et al. (1999, p. 683) argue that this type of construction can be paraphrased as a full RC if a general noun is taken as the head of the RC (as in *The thing that baffles me…*). Nominal RCs function as arguments in the main clauses that contain them, that is, as subjects and direct objects (Biber, et al., 1999, p. 193). In (27b), the nominal RC is the subject of the main clause.

[26] Because only RCs that are restrictive in function allow for non-expressed nominal nuclei, rephrasing of RCs as semi-free RCs is usually a test for restrictive function (see the 'one' test in section 4.2.2).

**28**

(a) La [$_{RC}$ que visitamos en Polonia] era espectacular. (Brucart, 1999, p. 446)

'The one that (we) visited in Poland was spectacular.'

(b) [$_{RC}$ Quien dice esto] miente. (Brucart, 1999, p. 449)

'Whoever says this lies.'

In (28a) the antecedent of the RC *que visitamos in Polonia* ('that (we) visited in Poland') is not fully expressed and must be retrieved from the context. Given the definite article *la*, we know that the antecedent is singular and feminine, which should constraint our search. In (28b) the RC *quien dice esto* ('whoever says this') has no explicit antecedent and functions as the subject of the main clause.

Including these two types of RCs in our study, given the possible difference in their contribution to textual cohesion, would cause us to increase the number of variables to be examined. In order to constrain our study to a manageable size, we have chosen to exclude free and semi-free RCs from the analysis[27,28].

We have seen then, that English and Spanish RCs share the same RC defining properties and the same RC formation processes identified by Downing (1978). In both languages, RC formation allows for the relativization of subject

[27] Biber et al. (1999, p. 195) take the view that sentential RCs (RCs that modify a proposition) are similar to nominal RCs because they can be paraphrased in the same way. We will include sentential RCs in our analysis, however, because they are functionally different from free RCs: While free RCs function as argument in clauses (Biber, et al., 1999, p. 193), sentential RCs are adverbial in function, that is, they comment on the content of the element they modify (Biber, et al., 1999, p. 867).

[28] Canac-Marquis and Tremblay (1997, p. 131) propose a reanalysis of non-restrictive RCs as restrictive RCs with a covert head that are in a relation of apposition to another NP. According to this view, non-restrictive RCs would also lack a linguistically expressed antecedent. We will include non-restrictive RCs in our analysis, as they perform a variety of discourse functions that play a role in textual cohesion (section 4.2.3).

and non-subject NPs within the RC. In both languages, relativizers may take the form of a relative pronoun or a relative adverb. And in both languages, antecedent NPs may be explicitly expressed in the form of nominals, pronominals or sentences, or partially or non-expressed in the case of semi-free and free relatives. There are however, some structural differences between the two languages. Spanish RCs must always be introduced with a relativizer, whereas English RCs allow zero-relativizers in some contexts. English allows relativization of all positions in the NPAH, but Spanish restricts relativization to the first five positions in the hierarchy and disallows the relativization of the object of the comparative. These differences notwithstanding, English and Spanish RCs can be said to be structurally similar. Whether the similarity also extends to their discourse function is the topic of the next section.

## 4.2 Functional properties of relative clauses

Functionally speaking, RCs in English and Spanish have been traditionally assigned to one of two types: restrictive (RRC) or non-restrictive (NRRC)[29]. In both languages, RRCs are said to identify the referent of the NP, whereas NRRCs are said to add descriptive information of a referent that has already been identified (Biber, et al., 1999, pp. 602-3; Butt & Benjamin, 2000, p. 494; Gervasi, 2000, pp. 4-10). This distinction, however, is not universally accepted. Biber et al. (1999, p. 602) noticed that there may be instances in which a distinction between RRCs and NRRCs cannot be made. In fact, Fox and Thompson (1990) and Weinert (2004) claimed that they were unable to distinguish between these

[29] Non-restrictive RCs have also been called appositive RCs. We will adopt the term non-restrictive in our study, and distinguish among NRRCs on the basis of their function.

two types of RCs on intonational grounds in their corpus study of English RCs (Fox & Thompson, 1990) and of English and German RCs (Weinert, 2004). The focus of these two studies was on a different functional classification of RCs. Fox and Thompson (Fox & Thompson, 1990) argued that discourse factors such as information status, grounding, humanness, definiteness and RC function play a role in determining the grammatical configuration of the RC, that is, the grammatical role of the head noun and the grammatical role of the relative clause gap. In particular, they observed that non-human subject head nouns tend to occur with object-gap RCs. This tendency was explained in terms of anchoring: The object-gap RC tends to have a human subject, in the form of a pronoun, that makes the introduction of the head noun relevant to the discourse. In other words, non-human referents are made relevant in relation to the human referents that manipulate them. Fox and Thompson's findings have had an impact on the study of RCs and have served as the basis for corpus studies in a variety of languages as well as studies in first and second language acquisition (e.g., Breivik, 1999; Collier-Sanuki, 1993; Diessel & Tomasello, 2000; Gervasi, 2000; Hadic Zabala, 2004; Kidd, Brandt, Lieven, & Tomasello, 2007; Reali & Christiansen, 2007; Weinert, 2004). Even though there is some psycholinguistic evidence that the structural configuration of the RC and its relation to information flow affect the processing of RCs in discourse (e.g., Gordon, Hendrick, & Johnson, 2004; Kidd, et al., 2007; Reali & Christiansen, 2007), it is not possible for us to include all possible factors in our analysis. Our study examines a different type of RC function, namely, the relationship between the RC and its head noun. While we

take referent identification as the starting point for our classification, the proposed RC taxonomy is not exclusively limited to the notion of restrictiveness.

This notion of restrictiveness, that is, the restrictive/non-restrictive distinction, is the starting point for the functional taxonomy of RC we propose in Table 15 at the end of the chapter. The traditional distinction between RRCs and NRRCs is complemented with a functional classification that brings together previous work on the functional properties of RCs (Chafe, 1988; De Haan, 1987; Halliday & Matthiessen, 2004; Lambrecht, 1988; Loock, 2007; Pérez González, 2006; Tao & McCarthy, 2001). The different functional types of RCs are discussed in detail in the following sections.

### 4.2.1 De Haan's (1987) relative clause taxonomy

The starting point for our functional taxonomy is De Haan's (1987) classification of RCs. De Haan (1987, p. 173) distinguished among three functions of RCs: (a) identifying RCs enable reference identification; (b) classifying RCs create new reference subclasses; and (c) describing RCs provide additional information. Examples of each are provided in Table 11 below (adapted from De Haan, 1987, p. 173).

**Table 11. Functions of relative clauses**

| Function | Example |
|---|---|
| Identifying | Which table did you buy? The table [$_{RC}$ that can be folded up]. |
| Classifying | What sort of table are you looking for? A table [$_{RC}$ that can be folded up]. |
| Describing | What sort of table did you buy? A table [$_{RC}$ that can be folded up]. |

De Haan's functional classification is complemented with a formal classification that distinguishes between definite and indefinite head nouns. The combination of these two parameters results in the RC taxonomy provided in Table 12.

**Table 12. De Haan's (1987) relative clause taxonomy**

| | | Function | | |
|---|---|---|---|---|
| **Definiteness** | | **identifying** | **classifying** | **describing** |
| | **indefinite** | -- | restrictive | non-restrictive |
| | **definite** | restrictive | -- | non-restrictive |

According to De Haan (1987, p. 174) RCs that modify definite noun phrases can have an identifying function, in which case they are restrictive in form, or they can have a describing function, in which case they are non-restrictive in form. RCs that modify indefinite noun phrases on the other hand, cannot have an identifying function. They can have a classifying function, in which case they are restrictive in form, or they can have a describing function, in which case they are non-restrictive in form[30].

[30] It is important to note that earlier work by (Smith, 1964) had distinguished determiners in RCs not only in terms of definiteness (definite and indefinite), but also in terms of specificity (unique, specified and unspecified). Unspecified determiners (such as *any* or *all*) occur with RRCs; specified determiners (such as *a*, *the*, and ø) with both RRCs and NRRCs; and unique determiners (ø of proper names) only with NRRCs.

We will look at the different sub-types of NRRCs in section 4.2.3. At this point, we focus on the distinction between identifying and classifying RRCs, a distinction that can be supported with the findings of Prince (1990). Prince (1990) explored the relationship between definiteness and RC function in a corpus of gap-containing and pronoun-containing (resumptive) RCs in Yiddish and English. Her findings about the distribution of resumptive pronouns in RCs reveal a distinction between RRCs with definite heads on one hand, and RRCs with indefinite heads and NRRCs on the other hand. Prince found that RRCs modifying definite heads do not allow resumptive pronouns, whereas RRCs modifying indefinite heads and NRRCs do allow them. Prince interpreted these findings in terms of Heim's (1983) file card account (cited in Prince, 1990). In the case of RRCs with definite heads, the definiteness of the head represents known information: The hearer already has file card for that entity and all he/she has to do is activate it. The RC contains information that must already be in the file card (i.e., old information), thus the RRC aids in the selection of the appropriate card. In the case of NRRCs, the hearer activates the relevant file card based on the contents of the head noun; the RC provides information that is presumably not present in the file card and must therefore be added to it. Finally, when it comes to RRCs with indefinite heads, the entity represented by the head is new information, therefore a new card must be created. The information contained in the RC is added to the card, much like the information contained in the NRRC. The information contained in the RRC cannot at this point be used to retrieve the card (as was the case with the RRC with a definite head), since there is no card to be retrieved.

Prince's (1990) findings thus support a distinction between identifying and classifying RRCs. Though a similarity is observed between classifying RRCs and NRRCs, our discussion in the next sections will justify keeping these two categories apart.

### 4.2.2 Lambrecht's (1988) relative clause taxonomy

A different RC taxonomy was introduced by Lambrecht (1988). Lambrecht distinguished among four types of RCs: (a) presentational relative constructions; (b) restrictive relative constructions; (c) appositive relative constructions; and (d) continuative relative constructions. The four types are illustrated in Example (29), (a), (b), (c) and (d) respectively (from Lambrecht, 1988, pp. 322-4, 328).

**29**

(a) Once upon a time, there was an old cockroach [$_{RC}$ who lived in a greasy paper bag].

(b) The cockroach [$_{RC}$ who lived in the paper bag] was very arrogant.

(c) The cockroach, [$_{RC}$ who was very arrogant], was hated by all his neighbours.

(d) The cockroach was very arrogant, [$_{RC}$ which is surprising], since cockroaches are known to be humble beings.

These four types of RCs can be assigned to two different groups, based on their constituency: In restrictive relative constructions and appositive relative constructions, the RC is a nominal modifier embedded in the NP structure, whereas in presentational relative constructions and continuative relative constructions, the RC is at the same structural level as the main clause.

As shown in Figures 1 and 2 (from Lambrecht, 1988, pp. 323-4), the RCs in restrictive relative constructions and appositive relative constructions are noun modifiers and form a complex NP with the head noun they modify.

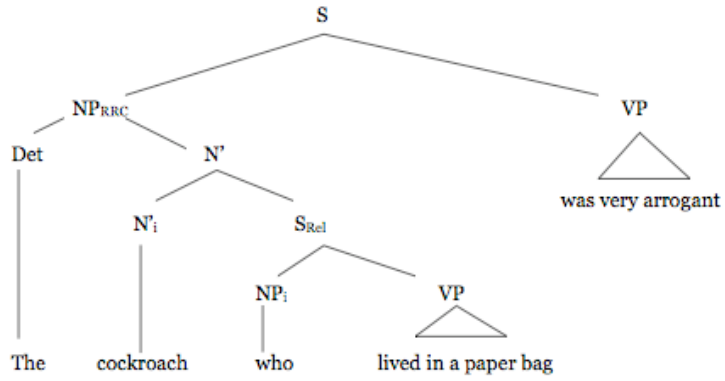


**Figure 1. Structure of the restrictive relative construction**

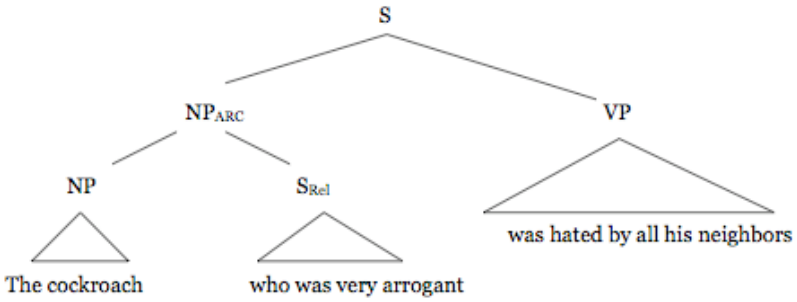


**Figure 2. Structure of the appositive relative construction**

There are functional and structural differences between restrictive and appositive relative constructions. Functionally, restrictive relatives restrict "the set of possible referents of the noun phrase" whereas appositive relatives add "a piece of parenthetical information to an NP referent" (Lambrecht, 1988, p. 328). Structurally, they differ in the position they occupy within the wider NP: While

the restrictive relative construction is a sister to N', the appositive relative construction is a sister to the NP. One way to distinguish between them structurally is with the *one*-substitution test shown in Example (30) (from Lambrecht, 1988, p. 324).

**30**

(a) The one who lived in the paper bag was very arrogant.

(b) *The one, who was very arrogant, was hated by his neighbours.

The noun *cockroach* can be replaced with *one* in (30a) but not in (30b). Given that *one* is said to replace N' (Lambrecht, 1988, p. 324)[31], the fact that this substitution is possible in (30a) but not in (30b) is taken as support for the structural distinction between restrictive relatives and appositive relatives.

Figure 3 (from Lambrecht, 1988, p. 329) shows that the two clauses that form the presentational relative construction are sisters at the structural level. The first clause, the presentational clause, introduces the new referent into the discourse and anchors it through the pseudo-locative *there*. The second clause, the RC, tends to have a subject gap and usually expresses a proposition about its antecedent, that is, the new referent. The new referent is the focus of the main clause, but the topic of the RC. In this way, the speaker can satisfy the information-structural requirements that are involved in introducing a new referent to the discourse and making an assertion about it.

[31] Lambrecht (1988) seems to draw from (McCawley, 1981, p. 103), who in turns cites Jackendoff (1977) and Baker (1978).

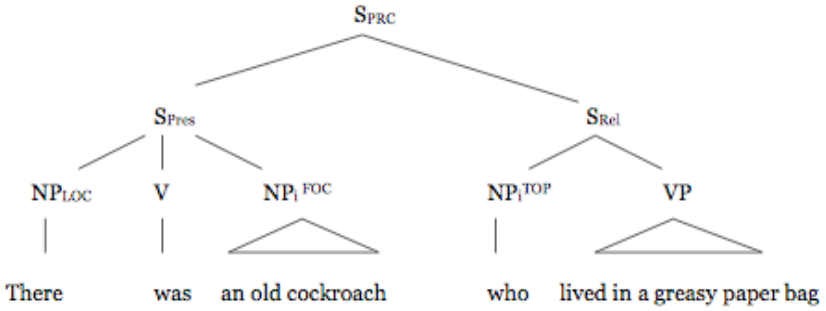


**Figure 3. Structure of the presentational relative construction**

The final type of RCs identified by Lambrecht, the continuative relative construction, is structurally similar to the presentational relative construction in that the main clause and RC are at the same structural level, as shown in Figure 4 (from Lambrecht, 1988, p. 329). Continuative relatives, however, have a markedly different discourse function. Drawing from Jespersen's work, Lambrecht (1988, p. 328) argued that continuative relatives do not serve as nominal modifiers but rather have the function of "continuing a narrative, or of establishing a temporal or logical link between two states of affairs". Continuative relatives are used to link states of affairs, that is, to create relational coherence. They characteristically occur in clause-final position and they tend to have propositional – rather than nominal – antecedents.

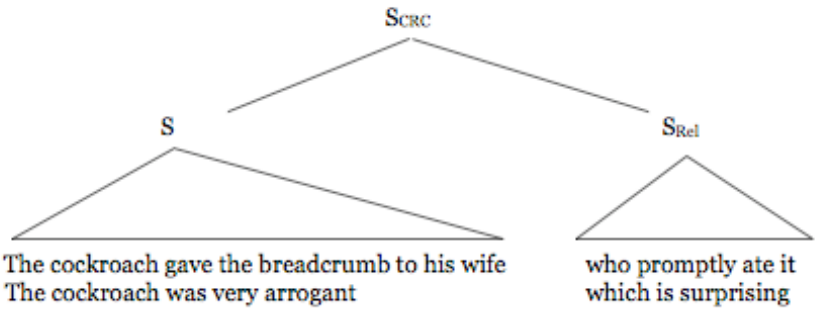


**Figure 4. Structure of the continuative relative construction**

It is important to note that two of Lambrecht's RC types have been subsumed in one category by other researchers. In his discussion of intonation unit linkages in spoken English, Chafe (1988, pp. 20-1) defined NRRCs as RCs that constitute a separate intonation unit from the 'main' clause. Chafe observed that these NRRCs can continue a narrative (like continuative relatives), or add additional information about a known referent (like appositive relatives). Furthermore, they can have nominal or propositional referents, to which they are linked with relative pronouns that are said to be "a one-word substitute for linkages of the type *and he*, *and it*, etc" (Chafe, 1988, p. 21).

Tao and McCarthy (2001), in a corpus study of English NRRCs, also made reference to Jespersen's (1909) distinction between appositive NRRCs and continuative NRRCs. Appositive NRRCs are embedded within the matrix clause, whereas continuative NRRCs follow the matrix clause. Tao and McCarthy faced difficulties in assigning spoken examples of NRRCs to either one of these two categories, and as a result, argued that it may be best to treat this distinction between appositive and continuative as a continuum.

To bring clarity into what constitutes a NRRC, the next section presents a classification of NRRCs based on functional properties that have been validated in corpus studies in both English and Spanish.

### 4.2.3 Non-restrictive relative clauses

#### 4.2.3.1 Evaluative, expanding and affirmative NRRCs

In a corpus study of English NRRCs, Tao and McCarthy (2001) identified three major functions of NRRCs: evaluation, expansion and affirmation. In a later study, Pérez González (2006) extended Tao and McCarthy's taxonomy to Spanish NRRCs. In the Spanish study however, data analysis was constrained to NRRCs with the relativizer *lo cual*, which is said to be always non-restrictive in Spanish (p. 406 and references therein). The distribution of NRRCs in Spanish was nevertheless similar to the one observed for English. Our discussion of Tao and McCarthy's categories will be complemented with Spanish examples taken from Pérez González (2006).

In both corpus studies, NRRCs expressing evaluation were the most frequent type of NRRCs. The function of evaluative NRRCs is to express the speaker's attitude or opinion towards the antecedent proposition. This function was sometimes explicitly marked, either by attitudinal discourse markers (e.g., *actually*), by modal expressions of certainty (e.g., *probably*) or by modal verbs (e.g., *would*) (Tao & McCarthy, 2001, p. 663). Structurally, they were mostly clause-final (vs. centre-embedded). Example of evaluative NRRCs are provided in (31) and (32). In (31), the speaker makes an evaluation on his/her brother's spending for Christmas. The evaluation is explicitly marked by *I think* and *silly*.

**31**

Context: Speaker is complaining about the materialism that dominates Christmas.

<speaker 1> I know my brother goes into debt for the kids every Christmas you know like if they don't spend two hundred pound on them you know it's not enough (<speaker 2> Mm) [$_{RC}$ which I think is silly] but that's the way of things today (Br.) (Tao & McCarthy, 2001, p. 663)

In (32), the RC *lo cual, ciertamente, merecían* expresses the belief that the defeat reported in the previous clause was entirely justified. This evaluative function is additionally expressed with an attitudinal discourse marker, namely the modal adverb *ciertamente* ('certainly').

**32**

derrotado en un, digamos, doscientos por ciento, [$_{RC}$ lo cual, ciertamente, merecían]. Ahora hay una nueva (Pérez González, 2006, p. 406)[32]

'defeated by, let's say, two hundred per cent, which (they) certainly deserved. Now there is a new'

Expanding NRRCs were the second most frequent type of NRRC observed by Tao and McCarthy (2001). The NRRCs that were assigned to the category of expansion added information that was seen by the speaker as topically relevant. They also tended to occur in clause-final position. Examples are provided in (33) and (34) below. In (33), the RC *which is aisle and next to it on the window side* provides additional relevant information on where the seats E and G are located. In (34), the RC *lo que- lo cual con su invalidez, pues no puede hacer* provides relevant information about the health of the person that explains why working as a maid is out of question.

[32] The examples in Pérez González (2006) were extracted with a corcondancer and for that reason, the context is not provided in full sentences.

**33**

Context: Speaker 2 is on the phone checking seat allocations with speaker 1, an airline employee, for a forthcoming plane trip.

<speaker 1> What we've got is E and G [$_{RC}$ which is aisle and next to it (<speaker 2> Yeah) on the window side (<speaker 2> Yeah) but not the window seat

<speaker 2> Mm (Br.) (Tao & McCarthy, 2001, p. 664)

**34**

trabajar como empleada de hogar, [$_{RC}$ lo que- lo cual con su invalidez, pues no puede hacer], y si piensa que (Pérez González, 2006, p. 406)

'to work as a maid, what- which with his/her handicap, (he/she) can not do, and if (he/she) thinks that'

Interestingly, Tao and McCarthy (2001) pointed out that though not inherently evaluative, these NRRCs are not devoid of evaluative or attitudinal information, as they occur in evaluative contexts and they contain information that is supportive of evaluation.

The third type of NRRCs identified by Tao and McCarthy (2001) are affirmative NRRCs. These clauses confirm that the action expressed in the antecedent proposition will or will not take place, or that it already has or has not taken place. Examples are provided in (35) and (36) below. In Example (35), the RC *which you did* confirms that the action in the main clause (*when you divide negative two you have to flip all the signs over*) has already taken place. This is also the function of the RC *lo cual hizo al día siguiente, verdad* in Example (36), where the RC confirms that the subject of the RC did meet with the minister, and that he/she did the next day.

**35**

Context: Speaker is discussing a mathematical operation

When she, you divide negative two you have to flip all the signs over… [$_{RC}$ which you did] (Am.) (Tao & McCarthy, 2001, p. 665)

**36**

tenía que ver al Ministro que es amigo personal, [$_{RC}$ lo cual hizo al día siguiente, verdad] ¿no? Bueno vamos, no co (Pérez González, 2006, p. 406)

'(He/she) had to see the Minister who is a personal friend, which (he/she) did the next day, true, right? Well let's see,'

In order to facilitate a comparison of the results, Table 13 below provides a tabulated version of the distribution of NRRCs in English and Spanish (from Pérez González, 2006, p. 407). Given the similarities in the distribution, we can assume a similarity in the function of NRRCs in Spanish and English texts, which allows us to code Spanish and English RCs with the same functional taxonomy.

**Table 13. Functional types of NRRCs in Spanish and English**

| **Functional Type** | ***Which* NRRCs** | | ***Lo cual* NRRCs** | |
|---|---|---|---|---|
| | N | % | N | % |
| Evaluation | 429 | 62.0 | 389 | 67.8 |
| Expansion | 215 | 31.1 | 146 | 25.5 |
| Affirmation | 25 | 3.6 | 22 | 3.9 |
| Other | 23 | 3.3 | 16 | 2.9 |

### 4.2.3.2 Continuative, relevance and subjectivity NRRCs

A different functional taxonomy of NRRCs was proposed by Loock (2007). In his taxonomy, Loock distinguished among continuative, relevance and subjectivity

NRRCs (which he calls appositive relative clauses, ARCs). The properties of these three types of NRRCs are summarized in Table 14.

**Table 14. Loock's (2007) taxonomy of NRRCs**

| |
|---|
| **Continuative NRRC:** |
| Creates narrative dynamism (i.e., moves narrative time forward) |
| Antecedent is an NP. |
| **Relevance NRRC:** |
| Levelling of the shared cognitive space (information is supplied to compensate any difference in the amount of knowledge shared by the participants); |
| Legitimacy of the antecedent (if the antecedent is completely new to the addressee, the NRRC legitimates it presence) |
| Explanation, justification, concession (the speaker can explain, justify or oppose the proposition expressed by the main clause; the link between the two clauses is to be inferred by the addressee) |
| Antecedent is a proper noun, usually in subject position. |
| **Subjectivity NRRC:** |
| Comments and judgments (speaker gives own opinion); |
| Correction (corrects or reformulates its antecedent); |
| Assessment and conclusion (provides interpretation of antecedent). |
| Antecedent is frequently sentential. |

The main function of continuative NRRCs is to move the narrative time forward. They are said to have the type of narrative dynamism normally associated with independent clauses, which is why this type of NRRC is seen as hierarchically independent from its main clause. This hierarchical independence what allows us to rephrase the NRRC in (37a) as an independent clause in (37b) (from Loock, 2007, pp. 340-2). There is, however, a semantic dependency relation between the main clause and the NRRC, which is what makes it impossible for us to negate the main clause without affecting the NRRC, as in

(37c). In other words, if she hadn't been airlifted to the hospital that would not have been the place where she died.

**37**

(a) (TABL_ARC301) She was found face down in the water and airlifted to hospital, [$_{RC}$ where she died hours later].

(b) (mTABL_ARC301) She was found face down in the water and airlifted to hospital, and she died there hours later./ She died there hours later.

(c) (mTABL_ARC301) #She wasn't airlifted to hospital, [$_{RC}$ where she died hours later].

The second type of NRRC, relevance NRRCs, are "used to make relevant the antecedent or the predicate in which it appears." (Loock, 2007, p. 346) This type of NRRC, then, corresponds to Tao and McCarthy's expanding category. In Example (38) (from Loock, 2007, p. 347), Tony Sewell's authority on black pupils is legitimized by the relevance NRRC. Without the NRRC, the reader may not know who Tony Sewell is, nor why he is in a position to make such a claim.

**38**

(QUAL_ARC9) Tony Sewell, [$_{RC}$ who has just finished an inquiry into soaring levels of exclusions among black pupils form a London school], claimed that too much concern with money and consumer goods was almost as damaging to black pupil's chances as racism.

Finally, the last type of NRRC identified by Loock (2007) is the subjectivity NRRC. Subjectivity NRRCs are used to convey an opinion or a judgment and are thus equivalent to Tao and McCarthy's evaluative category. In subjectivity NRRCs, the NRRC and the main clause are at two different levels: The main clause is at a referential level; the NRRC at a commentary level. In Example (39), the subjectivity NRRC *who might not have qualified anyway* expresses the

author's opinion and is just a comment on what actually happened, *that the men's 4 x 00 m team went out in the heats when they bungled a change-over, straying out of the prescribed area* (from Loock, 2007, p. 353).

**39**

(QUAL_ARC117) The men's 4 x 00 m team, [$_{RC}$ who might not have qualified anyway], went out in the heats when they bungled a change-over, straying out of the prescribed area.

## 4.3 A RC taxonomy

Having reviewed the structural and discourse properties of English and Spanish RCs, we now turn to the discussion of the taxonomy we will adopt for this study. First, we briefly discuss the classification of RCs in terms of SFL's system of TAXIS. We continue with our functional classification of RCs and conclude with our approach to RC segmentation, which we formulate, as we did before, in terms of SFL's system of TAXIS.

### 4.3.1 Relative clauses in Systemic Functional Linguistics

In SFL, RRCs (called defining relative clauses) are seen as embedded bound clauses whereas NRRCs (called non-defining relative clauses) are seen as dependent clauses in hypotactic relations. This distinction between TAXIS and embedding is of particular significance: "Whereas parataxis and hypotaxis are relations ***between*** clauses (…), embedding is not" (Halliday & Matthiessen, 2004, p. 426). In fact, embedding is a rank shift, "by which a clause or phrase comes to function within the structure of a group." (Halliday & Matthiessen, 2004, p. 426).

RRCs are downranked (i.e. embedded) clauses that function as qualifiers in nominal groups. They typically express the logico-semantic relation of expansion, that is, they "specify which member or members of the class designated by the Head noun (…) is or are being referred to" (Halliday & Matthiessen, 2004, p. 428).

NRRCs are clauses in a hypotactic relation to another clause. As we mentioned in section 3.2.3, in hypotactic relations, one clause is dominant and the other clause is dependent on the dominant clause. NRRCs are said to differ from RRCs in meaning and in expression. Instead of defining a subset (which is the meaning of a RRC), a NRRC "adds a further characterization of something that is taken to be already fully specific. This 'something', therefore, is not necessarily just a noun; the domain of a non-defining relative may be a whole clause, (…), or any of its constituents" (p. 400). Indeed, Halliday and Matthiessen (2004, pp. 400-1) identify three domains for NRRCs: the dominant clause or some part of it, a nominal group, or some expression of time or place.

Halliday and Matthiessen (2004, pp. 399-400) observe that most NRRCs express the logico-semantic relation of elaboration and have one of the following functions: (a) They introduce background information into the discourse; (b) they introduce a characterization; (c) they introduce an evaluation; or (d) they introduce an interpretation of some element of the dominant clause. Following Loock (2007), we can group functions (a), (b) and (d) under relevance NRRCs and place function (d) in the subjectivity NRRC category. NRRCs may also

express the logico-semantic relation of extension[33], in which case they are said to be semantically additive (Halliday & Matthiessen, 2004, p. 402). This type of NRRCs moves the narrative forward and could be seen to fit Loock's (2007) description of continuative NRRCs. We illustrate Halliday and Matthiessen's description of NRRCs with Example (40) (from Halliday & Matthiessen, 2004, p. 399).

**40**

Yu, [$_{RC}$ who has been visiting Taiwan this week], did not elaborate. (Text 13)

In this example, the NRRC *who has been visiting Taiwan this week* is in a hypotactic relation of elaboration with the dominant clause *Yu did not elaborate*. The domain of the NRRC is the nominal group *Yu*. The NRRC has the function of introducing background information.

### 4.3.2 A functional taxonomy of relative clauses

As we have seen throughout our discussion of the different approaches to RC classification, there is some overlap among the different categories. We noted, for example, that Chafe's (1988) description of NRRCs included Lambrecht's continuative as well as appositive relative constructions. In the preceding section, we saw that Loock's (2007) NRRC categories could be overlapped with Halliday and Matthiessen's (2004) NRRC functions. Drawing from all of these classifications, in this section we aim to provide a taxonomy of RCs that is functionally based and that minimizes overlapping between categories.

[33] She told it to the baker's wife, [$_{RC}$ who told it to the cook]. (Halliday & Matthiessen, 2004, p. 402)

The starting point for our taxonomy is the restrictive/non-restrictive distinction that has guided our discussion of RCs in this chapter. Following Halliday and Matthiessen (2004), we see RRCs as downranked clauses, as embedded clauses within a nominal group and NRRCs as dependent clauses in a hypotactic relation with a dominant clause.

Within RRCs, we follow De Haan's (1987) distinction between identifying RRCs and classifying RRCs. Identifying RRCs have a definite head; classifying RRCs have an indefinite head. Identifying RRCs overlap with Lambrecht's (1988) restrictive relative constructions. And Lambrecht's (1988) presentational relative construction can be seen as a type of classifying RRC, one in which the main clause is an existential *there* sentence[34].

Within NRRCs, we adopt Loock's (2007) functional classification, albeit with a minor label change, to avoid confusion. We distinguish among narrative NRRCs (Loock's continuative NRRCs), relevance NRRCs and subjectivity NRRCs. The category of narrative NRRCs subsumes Lambrecht's continuative RCs, and Tao and McCarthy's affirmative RCs. The latter were included in the narrative type as they tend to confirm that an event took place while at the same time adding information about when or where or how that event happened. The

[34] It is important to note that we do not adopt Lambrecht's structural description of presentational relative constructions. Lambrecht (1988) sees presentational relative constructions as bi-clausal, where both clauses are of equal status (see Fig. 2). In our view, presentational RCs constitute a subtype of RRCs and are thus embedded clauses. This is consistent with research in first language acquisition (Diessel & Tomasello, 2000) which has argued that these presentational constructions consist of one proposition. Diesel and Tomasello examined children's spontaneous production of RCs and found that the first RCs produced by English-speaking children are presentational constructions like the one in example below, (from Diessel & Tomasello, 2000, p. 137), where the relative clause is attached to the nominal predicate of the presentational copular clause (Here's), which in itself carries no propositional content.

Here's a tiger ***that's*** *gonna scare him.* (Nina 3;1)

category of relevance NRRCs includes both Lambrecht's appositive relative construction and Tao and McCarthy's expanding NRRCs. Our last category, subjectivity NRRCs encompasses Lambrecht's appositive (when it expresses an opinion and not when it provides background information) and Tao and McCarthy's category of evaluative NRRCs.

The complete taxonomy is provided in Table 15 below.

**Table 15. A functional taxonomy of relative clauses**

| Type | Function | Description |
|---|---|---|
| Restrictive | Identifying | RC adds information necessary for referent identification. |
| | Classifying | RC creates a new subclass within a reference class. |
| Non-restrictive | Narrative | RC moves the narrative time forward. |
| | Relevance | RC makes an antecedent relevant in context. |
| | Subjectivity | RC expresses opinion. |

### 4.3.3 Relative clause segmentation

In section 3.2.3, we formulated sentence-based and clause-based Centering in terms of SFL's system of TAXIS, as Paratactic-Clause Centering and Hypotactic-Clause Centering. The main goal of our research is to assess the contribution of RCs to textual cohesion in order to establish a segmentation approach that is functionally motivated. In order to incorporate RCs to the segmentation approaches, we need to include embedded clauses in our definitions. In (41) below, the segmentation approaches have been updated to include both RRCs and NRRCs. In order to exhaust all possibilities, a new approach to utterance segmentation is added to the approaches outlined in Chapter 3, embedded-clause

Centering, one in which all clauses, paratactic, hypotactic and embedded constitute center-updating units.

**41**

**Paratactic-Clause Centering**

A center-updating unit is an independent clause or a clause that is in a paratactic relation with another clause. Clauses that are in a hypotactic relation with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded.

**Hypotactic-Clause Centering**

A center-updating unit is an independent clause or a clause that is either in a paratactic relation or a hypotactic relation with another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded.

**Embedded-Clause Centering**

A center-updating unit is an independent clause or a clause that is either in a paratactic relation, in a hypotactic relation with another clause or embedded within another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Non-finite embedded clauses belong to the center-updating unit of the clause in which they are embedded.

Our reformulation of sentence-based and clause-based Centering in terms of SFL's paratactic, hypotactic and embedded clauses allows us to incorporate both RRCs (embedded clauses) and NRRCs (hypotactic clauses) to our

segmentation approaches. At the same time, we are also able to account for the other types of clauses identified in Kameyama's Intrasentential Centering, namely, reported speech complements[35] and nonreport complements[36] in a systematic way.

## 4.4 Summary

In this chapter, we have seen that English and Spanish RCs are both structurally and functionally similar. In spite of some structural differences concerning the presence of obligatory relativizers and the relativization of object of the comparative NPs, in both languages, subject and non-subject NPs may be relativized, relativizers may take the form of pronouns or adverbs, and antecedents be explicitly expressed or not. Functionally, the distinction between restrictive and non-restrictive RCs has been traditionally recognized in both languages. The results of recent corpus studies (Pérez González, 2006; Tao & McCarthy, 2001) have shown the distribution of NRRCs to be strikingly similar in the two languages.

The similarities we have observed allow us to apply the RC taxonomy we have proposed to both languages. This taxonomy takes the restrictive/non-restrictive distinction as its point of departure, a distinction that we ground in SFL's classification of clauses and we expand with findings of corpus studies. It is our incorporation of SFL's clause taxonomy what enables us to formulate three clear and systematic approaches to discourse segmentation. Armed with a

[35] Reported speech complements constitute paratactic clauses if they are quoted speech and hypotactic clauses if they are reported speech.

[36] Nonreport complements constitute embedded clauses.

taxonomy of RC functions and these systematic segmentation approaches, we can now investigate the contribution of the different types of RCs to textual cohesion in English and Spanish texts.

# CHAPTER 5: METHODOLOGY

## 5.1 Corpus linguistics

### 5.1.1 Corpus-based vs. intuition-based studies

The data for this study comes from a corpus, a collection of naturally occurring texts. The importance of a corpus-based study in order to identify the function of RCs in discourse cannot be denied and, in this particular case, a corpus-based study is preferred to speakers' intuitions.

The choice between corpus-based and intuition-based research has been greatly debated in the field of linguistics. Butler (2004, pp. 148-150) briefly summarizes the controversy around the nature of linguistic data for formalist and functionalist approaches. How we study language is a consequence of how we perceive language: If language is a mental representation, then the study of language is the study of linguistic competence or internalized language; if language is a communicative action, then the study of language is the study of language use.

In his book on *Knowledge of Language: Its Nature, Origin, and Use,* Chomsky (1986, p. 2) explains that starting in the mid-1950s there has been a shift in linguistics towards the rationalist tradition of Roger Bacon, Beauzée, John Stuart Mill, Ralph Cudworth, and James Harris. Following this rationalist tradition, Chomsky argues that linguistics is concerned with the study of " 'the language faculty,' which is understood to be a particular component of the human

mind" (p. 3). Knowledge of language is thus seen "as a certain state of the mind/brain, a relatively stable element in transitory mental states once it is attained; furthermore, as a state of some distinguishable faculty of the mind – the language faculty – with its specific properties, structure, and organization, one 'module' of the mind" (p. 13). Linguistics thus becomes the study of an element of the mind, that is, the study of internalized language (I-language) (p. 22). The pairing of forms and meanings that had traditionally been associated with linguistics and the study of language is now seen as externalized language (E-language) (p. 20) and beyond the scope of linguistic inquiry: "an epiphenomenon at best" (p. 25).

Researchers in the fields of functional linguistics and corpus linguistics (e.g., Biber, Conrad, & Reppen, 1998; Chafe, 1992; Fillmore, 1992; Halliday, 2004) see linguistics as the study of language use. This, however, does not mean that intuitive language is not necessary for linguistic analysis. Fillmore (1992, p. 35) makes the following two points:

> The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body.

In other words, the study of language requires both a reliance on intuitive knowledge and a reliance on natural corpora. The second is of particular importance given the authenticity of natural corpora in comparison with the

sometimes not-so-reliable linguistic intuition (p. 38). The study of natural corpora, however, cannot tell us what is not possible in language: “there are no corpora of starred examples” as noted by Fillmore (1992, p. 58). That is the realm of native-speaker intuition.

The choice between a corpus-based versus an intuition-based methodology lies on our understanding of language, and on the nature of our inquiry, that is, on the types of questions we ask. If we want to understand how language is used, then corpus research would be more insightful; if we want to describe the internalized linguistic knowledge Chomsky refers to, intuitions may be relevant. In this study, we want to understand how RCs contribute to the textuality (or texture) of a text. The nature of our question demands that we look at a corpus of RCs that have been found in naturally occurring discourse to see the different ways in which the different types of RCs (identified in Chapter 4) contribute to textual cohesion. In order to do so we measure the degree of cohesion between an utterance containing a RC and its immediate co-text, as modelled by Centering Theory. Before discussing the specifics of our corpus and the analysis undertaken, we address two theoretical issues concerning corpus studies: (a) the distinction between Corpus Linguistics as a field and corpus linguistics as a methodology; (b) the distinction between corpus-based and corpus-driven studies.

### 5.1.2 Corpus Linguistics vs. corpus linguistics

The status of corpus linguistics within the field of linguistics seems to be a matter of debate. Bas Aarts (2000, p. 7) sees corpus linguistics as a methodology in

linguistics: Corpus linguists use corpus data to make claims about language. From this perspective, there is no conflict between corpus linguistics and theoretical linguistics, inasmuch as corpus linguistics can provide linguistic evidence for theoretical claims. On the other side of the controversy, Jan Aarts (2002), Leech (1992), Teubert (2005) and Tognini-Bonelli (2001) (among others) see Corpus Linguistics as a discipline on its own and different from Theoretical Linguistics. Jan Aarts (2002, p. 14) explains:

> (...) whereas the former [theoretical linguistics] is concerned with the phenomenon of language in general (call it the language faculty, competence, I-language or universal grammar) and its object of study is therefore language-independent, corpus linguistics is concerned with language use, that is, with the way in which a given language system manifests itself in a particular context and cultural environment. It aims therefore not only at an understanding of the language system, but also of the parameters that determine the selection from the possibilities offered by the language system and thus shape the ultimate form that utterances must take in order to be appropriate within their micro- and macro-context.

In our view, Linguistics is the study of the language system as well as the study of the parameters that determine the selection from the possibilities. As a result, we do not distinguish corpus linguistics from theoretical linguistics on a philosophical basis, since we believe Linguistics to be the study of language use. The fact that different corpus studies make different contributions to Linguistic Theory, as we will see in the next section, suggests to us that corpus linguistics is more of a methodology than a field on its own.

### 5.1.3 Corpus-based vs. corpus-driven studies

Corpus studies can contribute to linguistic theory in two ways, either by illustrating or validating a theoretical point or by allowing a theoretical point to

emerge from data analysis. Based on their contribution, corpus studies can be corpus-based or corpus-driven (Tognini-Bonelli, 2001). In corpus-based studies, previously formulated theoretical notions inform and guide the analysis of the corpus. When the evidence from the corpus and the theoretical notions collide, theoretical notions prevail. In corpus-driven studies, theoretical notions do not predate but are derived from the corpus: "the theory has no independent existence from the evidence and the general methodological path is clear: observation leads to hypothesis leads to generalisation leads to unification in theoretical statement" (Tognini-Bonelli, 2001, pp. 84-5).

It is clear, following our discussion of RCs in Chapter 4, that certain theoretical notions, that is, the structural and functional properties of RCs, predate our analysis of the corpus. Ours, then, is not a corpus-driven study, but a corpus-based study. We do not, however, use the corpus just to find examples of our theoretical constructs, but rather to test hypotheses that will widen our understanding of the contribution of RCs to textual cohesion.

## 5.2 Our corpus

Even though, as noted by Butler (2004, p. 150), the term corpus could be applied to any body of texts, within the framework of corpus linguistics, a corpus refers to a collection of pieces of language that have certain properties. For example, EAGLES, the Expert Advisory Group on Language Engineering Standards, defines corpus as "a collection of pieces of a language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (EAGLES, 1996). Tognini-Bonelli (2001, p. 55) sees a corpus as "a

computerised collection of authentic texts, amenable to automatic or semi-automatic processing or analysis. The texts are selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language." Based on this definition, Tognini-Bonelli (2001, p. 54) identifies two issues that are particularly relevant in corpus studies: (a) the authenticity of the texts; and (b) the representativeness of the language. Corpora are authentic when they are collections of genuine communications that were not specifically produced for linguistic analysis. They are representative when a wide range of language variability is included in different text types. The importance of representativeness in a corpus study has been emphasized by many in corpus linguistics (e.g., Butler, 2004; Hunston, 2002; Johansson, 1995). If we want to be able to apply the statements that we derive from the corpus to a larger sample, and to the language as a whole, our corpus must be representative. Johansson (1995, p. 246) tells us, however, that it is not statistical representativeness that is mandatory. What is important, he argues, is that the corpus "be representative in some sense, by reflecting the variation of text types and linguistic choices in the language."

When collecting our corpus of RCs, we tried to achieve representativeness by carefully choosing the genres from which the texts would be selected. We drew the concepts of mode and tenor from SFL's Register Theory (Eggins, 1994, pp. 52-65) to distinguish between written and spoken texts on one hand, and formal and informal texts on the other. SFL identifies three dimensions of the context of

situation, that is, register[37], that affect text realization (Eggins, 1994, p. 52). These dimensions are field, mode and tenor. Field refers to what is talked about, tenor to the role relationships that hold between the participants, and mode to the role language plays in the interaction (Eggins, 1994, p. 52). Mode and tenor are of particular relevance to us. Following Martin (1984), Eggins (1994, p. 53) argues that the role of language in the interaction can be described in terms of the possibilities of immediate feedback (spatial/interpersonal distance) and the role of language as either action or reflection[38] (experiential distance). These two dimensions combined provide the characterization of spoken and written language provided in Table 16 below (adapted from Eggins, 1994, p. 55).

**Table 16. Characteristics of spoken and written language situations**

| Spoken discourse | Written text |
|---|---|
| interactive | non-interactive |
| face-to-face | not face-to-face |
| language as action | not language as action |
| spontaneous | not spontaneous |
| casual | not casual |

It is important to notice, though, that the characterizations provided above refer to typical or prototypical spoken and written language situations. The use of modern media (e-mail, text-message, chat) has blurred the distinction between these two types of situations, especially with respect to interactivity, spontaneity and casualness. Eggins (1994) argues that the different characteristics between

[37] We mentioned register, the context of situation, in our discussion of stages (section 3.1.2).

[38] Language as action accompanies a social process, for example, when playing a game of bridge. Language as reflection constitutes a social process, for example, when writing a novel (Eggins, 1994, p. 54).

spoken and written language situations are reflected in the language used in those situations. An example of that is the higher frequency of nominalizations found in written texts in comparison to spoken discourse. Eggins (Eggins, 1994, p. 56) explains that "these differences are not accidental, but are the functional consequence (the reflex) of the situational differences in mode." Following our note on new media, text-messaging is a written language situation that differs from the typical written language situation in that it is interactive, spontaneous and casual. Given that linguistic features are a functional consequence of the situational mode, we would expect the language of text message to resemble that of spoken discourse more than that of written texts.

The second dimension of register that was important in our text selection process was tenor. As we mentioned earlier, tenor refers to the relationship that holds between the participants in the interaction. Following Poynton (1985), Eggins (1994, p. 64) distinguishes three aspects of tenor: power, contact and affective involvement. Power can range from equal to unequal; contact from frequent to occasional; and involvement from high to low. The combination of these three dimensions allows us to characterize situations as either formal or informal, as displayed in Table 17 (from Eggins, 1994, p. 65).

**Table 17. Formal and informal language situations**

| Informal | Formal |
|---|---|
| equal power | unequal, hierarchic power |
| frequent contact | infrequent, or one off-contact |
| high affective involvement | low affective involvement |

Following these two classifications of language situations, we arrived at the types of text displayed in Figure 5.

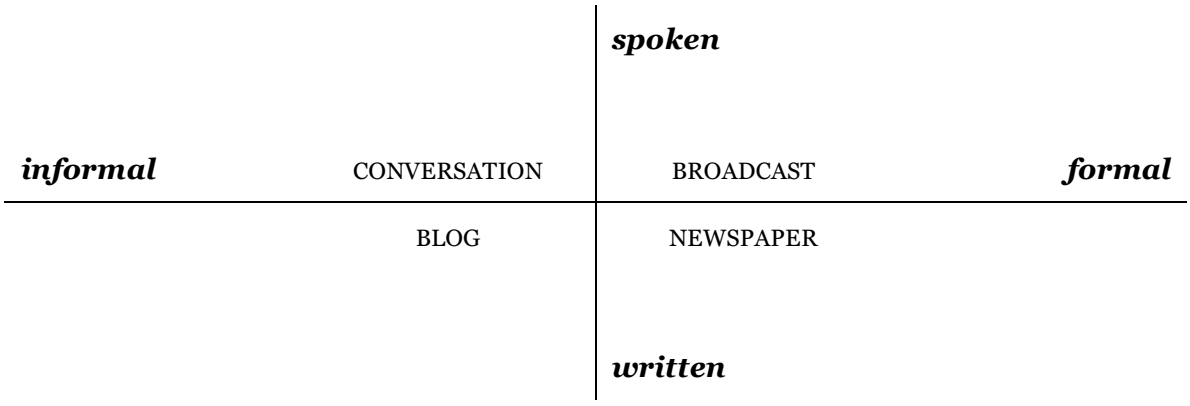


**Figure 5. Genres in our corpus study**

We see casual conversations and newspaper articles at the end of both continua: Conversations are spoken and informal; newspaper articles are written and formal. Blogs are not the prototypical written language situation: They allow for delayed feedback in the form of comments and they are more spontaneous and casual than newspaper articles. Their formality is somewhere between that of conversations and newspaper articles: Participants tend to be in an equal-power relationship and blog writers can assume (sometimes even see) that certain readers frequently read their pieces. Broadcast news are also not the prototypical spoken language situation, since they are not interactive, nor face-to-face, nor

spontaneous, nor casual. In terms of formality, they can be placed at the formal end of the continuum.

Our process of corpora selection also had to accommodate the fact that we are working with two languages, English and Spanish. The choice between parallel or comparable corpora was highly relevant. Hunston (2002, p. 15) explains that comparable corpora have similar designs, that is, they contain similar proportions of text types/categories, whereas parallel corpora contain translations from one language into the other. Given some methodological concerns with parallel corpora, in particular with respect to representativeness and quality of translations, comparable corpora tend to be preferred (Tognini-Bonelli, 2001, p. 7). This has also been our preference. The corpora used in this study are comparable, but not parallel, that is, the Spanish and English corpora contain similar proportions of RCs extracted form conversations, broadcast news, newspaper articles and blogs. The genres and sources are provided in Table 18 below. For this study, five RCs per type (identifying, classifying, narrative, relevance and subjectivity), per genre, per language were selected for analysis, resulting in a total of 200 RCs[39]. Unlike other corpus studies, the focus of our analysis was not on establishing the frequency of the different types of RCs in natural corpora, but to measure the contribution of different types of RCs to textual cohesion so as to determine the best way to segment discourse. For that reason, we chose to keep the number of analyzed tokens constant instead of

[39] It is important to note that not all types of RCs were equally frequent in the corpora. Relative frequencies of RCs per genre per language are provided in Appendix B.

analyzing every RC found in the texts that were extracted from the sources in Table 18.

**Table 18. Genres and sources for this study**

| Language | Genre and source |
|---|---|
| English | Conversation: CALLHOME American English Transcripts (LDC97T14) |
| | Blog: http://desperatehousewivesfan.blogspot.com/, http://new.ca.music.yahoo.com/blogs/realityrocks/ http://greys.wfaa.com/, and http://www.desperateblog.com/ |
| | Broadcast news: 1996 English Broadcast News Transcripts (HUB4) (LDC97T22) |
| | Newspaper: American National Corpus (ANC) Second Release (LDC2005T35) |
| Spanish | Conversation: CALLHOME Spanish Transcripts (LDC96T17) |
| | Blog: http://es.movies.yahoo.com/blog/ |
| | Broadcast news: 1997 Spanish Broadcast News Transcripts (HUB4-NE) (LDC98T29) |
| | Newspaper: TREC Spanish (LDC2000T51) |

The CALLHOME American English Transcripts corpus (Kingsbury, et al., 1997) contains the transcripts of 120 unscripted conversations between English native speakers. Speakers agreed to have their conversations recorded in exchange for a free long-distance phone call. For most speakers, that meant placing a call outside the United States, to family and friends residing abroad. The phone calls were approximately 30 minutes long. Only 10 minutes of each conversation were transcribed. The CALLHOME Spanish Transcripts corpus (Wheatley, 1996) contains the transcripts of 120 unscripted conversations between native speakers of Spanish and was collected in the same manner as the American English conversations.

The blog texts analyzed in our study came from a variety of sources. For English, our sources were blogs in which the authors commented on their favourite TV shows (http://desperatehousewivesfan.blogspot.com/,

http://new.ca.music.yahoo.com/blogs/realityrocks/ http://greys.wfaa.com/, and http://www.desperateblog.com/). For Spanish, the source was a blog in which authors commented on movie releases.

The 1996 English Broadcast News Transcripts (HUB4) corpus (Graff & Alabiso, 1997) contains the transcripts of 104 hours of television and radio broadcasts. The texts for our study were selected from transcripts of CNN World Today broadcasts. The 1997 Spanish Broadcast News Transcripts (HUB4-NE) corpus (Munoz, Alabiso, & Graff, 1998) contains the transcripts of 30 hours of broadcasts news obtained from Televisa, Univision and VOA. The texts for our study were selected from transcripts of Noticias Eco broadcasts (owned by Televisa).

The American National Corpus (ANC) (Reppen, Ide, & Suderman, 2005) contains texts from a variety of genres including transcripts of telephone conversations, travel guides, fundraising texts and newspaper articles, totalling over 20 million words. The texts for our study were selected from newspaper articles that were published in the New York Times. The TREC Spanish corpus (Rogers, 2000) contains texts from newspaper articles obtained from the Mexican newspaper El Norte and from the Agence French Presse. The texts used in our study were selected from the Agence French Presse.

Like any other methodology, a corpus-based study has both its advantages and its disadvantages. According to Chafe (1992, p. 88), the greatest advantages of corpus studies are: (a) Since they are based on overt behaviour, they are verifiable; and (b) since they contain collections of natural language, they are

closer to reality (see Biber, et al., 1998, pp. 169-170, for a similar perspective). Even if linguistic behaviour is an imperfect window to the mind, Chafe (1992, p. 88) reminds us that "[i]t is certainly the best single window available to us." While some may see the fact that some linguistic phenomena are not readily found in corpora as a disadvantage, Chafe (1992, p. 88) argues that "the frequent occurrence or the non-occurrence of some phenomenon is in itself an interesting fact in need of explanation."

A key limitation in corpus studies concerns the hazardous assumptions corpus linguists make when moving from data description to language description (Hunston, 2002; Leech, 2004). Questions regarding the size of the corpus, the diversity of texts in the corpus, the source of the statistical significance of the results, the reliability of the grammatical categories used in the coding and the accuracy of the analysis should be addressed if generalizability is sought. Leech (2004) suggested a few measures to address the first three issues: (a) the use of statistical tests; (b) testing trends across different items; and (c) testing trends across different subcorpora. We have incorporated different subcorpora as well as statistical tests in this study so as to minimize limitations. In addition, a reliability study was performed to test the validity of our coding categories and our coding procedure.

## 5.3 Categories

### 5.3.1 Segmentation

In our discussion of RCs in section 4.3.3, we outlined the segmentation approaches to be tested in our study. Following Kameyama (1998), we will

distinguish between sequential intrasentential Centering and hierarchical intrasentential Centering: Sequential intrasentential Centering results in a flat sequence, that is, "there is always a single centering state, and the output of a complex sentence is the output of the last subsentential unit" (Kameyama, 1998, p. 101). Hierarchical intrasentential Centering results in a tree structure. This means that there will be "multiple [C]entering states simultaneously active at different depths of embedding" (p. 101). As a result, the two approaches to segmentation that treat hypotactic and/or embedded clauses as center-updating units will be further subdivided to accommodate the sequential/hierarchical distinction. This distinction will allow us to assess the contribution of RRCs and NRRCs to the cohesion of a text. In the hierarchical condition, NRRCs in Hypotactic-Clause Centering and RRCs and NRRCs in Embedded-Clause Centering will not be accessible to the following utterance, meaning that the referring expressions in the following utterance will not be allowed to search for an antecedent in the RCs. If, as proposed by Miltsakaki (2003, 2005), RCs do not contribute to textual cohesion, the fact that the entities in RCs are not accessible to the following utterance should not result in less cohesive transitions.

As a result of including the sequential/hierarchical distinction, for each discourse segment containing a RC, we will compute five different realizations of the Centering algorithm. These five segmentation approaches are defined and illustrated in (42) – (46) below. For each segmentation approach, a Centering analysis of the utterance containing the RC as well as of its immediate co-text is provided. For Paratactic-Clause Centering (42), we illustrate the segmentation approach for both a NRRC and a RRC.

**42**

| **Paratactic-Clause Centering** | | | |
|---|---|---|---|
| A center-updating unit is an independent clause or a clause that is in a paratactic relation with another clause. Clauses that are in a hypotactic relation with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded. | | | |
| B: and uh we're going to try for Germany | | | |
| Cf: B+>GERMANY | Cp: B+ | Cb: B+ | Ct: EST-CONTINUE |
| B: If Germany doesn't happen we're going to try for ((Beal)) possibly ((Nollis)) Peterson [$_{RC}$ which is in Colorado] | | | |
| Cf: B+>BEAL> NOLLIS PETERSON> GERMANY> PETERSON> COLORADO | Cp: B+ | Cb: B+ | Ct: CONTINUE |
| A: ooh Colorado would be neat | | | |
| Cf: COLORADO | Cp: COLORADO | Cb: COLORADO | Ct: SMOOTH |
| A: I don't know | | | |
| Cf: A | Cp: A | Cb: A | Ct: CONTINUE |
| A: I'm trying to find an American university [$_{RC}$ that has master's programs] | | | |
| Cf: A> UNIVERSITY> UNIVERSITY> MASTERS | Cp: A | Cb: A | Ct: CONTINUE |
| B: ((breath)) NYU has one for one year ((breath)) | | | |
| Cf: NYU> MASTERS> ONE YEAR | Cp: NYU | Cb: UNIV. | Ct: SMOOTH |

With respect to the NRRC, the independent clauses and *uh we're going to try for Germany and ooh Colorado would be neat* and the clause complex *If Germany doesn't happen we're going to try for Beal possibly Nollis Peterson which is in Colorado* constitute center-updating units. For the RRC, all three

independent clauses (*I don't know, I'm trying to find an American university that has master's programs*, and *NYU has one for one year*) constitute center-updating units.

For the two hypotactic approaches (43 and 44), we illustrate the segmentation of a NRRC.

**43**

| **Sequential Hypotactic-Clause Centering** | | | |
|---|---|---|---|
| A center-updating unit is an independent clause or a clause that is either in a paratactic relation or a hypotactic relation with another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is sequential: The output of one center-updating unit constitutes the input for the next center-updating unit. | | | |
| B: If Germany doesn't happen | | | |
| Cf: GERMANY | Cp: GERMANY | Cb: GERMANY | Ct: SMOOTH |
| we're going to try for ((Beal)) possibly ((Nollis)) Peterson | | | |
| Cf: B+>BEAL> NOLLIS PETERSON | Cp: B+ | Cb: 0 | Ct: NOCB |
| [$_{RC}$ which is in Colorado] | | | |
| Cf: PETERSON> COLORADO | Cp: PETERSON | Cb: PETERSON | Ct: EST-CONTINUE |
| A: ooh Colorado would be neat | | | |
| Cf: COLORADO | Cp: COLORADO | Cb: COLORADO | Ct: SMOOTH |

In Sequential Hypotactic-Clause Centering, the three clauses in the clause complex (a) *If Germany doesn't happen* (b*) we're going to try for Beal possibly Nollis Peterson* (c) *which is in Colorado* become separate center-updating units. In the hierarchical approach, the hypotactic (dependent) clauses *If Germany*

*doesn't happen* and *which is in Colorado* are not accessible for the computation of transitions to the following utterance, hence the NOCB transition to the last utterance in (44).

**44**

| **Hierarchical Hypotactic-Clause Centering** | | | |
|---|---|---|---|
| A center-updating unit is an independent clause or a clause that is either in a paratactic relation or a hypotactic relation with another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is hierarchical: Hypotactic clauses constitute embedded Centering units. Their output is not accessible to the following center-updating unit. The output of their dominant clause is used to compute the transition to the following center-updating unit. | | | |
| B: and uh we're going to try for Germany | | | |
| Cf: B+> GERMANY | Cp: B+ | Cb: B+ | Ct: EST-CONTINUE |
| ⇒ B: If Germany doesn't happen | | | |
| Cf: GERMANY | Cp: GERMANY | Cb: GERMANY | Ct: SMOOTH |
| we're going to try for ((Beal)) possibly ((Nollis)) Peterson | | | |
| Cf: B+>BEAL> NOLLIS PETERSON | Cp: B+ | Cb: B+ | Ct: CONTINUE |
| ⇒ [$_{RC}$ which is in Colorado] | | | |
| Cf: PETERSON> COLORADO | Cp: PETERSON | Cb: PETERSON | Ct: SMOOTH |
| A: ooh Colorado would be neat | | | |
| Cf: COLORADO | Cp: COLORADO | Cb: 0 | Ct: NOCB |

For the two embedded approaches, we illustrate the segmentation of a RRC. In Sequential Embedded-Clause Centering, the embedded clause *that has*

*master's programs* is separated from its matrix clause *I'm trying to find an American university*.

**45**

| **Sequential Embedded-Clause Centering** | | | |
|---|---|---|---|
| A center-updating unit is an independent clause or a clause that is either in a paratactic relation, in a hypotactic relation with another clause or embedded within another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Non-finite embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is sequential: The output of one center-updating unit constitutes the input for the next center-updating unit. | | | |
| A: I'm trying to find an American university | | | |
| Cf: A> UNIVERSITY | Cp: A | Cb: A | Ct: CONTINUE |
| [$_{RC}$ that has master's programs] | | | |
| Cf: UNIVERSITY> MASTERS | Cp: UNIV. | Cb: UNIV. | Ct: SMOOTH |
| B: ((breath)) NYU has one for one year ((breath)) | | | |
| Cf: NYU> MASTERS> ONE YEAR | Cp: NYU | Cb: UNIV. | Ct: CONTINUE |

In Hierarchical Embedded-Clause Centering, the same embedded clause is not accessible for the computation of the transition to the following clause, leading to a second SMOOTH SHIFT transition.

**46**

| **Hierarchical Embedded-Clause Centering** | | | |
|---|---|---|---|
| A center-updating unit is an independent clause or a clause that is either in a paratactic relation, in a hypotactic relation with another clause or embedded within another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Non-finite embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is hierarchical: Hypotactic and embedded clauses constitute embedded Centering units. Their output is not accessible to the next center-updating unit. The output of their dominant clause (for hypotactic clauses) or their matrix clause (embedded clauses) is used to compute the transition to the following center-updating unit. | | | |
| A: I'm trying to find an American university | | | |
| Cf: A> UNIVERSITY | Cp: A | Cb: A | Ct: CONTINUE |
| ⇒ [$_{RC}$ that has master's programs] | | | |
| Cf: UNIVERSITY> MASTERS | Cp: UNIV. | Cb: UNIV. | Ct: SMOOTH |
| B: ((breath)) NYU has one for one year ((breath)) | | | |
| Cf: NYU> MASTERS> ONE YEAR | Cp: NYU | Cb: UNIV. | Ct: SMOOTH |

### 5.3.2 Properties of relative clauses

Each RC in the corpus was coded following the RC taxonomy discussed in section 4.3.2 and reproduced below in Table 19.

**Table 19. A functional taxonomy of relative clauses**

| Type | Function | Description |
|---|---|---|
| Restrictive | Identifying | RC adds information necessary for referent identification. |
| | Classifying | RC creates a new subclass within a reference class. |
| Non-restrictive | Narrative | RC moves the narrative time forward. |
| | Relevance | RC makes an antecedent relevant in context. |
| | Subjectivity | RC expresses opinion. |

RRCs and NRRCs were identified following the *one*-substitution test (see section 4.2.2). Identifying RRCs were those that had a definite head, while classifying RRCs were those with an indefinite head. The types of NRRCs were identified following the tests suggested by Loock (2007, pp. 340, 347, 353). Narrative NRRCs were identified by inserting a temporal adverbial, such as *later* in (47)[40]. In relevance NRRCs, the relativizer could sometimes be replaced by a connector that makes the relation between the two clauses explicit. Other times, as in Example (48), the function of the NRRC was recognizable from the context: the RC *who is a practicing pathologist at Temple Community Hospital* legitimizes Posey's authority to make a statement regarding the feelings of those who have lost a loved one. Finally, subjectivity NRRCs were identified by inserting parentheticals such as *in my opinion* in Example (49).

**47**

(a) So get your fix now becasue you arnt going to be seeing these lovely ladies for awhile, unless you buy the first season on DVD [$_{RC}$ (which you can do threw my site! ;P )]

(b) So get your fix now becasue you arnt going to be seeing these lovely ladies for awhile, unless you buy the first season on DVD [$_{RC}$ (which you can do **later** threw my site! ;P )]

[40] Examples are reproduced verbatim, including typos.

**48**

(a) Posey said Herrera's autopsy services are important because it provides closure for many families that are conflicted with a loved one's death. "We also help allay guilty feelings," said Posey, [$_{RC}$ who is a practicing pathologist at Temple Community Hospital].

**49**

(a) I had a Reuben [$_{RC}$ which was a little bit greasy]

(b) I had a Reuben [$_{RC}$ which, **in my opinion**, was a little bit greasy]

## 5.4 Coding

### 5.4.1 Procedure

Having discussed our coding categories, we can provide an overview of the coding procedure. Our coding procedure consisted of the following steps: (a) text selection; (b) RC selection; (c) functional coding; (d) Centering coding; and finally, (d) evaluation. Texts were selected from the corpora discussed in 5.2 (Table 18). The criterion for selection was the degree to which they conformed to the generic structure usually observed for texts realizing a specific genre. Choosing texts that conform as closely as possible to the prototype also facilitated the second step in our coding procedure namely, RC selection. In order to eliminate discourse segment boundaries as a confounding factor, the RCs selected for analysis were not found at stage boundaries, that is, they were stage-internal. Below we discuss how the stages in the four different genres were identified.

#### 5.4.1.1 Generic structure

Identifying the stages of a genre is one of the six steps Eggins and Slade (1997, pp. 231-5) observed in the process of generic analysis. These six steps are: (a) recognizing a text; (b) defining the social purpose of the text; (c) identifying stages; (d) identifying obligatory and optional stages; (e) devising a structural formula, and (f) analysing the linguistic properties of each stage. Given that our motivation for the inclusion of a generic analysis is to avoid confounding factors in our analysis of the contribution of RCs to cohesion, our focus is on steps (b) and (c) and consequently, steps (a), (d), (e) and (f) fall outside of the scope of our study. Following Taboada's (2004) interpretation of Eggins and Slade (1997), the relevant two steps involve defining the purpose of the text by identifying its overall function and identifying and labelling stages on the basis of functional criteria.

Previous work on genre has provided us with useful guidelines to identify stages in texts realizing genres so diverse as casual conversations, on-line blogs, newspaper articles and televised news broadcasts. In order to identify and label the stages found in casual conversations, we followed Eggins and Slade's (1997) work. In their analysis, Eggins and Slade distinguished between segments of text that constitute chunks and segments of text that constitute chat. Only chunk segments can be analyzed for their generic structure (Eggins & Slade, 1997, p. 231) and they can be recognized in two ways, either by a speaker's control of the conversation floor for an extended period of time or by the presence of predictable stages (p. 231). Eggins and Slade argued that these conversation chunks could be identified as instances of the genres of narrative, anecdote,

recounts, exemplum, observation/comment, opinion and gossip (p. 265). Table 20 provides a description of the purpose of each of these genres as well as their generic structure, summarizing work by Eggins (1997) and references therein (e.g., Horvath & Eggins, 1995; Martin & Rother, 1986; Plum, 1988; Slade, 1995).

**Table 20. Genres in casual conversation (adapted from Eggins & Slade, 1997)**

| Genre | Description and Generic Structure |
|---|---|
| Narrative | One of the participants narrates a story in which the protagonists face a crisis they must resolve. The participant may signal they are about to tell a story (abstract), will then set up the protagonists of the story (orientation), list the actions that lead to the problem (complication), and then present the problem (evaluation) and the actions that resolve it (resolution). He or she may make a point about the text (coda).<br>(Abstract) ^ Orientation ^ Complication ^ Evaluation ^ Resolution ^ (Coda) |
| Anecdote | One of the participants narrates a story in which the protagonist faces a problem, a remarkable event, which, however, has no explicit resolution. The telling of the remarkable event is followed by the expression of some kind of reaction.<br>(Abstract) ^ Orientation ^ Remarkable Event ^ Reaction ^ (Coda) |
| Recount | One of the participants recounts certain events in chronological order. The focus is on the succession of events.<br>(Abstract) ^ Orientation ^ Record of Events ^ Reorientation ^ (Coda) |
| Exemplum | One of the participants tells a story that is intended to be interpreted as a normative example.<br>(Abstract) ^ Orientation ^ Incident ^ Interpretation ^ (Coda) |
| Observation /Comment | One of the participants makes an observation which is followed by a comment. The focus is on factuality (not attitude).<br>Observation ^ Comment |
| Opinion | One of the participants expresses an opinion or judgment on a person, a thing or an event. Other participants react either in agreement or in disagreement. If they react in disagreement, statement of evidence may be exchanged until resolution is reached.<br>Opinion ^ Reaction ^ (Evidence) ^ (Resolution) |
| Gossip | An absent person and his/her behaviour is criticized by one participant and may be defended by another participant.<br>Third person focus ^ Substantiating Behaviour ^ (Probe)/Pejorative Evaluation ^ (Defence) ^ (Response to Defence) ^ (Concession) ^ (Wrap up) |

Note. Parentheses indicate optionality; the caret symbol indicates linear order; / indicates alternatives.

Example (3) introduced in Chapter 2 and reproduced here as Example (50) constitutes a realization of the narrative genre in a CallHome conversation.

**50**

ENG5208 (Conversation)

| | | |
|---|---|---|
| | | **Abstract** |
| 1 | B: | I want to tell you |
| | | **Orientation** |
| 2 | | that when I got to Chicago |
| 3 | B: | My sister Agnes was right there to meet me so that was very nice |
| 4 | A: | oh that's nice |
| 5 | B: | and then she waited until I got on the plane |
| 6 | B: | and then after I got on the plane %um |
| | | **Complication** |
| 7 | B: | there was a mixup because |
| 8 | B: | the people in Chica- I didn't realize this until I got to San Francisco |
| 9 | B: | that the people in Chicago they took the wrong ticket from me |
| 10 | A: | oh |
| 11 | B: | they took the ticket [$_{\text{RC-id}}$ that says from San Francisco to Taipei] ((laugh)) and so when they |
| 12 | A: | oh and then what happened? |
| | | **Evaluation** |
| 13 | B: | when they got to San Francisco |
| 14 | B: | the lady said I could really be charged for a second ticket and I said oh no ((laugh)) |
| 15 | A: | oh |
| | | **Resolution** |
| 16 | B: | but she didn't she didn't charge me she she called Chicago and she recognized the mistake [$_{\text{RC-id}}$ they made] |
| | | **Evaluation** |
| 17 | B: | well that caused great confusion for me because when I got on the plane in Chicago |
| 18 | B: | I had the wrong seat number somebody else had that seat number that's because I was ((laugh)) dealing with a different ticket. |
| 19 | A: | oh my goodness |
| | | **Coda** |
| 20 | B: | so anyhow it all worked out but it was |
| 21 | A: | well I'm glad |
| 22 | B: | a little confusion |

In the abstract (line 1), speaker B prepares the listener for what's going to happen. In the orientation, information about the who, where and when of the

event is provided. The usual or expected sequence of events is provided in temporal sequence (*and then… and then…*). In the complication (lines 7 to 12), there is a shift from temporal sequences to counter-expectations: There was a mix-up, something unexpected that the speaker did not realize until she arrived in San Francisco. The evaluation presents both a crisis (i.e., the speaker could be charged for a second ticket) and expressions of incredulity and disbelief (*oh.. oh no*). This crisis is avoided in the resolution (line 16), which is followed by a stage of evaluation in which B tries to provide an explanation for the crisis and A expresses her disbelief (*oh my goodness*). In the coda (lines 20-22) we find closure (*so anyhow, it all worked out*).

The descriptions and generic structures in Table 20 served as a guideline in the selection of texts and the identification of stages in our analysis of the CallHome conversations. One small addition to the list of possible genres in conversation was allowed to accommodate this special type of conversation, which has been described as update conversation in previous work (Taboada & Wiesemann, in press). During the conversation, the participants update each other on the latest events and future plans in their lives or the lives of others. Obligatory in this updating genre are either a recount of past events and/or an outline of future plans. Optionally, one of the speakers may inquire about the other participant's life in an orientation stage, and the updater may sum up the current events in his or her life in a coda stage. The generic structure of the updating genre is provided in (51), an example in (52) (from CallHome English 4677, (Kingsbury, et al., 1997).

**51**

(Orientation) ^ Recount of past events / Outline of future plans ^ (Coda)

**52**

ENG4677 (Conversation)

**Orientation**

1 B: are you doing anything exciting in the next few days
2 A: yep
3 B: what

**Outline of future plans**

4 A: we're leaving
5 B: ((exhale)) besides that
6 A: we're leaving tomorrow
7 A: I mean we're leaving for Framingham tomorrow [$_{\text{RC-rel}}$ where my cousins live]
8 A: and then we're staying there for the day and then we're leaving the next day
9 B: okay

**Coda**

10 A: like that so nothing much exciting
11 A: except for going out on the sunfish today again
12 B: oh
13 A: just ((faint))
14 B: %um
15 A: yeah

Lines 1 to 3 were identified as the orientation, in which speaker B inquires about A's future plans. In lines 4 to 9, A provides an outline of their future plans, which include visiting some cousins in Framingham. In lines 10 and 11, A makes an evaluation about the future plans (*nothing much exciting*) and in the next few lines the topic is abandoned and the update event completed.

The second genre that concerns us, online blogs, is a relatively new addition to the class of communication media. As a result, literature on their generic structure is scarce. To our knowledge, one of the few studies on the genre of blogs was published by Herring, Scheidt, Wright and Bonus (2005). Herring et

al. (2005, p. 142) define weblogs as "frequently modified web pages in which dated entries are listed in reverse chronological sequence." Their study is a corpus analysis of 203 English text-based blogs in which they focused on the purpose and structural characteristics of blogs as well as on the demographics of blog authors. Based on purpose, Herring et al. (2005, following Blood, 2002) distinguished among personal journals, filters and k-logs. In personal journals, bloggers report on events in their personal lives. In filter blogs, bloggers comment on public events and in k-logs or knowledge-logs, bloggers provide information on a topic or product (Herring, et al., 2005, p. 147). The type of blog that concerns us here, namely, blogs in which television shows or movies are commented and reviewed belongs to the second type of blog, filter blogs.

Regardless of their purpose, blogs share structural characteristics. Herring et al.'s analysis revealed that the blog entry is composed of three elements: (a) entry headers and footers, (b) entry body features, and (c) entry body text (p. 154). Entry headers usually include the date and title of the entry, while entry footers display the time of the posting, the author and a link to a permanent copy of the entry. The header may also include a link to add comments, but as observed by the authors, the presence or absence of certain features commonly associated with blogs, such as comment boxes, depends on the type of software used (Herring, et al., 2005, p. 153). The entry body may consist of text, images and links to other sites. Herring et al.'s study focused on text-based blogs, so the incidence of images in their study is very low as a result of the sampling. Finally, the entry body text tends to be short, containing 210 words in average (p. 155).

Given Herring et al.'s analysis of weblogs, we propose a generic structure for blogs in (53) and illustrate with an example in (54) (downloaded from http://www.beloblog.com/greys/ on March 13, 2009).

**53**

Header ^ Entry Body ^ Footer

**54**

| | |
|---|---|
| | http://www.beloblog.com/greys/ (Blog) |
| | **Header** |
| 1 | Paging Dr. Sloan |
| | **Entry Body** |
| 2 | The man [$_{\text{RC-id}}$ who plays an over-the-top handsome, egotistical, womanizing plastic surgeon] recently underwent a procedure to battle skin cancer. |
| 3 | Eric Dane [$_{\text{RC-rel}}$ who plays Dr. Mark Sloan on Grey's Anatomy] revealed recently to OK! Magazine that he had developed skin cancer above his lips. |
| 4 | He had the growth removed. |
| 5 | Doctors apparently froze it off using liquid nitrogen. |
| 6 | >>OK! Magazine |
| | **Footer** |
| 7 | Posted by Katharyn at 3:11 PM \| Permalink \| Comments (0) |

The header in line 1 contains the title of the entry, while the footer in line 7 contains information about the author of the blog (*Katharyn*) as well as the time the blog was posted (*3:11 PM*). There are also links to the archive and to the comment section. The entry body of the blog in this case consists of text (lines 2 - 5) and of a link to a magazine (*OK! Magazine*).

For this study, RCs were selected from the entry body text, as long as these RCs were stage-internal. Thus our analysis concentrated on the textual component of the blogs. It is important to note however, that a full analysis of the blog genre needs to explore multimodal communication, that is the interaction

between textual and non-textual elements (as in Knox, 2007, for online newspaper articles).

The third genre in our study is newspaper articles that have appeared in print (and not online). The term newspaper articles covers a wide range of text types including but not limited to news stories, expert interviews or feature articles. Our analysis of newspaper articles will be limited to the news story, which, as defined by White (1998, p. 171), are "those news reporting texts which act to represent material events or activity sequences." White notes that the activity sequences reconstructed in news stories tend to be associated with crime, misadventure and warfare. In our data set, they included stories about politics (at the national and international level), economy, general interest and sports in both languages, as well as news stories on warfare and crime in Spanish.

Ungerer (2004, p. 308) tells us that news stories have the principal goal of providing information and are prototypically organized according to a top-down principle of decreasing importance. The prototypical textual elements of a news story are: (a) the headline, which tends to contain the most important information, (b) the lead, which provides a summarized account of the news item, and (c) the body copy, which provides detailed information about the news items consisting of history, consequence and evaluation (p. 308).

White's (1998) proposal for the generic structure of the news story differs slightly from Ungerer's headline-lead-body structure. According to White (1998), the headline and the lead are both components of one stage, the opening phase in the news story genre (p. 185). The headline/lead stage may serve different

functions: It may provide a synopsis, a description of some aspects of the event (p. 187), it may serve as an abstraction, providing a generalization of the event by means of a generalizing label (p. 190), or it may provide both a synopsis and a generalization (p. 190). The second stage in the news story genre is the body of the story where the information presented in the opening headline/lead is further developed. Depending on how the information is further developed, the body of the story may be subdivided in the subcomponents of elaboration (providing detailed information), cause-and-effect (providing causes or reasons), contextualization (setting the event in a temporal, spatial or social context) and appraisal (providing a judgment) (p. 193). These subcomponents do not stand in relation to each other but to the opening headline/lead stage, in accordance with the orbital organizing principle.

Our generic analysis of newspaper articles differed slightly from the two analyses presented here. While we followed White (1998) in distinguishing two stages of generic structure, we included the lead with the body of the story, as shown in (55). Our rationale for separating the lead from the headline is functionally and formally motivated. Functionally, we believe that the lead is similar to the sub-components of the body of the story in that it elaborates the information contained in the headline. Formally, we follow Eggins (2004, p. 65), who advocates that functionally identified stages must be formally recognizable in their (linguistic) realizational patterns. In our opinion, there is a formal difference between the headline and the lead that that is not found between the lead and the body of the story: Headlines tend to be non-finite clauses, as in

Example (56), whereas utterances in the lead and body of the story are finite clauses (or clause complexes).

**55**

Headline ^ Body of the story [lead, elaboration/ cause-and-effect/ contextualization/ appraisal]

These stages are illustrated in Example (56). The headline (line 1) summarizes the news story in eight words: 'Murderer of thee teenagers executed in the US'. The body of the story further elaborates the information in the headline. Line 2 (the lead) presents the points that would be further developed in the text: (a) the execution, (b) the crime and (c) the trial.

**56**

AF940517 (Newspaper)

**Headline**

1 ASESINO DE TRES ADOLESCENTES EJECUTADO EN ESTADOS UNIDOS

**Body of the story**

*Lead*

2 BALTIMORE, EEUU, Mayo 17 (AFP) - Un hombre [$_{RC\text{-}cl}$ que asesinó a tres adolescentes en 1990], sin mostrar signos de arrepentimiento durante su juicio, fue ejecutado este martes con una inyección letal en el estado de Maryland (este de Estados Unidos).

*Elaboration (execution)*

3 John Frederick Thanos, 45 años, dijo "Adios" antes de que le administraran la dosis letal, al transformarse en el primer preso de Maryland ejecutado desde 1961, y en el 240 del país desde que la Corte Suprema autorizó a los Estados a restablecer la pena de muerte.

4 "Aquí viene", dijo Thanos al sentir la entrada en su corriente sanguínea del tóxico fatal, indicaron testigos de la ejecución.

*Elaboration (crime)*

5 Poco después de haber sido liberado de la cárcel por error hace 18 meses, [$_{RC\text{-}rel}$ donde había sido enviado por robo], Thanos asesinó a un joven de 18 años [$_{RC\text{-}cl}$ que hacía autostop], y una semana después a otros dos adolescentes de 16 y 14 años, durante el asalto a una estación de gasolina.

*Elaboration (trial)*

6 Durante su juicio Thanos se negó a reconocer la autoridad del tribunal y declaró al juez que si sus dos jóvenes víctimas pudiesen recuperar la vida, los volvería a asesinar.

7 "Lo hice, fui sentenciado, lo acepto", declaró Thanos el 6 de mayo, negándose a presentar una apelación.

8 Durante el juicio su madre y hermana intentaron probar que era mentalmente irresponsable.

**Headline**

1 Murderer of thee teenagers executed in the US

**Body of the story**

*Lead*

2 Baltimore, USA, May 17th (AFP) – A man who murdered three teenagers in 1990, who showed no signs of regret during his trial, was executed this Tuesday with a lethal injection in the state of Maryland (Eastern United States).

*Elaboration (execution)*

3 John Frederick Thanos, of 45 years of age, bid farewell before receiving the lethal dose and thus becoming the first prisoner of Maryland to be executed since 1961, and the 240th prisoner in the country since the Supreme Court reinstated the death penalty.

4 "Here it comes", said Thanos feeling the lethal toxic entering his blood stream, witnesses to the execution said.

*Elaboration (crime)*

5 Shortly after having been freed from prison by mistake 18 months ago, where he had been sent to for robbery, Thanos murdered an 18-year-old young man who was hitch-hiking, and a week later he murdered another two teenagers of 16 and 14 years of age during a gas station robbery.

*Elaboration (trial)*

6 During the trial Thanos denied the authority of the jury and told the judge that if his two young victims could come back to life, he would kill them again

7 . "I did it, I was sentenced, I accept it", said Thanos on May 6th and refused to appeal the sentence.

8 During the trial, his mother and sister tried to show that he was mentally incompetent.

As was the case with newspaper articles, the term 'news broadcasting' encompasses different types of genres, such as news story reading, news reviewing and news interview (Clayman & Heritage, 2002; Coupland, 2001). The news broadcast texts from which RCs were extracted belong to one particular genre of the news broadcasting: mainstream TV news, which, as we found out in our analysis, is in itself is a hybrid between news story reading and the news interview. This means that it shares generic properties with the news story: It has a headline stage in which the story is introduced and a body stage in which the

story is fully developed. While the headline stage is introduced by the news anchor, the body of the story may consist of news reading by the news anchor or by a reporter, or of a news interview, between the anchor (or a reporter) and an interviewee, and in this way, it shares generic properties with the news interview. As analyzed by Clayman and Heritage (2002)[41], the organizational structure of the news interview consists of 3 phases: an opening phase, an interview phase per se and a closing phase. The main characteristic of the opening phase is that it is not interactional since it consists of a monologue in which the interviewer announces the topic (headline), provides background information (background) and introduces the interviewee (lead-in) (Clayman & Heritage, 2002, pp. 58-65). Depending on the characterization of the interviewee (lead-in), Clayman and Heritage distinguished among three types of interview discourse (interview phase): (a) newsmaker interview, (b) background interview, and (c) debate interview (pp. 68-72). In the newsmaker interview, the interviewee is or has been a participant (or the protagonist) of a news event. In the background interview, the interviewee possesses relevant expertise to weigh in on a news event. And finally, in the debate interview, interviewees have relevant expertise on a news event and take opposing positions in an argument. All three types of interview discourse are characterized by an institutionalized question-answer format in which interviewers ask questions and interviewees respond to them (p. 97). The types of questions will vary according to the type of interview, but the interview phase or stage in a news interview is characterized by sequences of question-

[41] Clayman and Heritage (2002) provide an account of news interview from the perspective of Conversation Analysis. Given that the phases were identified functionally, it is possible for us to draw from their analysis in our characterization of the genre.

answer pairs. The last phase of the news interview identified by Clayman and Heritage (2002) is the closing phase in which the interviewer terminates the interview to fit pre-established time constraints. Clayman and Heritage identify two stages in the closing down phase: (a) a winding down stage in which the interviewer announces the need for termination or makes a general comment on about what has been discussed and the implications of the discussion (pp. 77-8), and (b) the termination stage in which the interviewer thanks the interviewees for their participation (p. 74).

Our analysis of the generic structure of news broadcasting combined the properties of the news story identified by White (1998) and the properties of the news interview identified by Clayman and Heritage (2002):

**Table 21. Stages in the news broadcast genre**

| **Genre** | **Stages** | | |
|---|---|---|---|
| News Story | Headline | Body of the Story (lead, elaboration/ cause-and-effect/ contextualization/ appraisal) | |
| News Interview | | Opening | Interview | Closing |
| News Broadcast | Headline | Story Reading and/or Interview | |

An example of how this generic structure was realized in our data is provided in (57).

**57**

h960514 (Broadcast News)
**Headline**
Speaker="Kathleen_Kennedy"

1 ((breath)) a homes test for the aids virus has received approval from the food and drug administration ((breath))
2 the test offers complete privacy
3 but critics worry about how some people may react to getting bad news over the phone
4 ((breath)) c. n. n. medical correspondent dan rutz has more in tonight's news for ((beep))

medicine

**Story reading**

Speaker="Dan_Rutz"

millions of americans are tested for h. i. v. every year

((breath)) up until now it required a trip to a clinic or blood bank

((breath)) now after years of debate the f. d. a. has approved a new option a kit [RC-cl that can be used in private at home]

Speaker="h960514_F_US_002"

this is a way for people [RC-cl who ((breath)) don't want to or can't ((breath)) go to a clinic or a doctor's office] to get tested

and we think having that choice ((breath)) is an important choice to have

Speaker="Dan_Rutz"

((breath)) a finger prick and a few drops of dried blood are all [RC-cl that takes] ((breath))

the card is then sent to a lab

((breath)) to protect privacy each test carries a code instead of a name

((breath)) buyers phone in that code for the results

((breath)) if the test is normal they get a recording ((breath))

if it's positive for the aids virus ((breath)) they are automatically connected to a counselor

((breath)) Reni vaughn has had to deliver that bad news in person at the aid atlanta anonymous test site

Speaker="Reni_Vaughn"

usually people are calm and just silent tears ((noise))

Speaker="Dan_Rutz"

vaughn says face to face counselling is important at such a time ((breath))

she sees a place for home aids ((noise)) testing

but like critics [RC-cl who've opposed it ((breath)) worries

about those [RC-id who will ((noise)) hear they have a ((noise)) life threatening infection ((noise)) over the phone]

Speaker="Reni_Vaughn"

((breath)) and i'm concerned that if someone hears they're positive over the phone their immediate reaction is to withdraw

and in doing that they hang up the phone

and then they lose contact ((breath))

Speaker="h960514_F_US_002"

we think that studies have been done to show ((breath)) that people can receive this kind of information over the phone

these are trained counsellors

they have ((breath)) good training

we know that they've ((breath)) been able to handle other serious problems like suicide or other kinds of ((breath)) hot-line prevention programs over the phone

Speaker="h960514_F_US_008"

of thi- this female she has not returned for her test results

and react ((r.)) means reactive [RC-rel which is h. i. v. positive] ((noise))

Speaker="Dan_Rutz"

31 many people [$_{RC\text{-}cl}$ who do get tested] never show up for the results

32 and according to one government estimate((breath)) over half of those at greatest risk have never been tested

33 ((breath)) public health officials hope home testing will lead to earlier diagnosis and treatment ((breath)) as well as safer sex practices ((breath)) to keep the virus from spreading

34 the new test called confide will be available in texas and florida next month and the rest of the country by early next year ((breath))

35 the drugstore price will around forty dollars

36 dan rutz c. n. n.

The headline in lines 1 to 4 summarizes the key points in the news story: (a) There is a new HIV test that can be taken from home; (b) this test is private; (c) but critics are worried about how patients with positive tests would react upon hearing the news over the phone. These three points are elaborated in the body of the entry or story reading. Lines 5 to 15 explain how the test works, including privacy issues (lines 11 to 13). Lines 16 to 28 deal with concerns about patients' reactions to this type of news delivered over the phone. The last part of the story reading highlights the possible benefits of the home test.

Once the texts in the corpus were analyzed for their generic structure following the guidelines outlined in this section, stage-internal RCs were selected from those texts. These RCs were coded for their functional properties following the functional taxonomy of RCs provided in Table 18. The functional coding was followed by a Centering analysis that included the five approaches to segmentation identified in section 5.3.1. The approaches to segmentation and the properties of RCs have been previously discussed. Here, we review the general properties of the Centering algorithm.

#### 5.4.1.2 Centering algorithm

To the exception of utterance segmentation, which is being examined here, the application of Centering Theory adopted in this study follows a coding manual specifically composed for this purpose (Appendix A), which in turn follows some of the guidelines proposed in Hadic Zabala and Taboada (2006). In any application of Centering, the key parameters to be specified are, in addition to the notion of utterance, the concept of entity realization, the population and ranking of entities in the list of forward-looking centers (Cf-list) as well as the computation of Centering transitions. These notions were introduced and discussed in Chapter 2. Here we focus on how these parameters were instantiated in our analysis.

The definition of realization proposed by Walker et al. (1998, p. 4) states that "[a]n utterance U realizes a center c if c is an element of the situation described by U, or c is the semantic interpretation of some subpart of U." Walker et al. (1998, p. 4) interpret this definition to include "pronouns, zero pronouns, explicitly realized discourse entities, and those implicitly realized centers that are entities inferable from the discourse situation." The inclusion of implicitly realized centers is of particular importance as research has shown that without the inclusion of indirect realization a considerable proportion of transitions would be null (e.g., Fais, 2004; Poesio, et al., 2004a). Following these findings as well as Walker et al.'s (1998) definition, our interpretation of realization includes both direct and indirect realization of entities (following Taboada & Hadic Zabala, 2008).

The incorporation of indirect realization in the Centering algorithm can vary. One way to do it is to include relationships of lexical cohesion (Halliday & Hasan, 1976) in the Cf-list, as in Fais (2004). Fais accounted for lexical cohesion in her transition types: If an entity of the current Cf is lexically related with an entity in the previous Cf, then the transition between the two utterances was labelled COHESIVE. If there was no such relation between entities in the two utterances, then the transition was labelled COMPLETE SHIFT.

An alternative way follows Prince's (1996) hearer-status instead of Halliday and Hasan's lexical cohesion (e.g, Strube & Hahn, 1999; Walker & Prince, 1996). For instance, Strube and Hahn (1999) included inferable, containing inferable and anchored-brand new entities in their Cf-ranking. In the basic version of the Cf-ranking, these entities are included with discourse-new entities, that is, they are ranked below discourse-old entities. In the extended version of the Cf-ranking, these entities constitute mediated discourse entities and are ranked between discourse-old entities and discourse-new entities.

We adopted the first approach and listed entities in a relationship of lexical cohesion in the Cf-list, following Fais (2004) and Hadic Zabala and Taboada (2006). Lexical cohesion was understood in terms of Halliday and Hasan's (1976) ties of grammatical and lexical cohesion: reference ties, substitution ties, ellipsis ties, synonymy ties, superordinate ties and general word ties (see Hadic Zabala & Taboada, 2006, pp. 12-14, for details). Unlike Fais (2004), who proposed new transition types, we computed standard Centering transitions, albeit with the following modifications: (a) If the current Cb is in a

cohesive relation with the previous Cb, the transition will be CONTINUE or RETAIN (depending on the realization of the Cp); (b) if the current Cb is in a cohesive relation with an entity other than the previous Cb, the transition will be SMOOTH or ROUGH SHIFT (depending on the realization of the Cp).

Once the entities that are realized in an utterance were identified, they were listed in the set of forward-looking centers for that utterance, ranked according to their discourse salience. We mentioned in Section 2.2 that the ranking of entities in the Cf-list is language-specific. Our Cf-ranking in the English data followed the standard method of ranking by grammatical function, whereas in the Spanish data, the ranking proposed by Taboada (2002, 2008) was adopted. The rankings are provided in (58) and (59) below. Entities to the left end of the ranking are said to be more salient than entities to the right end of the ranking.

**58 English Cf Template**

Subject > Indirect Object > Direct Object > Other

**59 Spanish Cf Template**

Experiencer > Subject > Animate Indirect Object > Direct Object > Other > Impersonal pronoun

Taboada (2002) observed that animacy and empathy are salient in Spanish discourse and included them in the ranking of entities in the Cf-list. The inclusion of animacy and empathy is seen in the placement of the experiencer of psychological verbs in the highest position, which relates to its linguistic realization as a clitic before the verb in Spanish sentences (as in Example (60), where the experiencer *me* is ranked higher than the grammatical subject of the

utterance, *el mar*). A later version of the template (Taboada, 2008) included impersonal and arbitrary pronouns towards the end of the ranking, given their low salience.

**60**

Me gusta el mar.

me pleases the sea

'I like the sea'

These Cf-rankings are clause-based. For the segmentation approaches that allow more than one clause, the ranking of entities in the dominant or matrix clause precedes the ranking of entities in the dependent or embedded clause (following Miltsakaki, 2002, 2003). For example, the utterance in (61) consists of a dominant clause (a), and two hypotactic clauses (b) and (c): a) *Justin Chambers ingresó en el ala psiquiátrica del Centro Médico UCLA (Los Ángeles) esta semana*, b) *que encarna al Dr. Alex Karev en Anatomía de Grey,* and c) *donde pasó tres días*. We ranked entities in multi-clausal utterances following the order: paratactic clause > hypotactic clause > embedded clause. If there were more than one hypotactic or embedded clauses, we followed linear order.

**61**

Justin Chambers, que encarna al Dr. Alex Karev en Anatomía de Grey, ingresó en el ala psiquiátrica del Centro Médico UCLA (Los Ángeles) esta semana, donde pasó tres días.

'Justin Chambers, who plays Dr. Alex Karev in Grey's Anatomy, was admitted to the psychiatric ward of the UCLA Medical Centre (Los Angeles) this week, where he spent three days.'

In addition to multi-clausal utterances, utterances with complex noun phrases, in particular entities in possessor-possessed relations, are problematic

for the Cf-ranking. Hadic Zabala and Taboada (2006, p. 17) followed Di Eugenio (1998) in coding possessor entities higher than possessed entities if the possessed is inanimate, and possessed higher than possessor if the possessed is animate. Genitive NPs that consist of multiple NPs, like the one in (62), are not accounted for with this possessor-possessed analysis.

**62**

Estamos cerca de cumplir tres meses con **el asunto de la huelga de guionistas**.

'We are about to mark three months with the subject of the writer's strike.'

Cf: (WRITER+READER)>3MESES>ASUNTO>HUELGA>GUIONISTAS

The NPs *el asunto*, *la huelga* and *guionistas* in the complex NP *el asunto de la huelga de guionistas* do not all stand in a possessor-possessed relation. Whereas the writer's strike is a possessed NP, the subject of the writer's strike is not. Because deciding the relation between two NPs in a complex NP would dramatically increase the complexity of coding (especially when coding reliability is sought and tested), we followed standard practice in Centering and ranked NPs within a complex NP linearly (e.g., Byron & Stent, 1998; Tetreault, 2001; Walker & Prince, 1996).

Once the Cf-list for each utterance was populated and the Cp and Cb were selected, the Centering transitions were computed. In our instantiation of the Centering algorithm, we introduced an adjustment to the traditional set of Centering transitions (which is reproduced in Table 22 below).

**Table 22. Centering transitions**

| | **$Cb(U_{i-1}) = Cb(U_i)$ or $Cb(U_{i-1}) = 0$** | **$Cb(U_{i-1}) \neq Cb(U_i)$** |
|---|---|---|
| **$Cb(U_i) = Cp(U_i)$** | CONTINUE | SMOOTH SHIFT |
| **$Cb(U_i) \neq Cp(U_i)$** | RETAIN | ROUGH SHIFT |

As we can see in Table 22, CONTINUE and RETAIN transitions are computed when the current Cb is the same as the previous Cb, or when there is no previous Cb. Poesio et al. (2004a, p. 315) call the latter type of transition CENTER ESTABLISHMENT (EST), after Kameyama's (1985) work. Kameyama (1985, p. 98) originally introduced what she called the Center-establishment rule to account for the use of unstressed pronouns following segment-initial utterances, which lack a Cb. This is illustrated in Example (63) (adapted from Kameyama, 1985, p. 97).

**63**

Who is Max waiting for? [Cf = Max; Cb = 0]

He is waiting for Rosa. [Cf= Max > Rosa; Cb = Max]

Kameyama (1985, p. 100) explained the use of *he* in terms of the Center-establishment rule: "If one of the Cfs in the previous utterance is made into the Cb of the current utterance, an unstressed pronoun is used." Kameyama distinguished this rule from the Center-retention rule, which stated that "[i]f the Cb of the current utterance is the same as the Cb of the previous utterance, an unstressed pronoun should be used" (Kameyama, 1985, p. 100). Poesio et al. (2004b, p. 55) observed that EST transitions are found in discourse-segment boundaries more frequently than CONTINUE transitions. We followed Poesio et al.

(2004a, 2004b) and adopted the EST transition to model the difference between retention and establishment postulated by Kameyama, and tested empirically by Poesio et al. We propose however, one further modification: We distinguished between EST-CONTINUE and EST-RETAIN transitions, to capture the difference in terms of salience, that is, the realization of the Cp and the Cb. Our Centering transitions are provided in Table 23.

**Table 23. Our Centering transitions**

| | **Cb ($U_{i-1}$) = Cb ($U_i$)** | **Cb ($U_{i-1}$) = o** | **Cb ($U_{i-1}$) ≠ Cb ($U_i$)** |
|---|---|---|---|
| **Cb ($U_i$) = Cp ($U_i$)** | CONTINUE | EST-CONTINUE | SMOOTH SHIFT |
| **Cb ($U_i$) ≠ Cp ($U_i$)** | RETAIN | EST-RETAIN | ROUGH SHIFT |

The inclusion of EST-CONTINUE and EST-RETAIN to our inventory of transitions requires that we modify our current version of Rule 2 of Centering (revising Walker, et al., 1998, p. 4): Transition states are ordered. The CONTINUE transition is preferred to the EST-CONTINUE transition, which is preferred to the RETAIN transition. The RETAIN transition is preferred to the EST-RETAIN transition, which in turn is preferred to the SMOOTH SHIFT transition, which is preferred to the ROUGH-SHIFT transition. Empirically testing the validity of our transition ranking, although highly desirable, extends beyond the scope of our study. We feel confident however, in ranking EST-CONTINUE and EST-RETAIN as preferable to the SHIFT transitions, since the EST- transitions are usually grouped together with the standard CONTINUE and RETAIN transitions. Our ranking of EST- transitions after their CONTINUE and RETAIN counterparts is motivated by the fact that they follow a NOCB transition, which signals a break in the cohesive ties of the text. A

text with no such breaks would be perceived to be more coherent, hence our ranking.

### 5.4.2 Reliability study

In order to test the reliability of our coding categories and our segmentation approaches, a sample of the data was analyzed by the author and a separate coder, following the coding manual provided in Appendix A. Checking the reliability of the coding procedure involved identifying RC function (identifying, classifying, narrative, relevance, subjectivity) and clause type (paratactic, hypotactic or embedded), segmenting clauses according to the different segmentation approaches (Paratactic, Sequential Hypotactic and Sequential Embedded) and finally, identifying subsequent mention of RC entities and co-reference to head or non-head nouns. The results of the reliability study are provided in Table 24.

It is important to note, that only the three segmentation approaches mentioned above (Paratactic, Sequential Hypotactic and Sequential Embedded) were included in the reliability study. The distinction between a sequential segmentation and a hierarchical segmentation concerns the accessibility of discourse segments: In a sequential approach each discourse unit is accessible to the next; in a hierarchical approach, embedded discourse units are not accessible to the next. This distinction in accessibility had no effect in the identification of RC function or of clause type. The segmentation of texts into clauses follows the same procedure for sequential and hierarchical approaches up to the computation of the Centering algorithm, which was not computed in this case

(see explanation below). With regard to the identification of subsequent mentions of RC entities, given that in hierarchical approaches RCs constitute embedded discourse units, entities in RCs are not available for subsequent mention (by definition). Given all of the above, the hierarchical variants of the hypotactic and embedded approaches were excluded from the reliability study.

**Table 24. Agreement in our reliability study**

| Category | Tokens | Agreement (n) | Agreement (%) |
|---|---|---|---|
| RC function | 52 | 46 | 89% |
| Clause type | 111 | 102 | 92% |
| Segmentation | 333 | 331 | 99% |
| Subsequent mention | 60 | 57 | 95% |
| Co-reference | 60 | 54 | 90% |

Disagreements in the identification of RC function were of two types: (a) RRCs were incorrectly identified as NRRCs following an incorrect interpretation of the head noun (n=3); (b) NRRC were correctly identified as non-restrictive, but as a different type of NRRC (n=3). Disagreements in the identification of clause type concerned mostly cases of reported (n=4) and quoted speech (n=3), which, following Halliday and Matthiessen (2004) are identified as clauses in hypotactic (reported) or paratactic (quoted) relations. Agreement in the segmentation approaches was almost perfect, except for two cases involving the first a typing error and the second an incomplete utterance. As for the identification of subsequent mentions, only one example proved problematic for the three segmentation approaches that were tested. It concerns the NP *el macizo* ('the massif') in (64b) and its referential link to other entities in (64a) and (64c).

One coder identified co-reference with *montaña* ('mountain'), whereas the other coder identified co-reference with *región* ('region'). This disagreement meant that for one coder, there was a subsequent mention of the head noun *el macizo*, whereas for the other coder, no subsequent mention of any RC entities was found.

**64**

AF940517 (Newspaper)

(a) Dentro de diez años, el agua dejará de corer por las laderas de la Sierra Nevada de Santa Marta, norte de Colombia, debido a la alarmante deforestación [$_{\text{RC-id}}$ que sufre ese parque natural constituido por la montaña intertropical más alta del mundo a orillas del mar]

'In ten years, water will stop running down the slopes of the Sierra Nevada of Santa Marta, north of Colombia, due to the alarming deforestation of this natural park that has the highest intertropical mountain by the sea in the world.'

(b) El macizo, de 92 km cuadrados, [$_{\text{RC-rel}}$ que tiene dos picos nevados de igual altitud, 5.770 metros -- el Simón Bolívar y el Cristóbal Colón --,] se yergue al borde del mar Caribe, en el departamento de Magdalena, 980 km al norte de Bogotá.

The massif, with a surface of 92 km2, which has two snowy peaks of equal altitude, 5,770 m. – the Simon Bolivar and the Christopher Columbus --, lies by the Caribbean Sea, in the department of Magdalena, 980 km north of Bogota.'

(c) Según la Fundación Prosierra Nevada, entidad no gubernamental [$_{\text{RC-cl}}$ que trabaja en la región desde 1986], de las 2.115.873 hectáreas de bosques primarios, apenas sobreviven 319.561, es decir el 18 por ciento.

'According to the Prosierra Nevada Foundation, a non-governmental organization that has been working in the region since 1986, of the 2,115,873 hectares of forests, only 319,561 still survive, that is 18%.'

Finally, disagreements concerning the identification of the head or non-head status (in the RC) of an entity subsequently mentioned in the discourse concerned this example above in addition to a case of incorrect identification of the head of a RC.

While the reliability of the application of the Centering algorithm was not directly evaluated in this reliability study, it has been evaluated in previous work with Dr. Maite Taboada (Taboada & Hadic Zabala, 2008, pp. 86-88). For that study, the transcripts of one Spanish and one English CallHome conversation were coded by the two authors and compared to the coding done by a third researcher. The study followed the coding manual in Hadic Zabala and Taboada (2006), which served as a guideline for the coding manual in this study. The results of the study showed the following percentages of agreement in unit segmentation (following a segmentation approach similar but not identical to Sequential Hypotactic-Clause Centering) and Cf-ranking.

**Table 25. Reliability of Centering analysis in Taboada and Hadic Zabala (2008)**

| Category | Language | Agreement |
|---|---|---|
| Unit segmentation | English | 91.89% |
| | Spanish | 92.89% |
| Cf-ranking | English | 77.34% |
| | Spanish | 76.13% |

Given the results of our study and the study reported in Taboada and Hadic Zabala (2008), we feel confident in the reliability of our coding categories and our application of the Centering algorithm.

## 5.5 Evaluation

The contribution of RCs to discourse cohesion, and as a result, the adequacy of the different segmentation methods to model this cohesion was evaluated with a theory-internal measure, Constraint 1 and Rule 2 of Centering, and a theory-external measure, the subsequent mention of RC entities.

### 5.5.1 Constraint 1 and Rule 2 of Centering

Following work by Poesio et al. (2000), Poesio et al. (2004a, 2004b) and Taboada and Hadic Zabala (2008), we evaluated the different methods with respect to their violations of Constraint 1 and their adherence to Rule 2 of Centering.

Constraint 1 states that each utterance must have precisely one Cb (the strong version), or that each utterance must have at most one Cb (the weak version). In section 2.2, we defined the Cb as the highest entity in the Cf-list of the previous utterance that is realized in the current utterance and noted that it serves a cohesive function: It links the current utterance to the previous utterance. The absence of a linking entity constitutes a violation of this constraint (the strong version) and results in a NOCB transition, which signals a break down in local cohesion.

Rule 2 states that transition states are ordered: CONTINUE transitions are preferred over RETAIN transitions and these in turn are preferred over SHIFT transitions. Following our modification of Rule 2, our preferred ranking of transition states is:

**65**

CONTINUE > EST-CONTINUE > RETAIN > EST-RETAIN > SMOOTH SHIFT > ROUGH SHIFT

In order to capture violations of Constraint 1 and adherence to Rule 2 quantitatively, we adopted a scoring system similar to Cristea et al.'s (1998) and Ballantyne's (2004) numerical score. In both scoring systems, transitions receive a score between 1 (ROUGH SHIFTS) and 4 (CONTINUE). Cristea et al. (1998) assign a

score of 0 to NOCB transitions. The scoring system used in this study accommodates our modification of Rule 2 and is provided in Table 26 below.

**Table 26. Scoring system**

| Transition | Score |
|---|---|
| CONTINUE | 6 |
| EST-CONTINUE | 5 |
| RETAIN | 4 |
| EST-RETAIN | 3 |
| SMOOTH SHIFT | 2 |
| ROUGH SHIFT | 1 |
| NOCB | 0 |

These scores were computed for two sets of transitions: the transition from a previous utterance to the utterance containing the RC (Pre-RC Degree of Cohesion), and the transition from the utterance containing the RC to the following utterance (Post-RC Degree of Cohesion). In general, the method with the highest numerical score was preferred.

### 5.5.2 Subsequent mention of RC entities

We noted in section 2.2 that Rule 2 of Centering has not been indisputably validated in psycholinguistic experiments or corpus studies. For this reason, we have chosen to complement Rule 2 with a theory-external evaluation measure. Following Miltsakaki (2003), we assess the role of RC entities in textual cohesion by calculating the frequency in which these entities are subsequently mentioned in the discourse (see section 3.2.1, for a discussion). The assumption here is that salient (topical) entities will continue to be the focus of attention in subsequent

utterances. Given that Centering models cohesion from one utterance to the next, we concentrate on subsequent mention in the next Centering utterance. We distinguish between the head noun of the RC and entities in the RC. The head noun belongs to both the RC and its superordinate clause (matrix or dominant clause), and is more topical than other entities in the RC because of this dual membership (Smith, 2003).

In testing the salience of RC entities, we aim to provide a theory-external evaluation of the contribution of RCs to textual cohesion. If entities in RCs are not subsequently mentioned in the discourse, we would conclude that their contribution to textual cohesion is rather minimal. If, on the other hand, RC entities are maintained in subsequent discourse, we would conclude that their role in textual cohesion is of significance. Either conclusion would have an impact on our choice of a segmentation approach.

## 5.6 Statistical analysis

To sum up our methodology so far, we mentioned that a total of 200 RCs were selected for analysis. These 200RCs were equally distributed across RC type, genre and language. Their distribution is provided in Table 27 below.

**Table 27. Distribution of RCs per type, genre and language**

| | RC Type | | | | | |
|---|---|---|---|---|---|---|
| **English** | Identifying | Classifying | Narrative | Relevance | Subjectivity | Total |
| Conversation | 5 | 5 | 5 | 5 | 5 | 25 |
| Blog | 5 | 5 | 5 | 5 | 5 | 25 |
| Newspaper | 5 | 5 | 5 | 5 | 5 | 25 |
| Broadcast | 5 | 5 | 5 | 5 | 5 | 25 |
| Total | 20 | 20 | 20 | 20 | 20 | 100 |
| **Spanish** | | | | | | |
| Conversation | 5 | 5 | 5 | 5 | 5 | 25 |
| Blog | 5 | 5 | 5 | 5 | 5 | 25 |
| Newspaper | 5 | 5 | 5 | 5 | 5 | 25 |
| Broadcast | 5 | 5 | 5 | 5 | 5 | 25 |
| Total | 20 | 20 | 20 | 20 | 20 | 100 |

The 200 RCs and their immediate environment were analyzed for the five different segmentation approaches outlined in 5.3.1, namely, Paratactic-Clause Centering, Sequential Hypotactic-Clause Centering, Hierarchical Hypotactic-Clause Centering, Sequential Embedded-Clause Centering and Hierarchical Embedded-Clause Centering. The adequacy of the segmentation approaches was evaluated with the measures of cohesion discussed in 5.5.1 and 5.5.2. We used a scoring system (Table 26) to compute violations of Constraint 1 and Rule 2 of Centering in order to better assess the contribution of RCs to textual cohesion. We measured this Degree of Cohesion for both utterances containing the RC and utterances following the RC. We labelled these measures Pre-RC Degree of Cohesion and Post-RC Degree of Cohesion respectively. We used a simple scoring

system to compute the presence (value=1) or absence (value=0) of subsequent mentions.

An analysis of variance (ANOVA) was then conducted on these data. ANOVA is a test of statistical significance used to study the effects of two or more treatment variables. Specifically, it tests for significant differences in the mean values obtained from several populations (Woods, Fletcher, & Hughes, 1986, p. 194). In this study, we wanted to test for significant differences in the mean values of Pre-RC Degree of Cohesion, Post-RC Degree of Cohesion and Subsequent Mention obtained for RCs of five different types, selected from four different genres in two languages that were analyzed with five different segmentation approaches. For this purpose, a factorial-mixed design ANOVA was applied, with Degree of Cohesion (Pre-RC and Post-RC) and Subsequent Mention as dependent variables and segmentation approach as within-subject factor and language, genre and RC type as (independent) between-subject factor variables. The factors in a factorial ANOVA are the different criterion variables, and each factor has several levels (Woods, et al., 1986, p. 203). In our study, the factors and factor levels (in parenthesis) were as follows: segmentation approach (Paratactic, Sequential-Hypotactic, Hierarchical-Hypotactic, Sequential-Embedded, Hierarchical-Embedded), language (English, Spanish), genre (conversation, broadcast news, blog, newspaper article) and RC type (identifying, classifying, narrative, relevance subjectivity). The ANOVA had a mixed design, given that the tests for segmentation approach were carried out within subjects whereas the tests for the other factors (language, genre and RC type) were carried

out between subjects. In other words, the evaluation of segmentation approach required repeated measures of the same RCs.

We tested for main effects (of language, genre, RC type and segmentation approach) as well as interaction effects (combinations of the above). Main effects show differences across levels of a factor; these differences are independent of other factors (Woods, et al., 1986, p. 205). Interaction effects show differences that appear when two or more factors are examined: "An interaction between two factors is said to exist if the mean differences among levels of factor *A* are *not* constant across levels (categories) of factor *B*." (Glass & Hopkins, 1984, p. 403)

Global tests (tests of within- and between-subjects effects across all factor levels) were performed first, as they allow us to identify which differences were significant. The alpha level or *p* level was set at .05. Significant results in the global tests indicate that the mean values for the levels of a factor are not equal (Glass & Hopkins, 1984, p. 434). When main effects were observed, given the number of levels in the factors of our study, multiple comparisons (post-hoc tests) were required to identify which levels were associated with the main effect (Glass & Hopkins, 1984, p. 434). In other words, post-hoc tests were only performed and are only reported if the main effect was significant. Where interaction effects were observed, marginal means were plotted to facilitate the interpretation of the interaction. Interaction effects are included in the discussion, as they influence the interpretation of main effects if not taken into account (Glass & Hopkins, 1984, p. 408; Hatch & Lazaraton, 1991, p.380). Note

that reported estimated marginal means are corrected for all possible interaction effects.

## 5.7 Summary

As mentioned throughout this dissertation, the goal of this study is to establish the best approach to RC segmentation based on the contribution of RCs to textual cohesion. For this purpose, a corpus of 200 RCs was collected following the procedures described in this chapter: We selected RCs of five different functional types in texts that realized four different genres obtained from two comparable corpora in English and Spanish. In order to assess the contribution of RCs to cohesion we drew from Centering Theory. In this chapter, we have reviewed the specific coding procedures used in our application of the Centering algorithm, we have outlined and illustrated the five segmentation approaches to be evaluated and we have introduced the measures used to evaluate them. Given the factors and levels that have to be taken into account, we conducted a factorial mixed-design ANOVA to identify significant differences in Degree of Cohesion scores and Subsequent Mention scores between languages, among genres, among RC types and among segmentation approaches.

The results of this factorial mixed-design ANOVA, which we present in the next chapter, will allow us to answer the research questions we formulated in the Introduction. Our first research question inquired rather generally about the contribution of RCs to textual cohesion:

> Q1: What is the contribution of RCs to textual cohesion, which is understood as the textual connectivity that results from co-text dependent nominal interpretation?

It was divided into two sub-questions. Q1a asked whether the segmentation of discourse in different approaches results in different scores of Pre-RC and Post-RC Degree of Cohesion and whether these scores vary depending on RC type:

> Q1a: If RCs are treated as separate utterances, is the model of discourse that results from a Centering analysis more or less cohesive? Do we get more or less violations of Constraint 1 and Rule 2 of Centering? Does the Centering algorithm give us a model of discourse that is more cohesive (or less cohesive) for some types of RCs than other types of RCs?

Q1b asked first whether RC entities are subsequently mentioned and whether those subsequent mentions co-refer with head or non-head RC entities. In addition, it asked whether scores of Subsequent Mention differ across RC type:

> Q1b: Are entities in RCs mentioned in subsequent discourse? Are subsequent mentions co-referential with the antecedent of the RC or with an entity inside the RC? Are there differences among the different types of RCs, so that entities in some types of RCs are more likely to be subsequently mentioned than entities in other types of RCs?

Finally, Q2 asks whether the results obtained for Pre-RC and Post-RC Degree of Cohesion as well as Subsequent Mention allow us to identify the best way to segment RCs in discourse.

> Q2: Given the answers to Q1a and Q1b, what is the approach to discourse segmentation that best captures the functional properties of RCs?

We return to these questions in the discussion of the statistical results of the study.

# CHAPTER 6: RESULTS

## 6.1 Overview of results

In this chapter, we report on the results of the factorial mixed-design ANOVA that was performed on our data. Specifically, we report tests of within-subjects effects for segmentation approach and all interaction with language, type and genre, as well as tests of between-subjects effects for language, type and genre and their interactions. The results of these tests are provided in Table 28 below. Significant results are marked in bold.

**Table 28. Test of between- and within-subjects effects**

| | Pre-RC Degree of Cohesion | | | Post-RC Degree of Cohesion | | | Subsequent Mention | | |
|---|---|---|---|---|---|---|---|---|---|
| Source | Df | F | Sig. | df | F | Sig. | df | F | Sig. |
| Between-subjects effects | | | | | | | | | |
| Language (L) | 1 | 0.2 | .658 | 1 | 1.2 | .267 | 1 | 1.0 | .310 |
| Genre (G) | 3 | 3.6 | **.015** | 3 | 0.5 | .687 | 3 | 2.6 | .051 |
| RCType (T) | 4 | 1.3 | .292 | 4 | 0.9 | .474 | 4 | 0.7 | .561 |
| L×G | 3 | 0.4 | .746 | 3 | 0.1 | .980 | 3 | 2.0 | .111 |
| L×T | 4 | 1.5 | .219 | 4 | 1.0 | .426 | 4 | 0.4 | .822 |
| G×T | 12 | 1.2 | .288 | 12 | 1.2 | .288 | 12 | 1.1 | .351 |
| L×G×T | 12 | 1.0 | .475 | 12 | 0.7 | .724 | 12 | 1.0 | .425 |
| Within-subjects effects | | | | | | | | | |
| Segmentation (S) | 4 | 5.7 | **<.001** | 4 | 5.5 | **<.001** | 4 | 154.7 | **<.001** |
| S×L | 4 | 0.8 | .556 | 4 | 1.3 | .257 | 4 | 0.9 | .468 |
| S×G | 12 | 0.5 | .893 | 12 | 1.7 | .069 | 12 | 1.5 | .116 |
| S×T | 16 | 0.9 | .515 | 16 | 0.4 | .975 | 16 | 7.0 | **<.001** |
| S×L×G | 12 | 1.0 | .415 | 12 | 1.5 | .117 | 12 | 1.3 | .234 |
| S×L×T | 16 | 1.1 | .335 | 16 | 0.9 | .523 | 16 | 1.0 | .497 |
| S×G×T | 48 | 1.3 | .083 | 48 | 0.8 | .763 | 48 | 1.3 | .094 |
| S×L×G×T | 48 | 1.0 | .496 | 48 | 1.2 | .173 | 48 | 1.1 | .290 |

The results of the tests of within-subjects effects on the bottom half of Table 28 show a significant main effect of segmentation approach for all three

measures of cohesion. In addition to this significant main effect, a significant interaction effect between segmentation approach and RC type was observed, albeit only for the dependent variable Subsequent Mention. As for the results of the tests of between-subjects effects, genre was found to have a significant main effect for the dependent variable Post-RC Degree of Cohesion and an almost significant main effect for Subsequent Mention. No other main effects or interaction effects were significant.

We evaluate all main effects and all significant interaction effects in more detail in the following sections. We report effects by factor. For each factor, we provide estimated marginal means and 95% confidence intervals (CI). Estimated marginal means are corrected values, that is, they are mean values that have been corrected for differences due to other variables in order to eliminate confounding[42]. 95% confidence intervals are estimations of the mean value of the population (Griffiths, Stirling, & Weldon, 1998, p. 295)[43]. For those factors for which a significant main effect was observed, we report on the results of post-hoc multiple comparisons tests performed to identify the levels that were associated with the main effect.

[42] "Marginal Means displays estimated marginal means of the dependent variable in the cells (with covariates held at their overall mean value) and their standard errors of the means for the specified factors. These means are predicted, not observed means. The estimated marginal means are calculated by using a modified definition by Searle, Seed and Milliken (1980)." (SPSS, 1998).

[43] In other words, if we were to collect similar samples (same n) of data again and again, 95% of the confidence intervals obtained would contain the actual population mean.

## 6.2 Segmentation approach

Table 29 shows the estimated marginal mean values for Pre-RC Degree of Cohesion, Post-RC Degree of Cohesion and Subsequent Mention for RCs analyzed with five different segmentation approaches. These results show a clear preference for utterances segmented according to Paratactic-Clause Centering to obtain higher scores in terms of Degree of Cohesion (Pre-RC and Post-RC) in comparison to utterances segmented according to Sequential Hypotactic-Clause Centering and Hierarchical Hypotactic-Clause Centering. These last two segmentation approaches in turn, show a tendency to obtain higher scores than utterances segmented according to Sequential Embedded-Clause Centering and Hierarchical Embedded-Clause Centering. As for the presence of subsequent mentions, the sequential approaches (Paratactic, Sequential Hypotactic and Sequential Embedded) show higher mean scores of Subsequent Mentions than the hierarchical approaches.

**Table 29. Estimated marginal means by segmentation approach**

| | Dependent Variable | | |
|---|---|---|---|
| | Pre-RC Degree of Cohesion | Post-RC Degree of Cohesion | Subsequent Mention |
| Segmentation | Mean (95% CI)* | Mean (95% CI)* | Mean (95% CI)* |
| Paratactic | 3.42 (3.12 – 3.72) | 3.09 (2.77 – 3.40) | 0.55 (0.48 – 0.61) |
| Sequential-Hypotactic | 3.08 (2.75 – 3.41) | 2.90 (2.53 – 3.26) | 0.54 (0.47 – 0.60) |
| Hierarchical-Hypotactic | 3.08 (2.74 – 3.41) | 2.76 (2.42 – 3.09) | 0.18 (0.13 – 0.22) |
| Sequential-Embedded | 2.84 (2.49 – 3.19) | 2.39 (2.03 – 2.76) | 0.50 (0.43 – 0.57) |
| Hierarchical-Embedded | 2.88 (2.55 – 3.21) | 2.58 (2.24 – 2.92) | 0.00 (0.00 – 0.00) |

As mentioned before, ANOVA tests of within-subjects effects showed these differences in segmentation approach to be statistically different for all three variables: Pre-RC Degree of Cohesion ($F=5.7$, $df=4$, $p<.001$), Post-RC Degree of Cohesion ($F=5.5$, $df=4$, $p<.001$), and Subsequent Mention ($F=154.7$, $df=4$, $p<.001$). The estimated marginal means of Pre-RC Degree of Cohesion and Post-RC Degree of Cohesion for the five segmentation approaches are plotted in Figure 6 to illustrate the discussion.

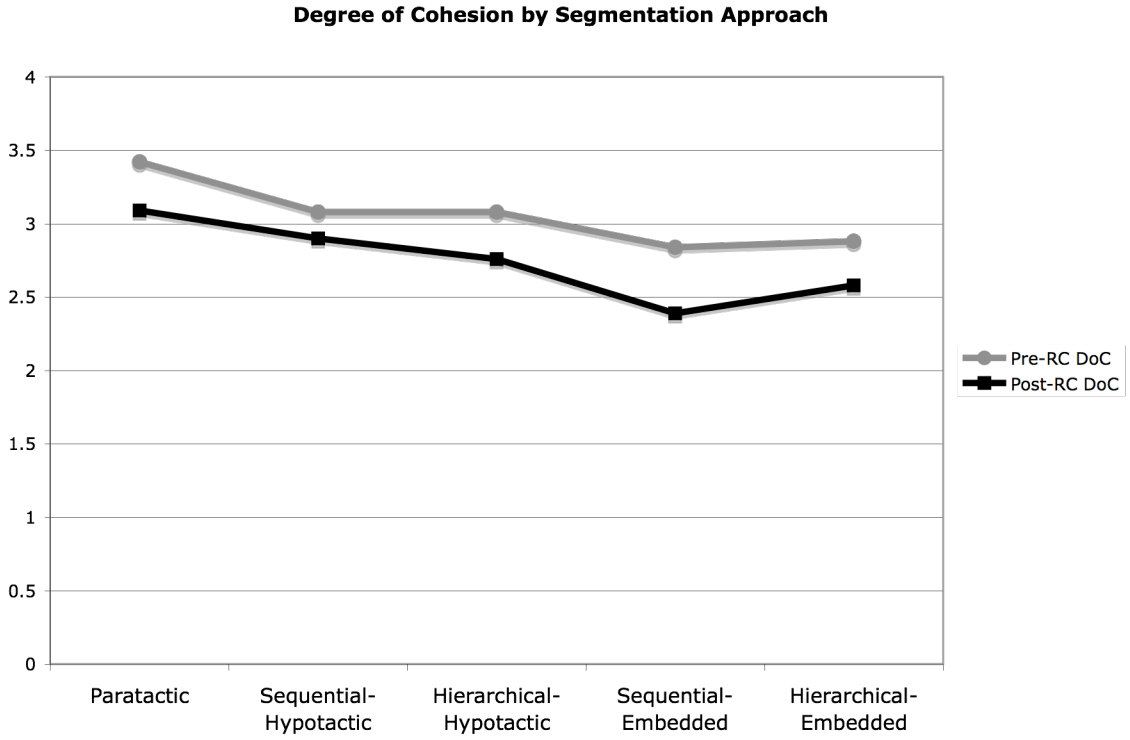


**Figure 6. Plot of estimated marginal means of Pre- and Post-RC Degree of Cohesion by segmentation approach**

Post-hoc multiple comparisons tests showed significant differences in Pre-RC Degree of Cohesion: (a) between Paratactic-Clause Centering and all other approaches; (b) between Sequential Hypotactic-Clause Centering and Sequential Embedded-Clause Centering ($p<.05$) and (c) between Hierarchical Hypotactic-Clause Centering and Hierarchical Embedded-Clause Centering ($p<.05$). In other words, the transitions to the utterance containing the RC when the segmentation of utterances followed Paratactic-Clause Centering were more cohesive than the transitions obtained with Sequential Hypotactic-Clause Centering ($p<.01$), Hierarchical Hypotactic-Clause Centering ($p<.001$), Sequential Embedded-Clause Centering ($p<.01$) and Hierarchical Embedded-Clause Centering ($p<.001$).

Transitions to the utterance containing the RC when the segmentation of utterances followed Sequential Hypotactic-Clause Centering were more cohesive than the transitions obtained with Sequential Embedded-Clause Centering. And finally, transitions to the utterance containing the RC when the segmentation of utterances followed Hierarchical Hypotactic-Clause Centering were more cohesive than transitions obtained with Hierarchical Embedded-Clause Centering. Differences between the two hypotactic approaches or between the two embedded approaches were not significant.

Similar (but not identical) findings emerged for the dependent variable Post-RC Degree of Cohesion and are displayed in Figure 6. Post-hoc multiple comparisons tests revealed significant differences in Post-RC Degree of Cohesion: (a) between Paratactic-Clause Centering and Hierarchical Hypotactic-Clause Centering ($p<.01$), Sequential Embedded-Clause Centering ($p<.01$) and Hierarchical Embedded-Clause Centering ($p<.001$); (b) between Sequential Hypotactic-Clause Centering and Sequential Embedded-Clause Centering ($p<.01$) and (c) between Hierarchical Hypotactic-Clause Centering and Hierarchical Embedded-Clause Centering ($p<.05$). In other words, the transitions to the utterance following the RC were more cohesive when the segmentation of utterances followed Paratactic-Clause Centering than when it followed Hierarchical Hypotactic-Clause Centering, Sequential Embedded-Clause Centering and Hierarchical Embedded-Clause Centering. Transitions to the utterance following the RC were more cohesive when the segmentation of utterances followed Sequential Hypotactic-Clause Centering than when it followed Sequential Embedded-Clause Centering. And finally, transitions to the

utterance containing the RC were more cohesive when the segmentation of utterances followed Hierarchical Hypotactic-Clause Centering than when it followed Hierarchical Embedded-Clause Centering. Differences between the two hypotactic approaches or between the two embedded approaches were once again not significant. Interestingly, mean values of Post-RC Degree of Cohesion in Paratactic-Clause Centering were not found to be statistically different from mean values of Post-RC Degree of Cohesion in Sequential Hypotactic-Clause Centering.

Finally, for the dependent variable Subsequent Mention post-hoc multiple comparisons tests showed significant differences in the presence of subsequent mentions: (a) between Hierarchical Embedded-Clause Centering and all other segmentation approaches ($p<.001$); (b) between Sequential Embedded-Clause Centering and both hypotactic approaches (Sequential Hypotactic-Clause Centering, $p<.05$, Hierarchical Hypotactic-Clause Centering, $p<.001$); and (c) between Hierarchical Hypotactic-Clause Centering and Sequential Hypotactic-Clause Centering ($p<.001$) as well as Paratactic-Clause Centering ($p<.001$). In other words, no significant differences were observed between Paratactic-Clause and Sequential Hypotactic-Clause Centering or between Paratactic-Clause and Sequential Embedded-Clause Centering[44]. As shown in Figure (7), the presence of subsequent mentions of RC entities was significantly lower when utterances were

[44] Sequential Embedded-Clause Centering is almost significantly different from Paratactic-Clause Centering ($p=.06$) and barely significantly different from Sequential Hypotactic-Clause Centering ($p=.03$). Given the values in the table of estimated marginal means, these three approaches should not be significantly different for the measure Subsequent Mention. The differences in p value are due to a smaller Std Error for Sequential Embedded-Clause Centering vs. Sequential Hypotactic-Clause Centering.

segmented following a hierarchical segmentation approach and the lowest when utterances were segmented following Hierarchical Embedded-Clause Centering.

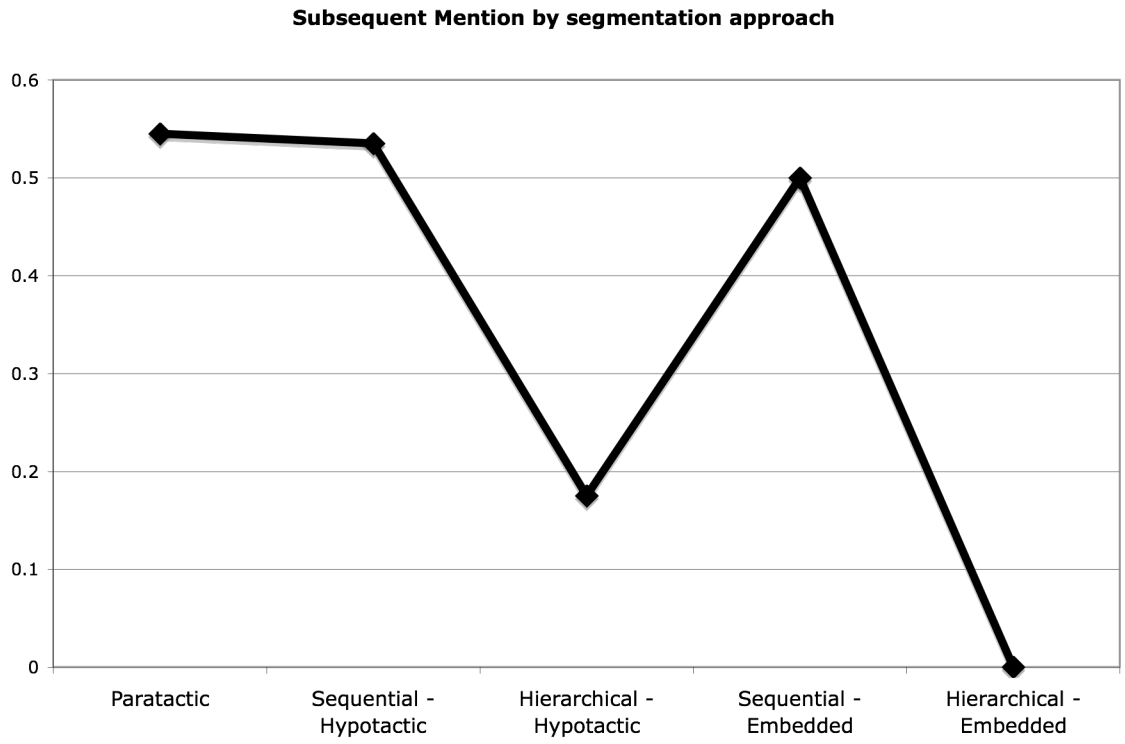


**Figure 7. Plot of estimated marginal means of Subsequent Mention by segmentation approach**

For the dependent variable Subsequent Mention however, an interaction effect for segmentation approach and RC type was observed ($F=7.0$, $df=16$, $p<.001$). The plot of estimated marginal means for segmentation approach and RC type (provided in Figure 8) shows that the main effect of segmentation approach whereby sequential approaches were found to be more cohesive than hierarchical approaches does not hold for identifying and classifying RCs in the Hierarchical Hypotactic-Clause Centering approach.

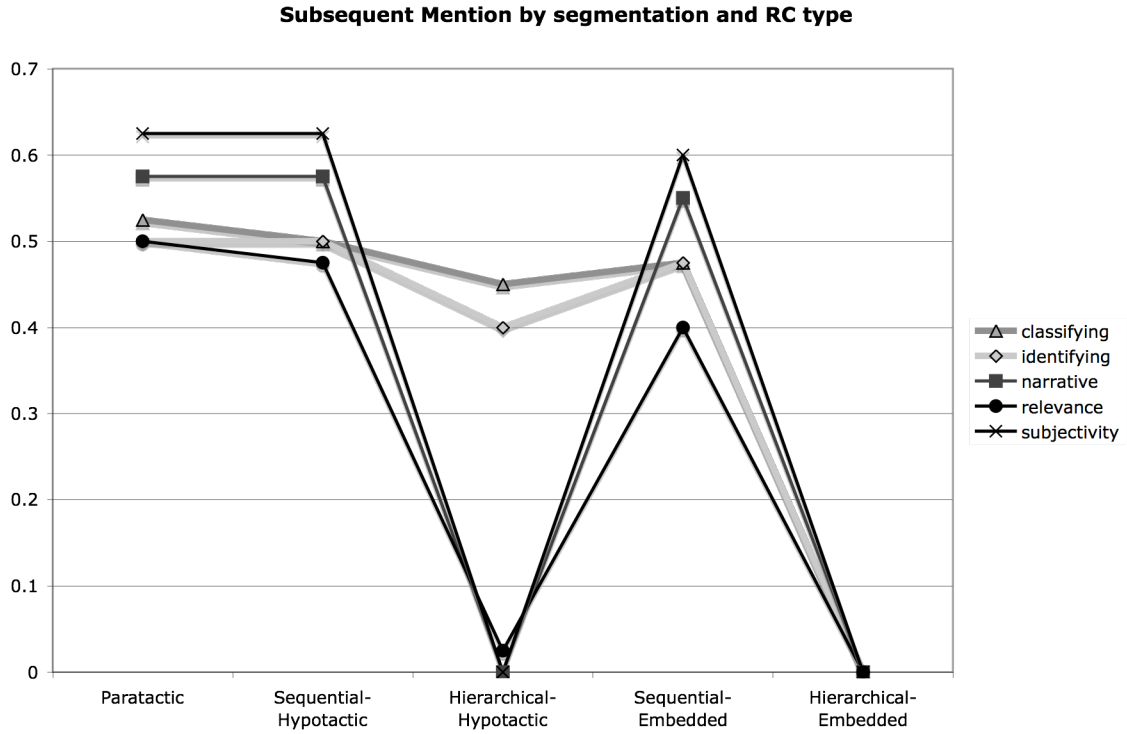


**Figure 8. Plot of estimated marginal means of Subsequent Mention for segmentation approach and RC type**

This interaction is not unexpected: Identifying and classifying RCs are embedded clauses and by definition, they are processed with the matrix clause in this segmentation approach. An account is provided in 6.2.1 below. We return to the discussion of the ANOVA results in 6.3.

### 6.2.1 Subsequent mention of head or non-head RC entities

The dependent measure Subsequent Mention assesses the role of RC entities in textual cohesion by calculating the frequency in which these entities are subsequently mentioned in the discourse. The rationale behind this measure of cohesion (as discussed in section 5.5.2) was the assumption that salient (topical) entities will continue to be the focus of attention in subsequent utterances. For

each RC in Table 27, we computed the presence and the type (head/non-head) of subsequent mention in all five segmentation approaches. These computations yielded a total of 351 RCs with subsequent mentions, out of a possible 1000 RCs with subsequent mention (i.e., 100 RCs per segmentation approach (x5) per language (x2)). Table 30 provides the distribution of subsequent mentions by language and segmentation approach.

**Table 30. Type of subsequent mention by language and segmentation approach**

| Language | Segmentation approach | RCs with subsequent mention | Subsequent mention of Head | Subsequent mention of Non-head | Total subsequent mentions |
|---|---|---|---|---|---|
| English | Paratactic | 59 | 39 | 32 | 71 |
| | Sequential-Hypotactic | 55 | 36 | 29 | 65 |
| | Hierarchical-Hypotactic | 19 | 13 | 9 | 22 |
| | Sequential-Embedded | 54 | 33 | 30 | 63 |
| | Hierarchical-Embedded | 0 | 0 | 0 | 0 |
| | Subtotal | 187 | 121 | 100 | 221 |
| Spanish | Paratactic | 50 | 33 | 24 | 57 |
| | Sequential-Hypotactic | 52 | 36 | 22 | 58 |
| | Hierarchical-Hypotactic | 16 | 12 | 6 | 18 |
| | Sequential-Embedded | 46 | 33 | 17 | 50 |
| | Hierarchical-Embedded | 0 | 0 | 0 | 0 |
| | Subtotal | 164 | 114 | 69 | 183 |
| Total | | 351 | 235 | 169 | 404 |

Note that the total number of subsequent mentions exceeds 351, as subsequent mentions of both head and non-head RC entities were observed in some utterances. The presence of a subsequent mention was already measured with the dependent variable Subsequent Mention in the ANOVA results above,

and was found to distinguish between hierarchical and sequential approaches. As mentioned before, this is not unexpected, given that by definition, entities in embedded segments are not accessible to following discourse. It also accounts for the lower frequency of subsequent mentions in the hierarchical approaches (e.g., 0 – 19), shown in Table 30. Given that entities in embedded segments are not accessible to following discourse, referring expressions in a following utterance cannot resolve to an antecedent in the RC, when the RC constitutes an embedded discourse segment in hierarchical Centering (see section 5.3.1). This means that no RCs entities were accessible to the following utterance in the Hierarchical-Embedded segmentation approach. In the Hierarchical-Hypotactic segmentation approach, only NRRCs constitute embedded discourse segments. Referring expressions in utterances following RRCs (identifying and classifying) were able to search for antecedents in the RC. This restriction explains the lower number of subsequent mentions obtained for the RCs in the Hierarchical-Hypotactic approach in relation to the sequential approaches.

Having explained the lower incidence of subsequent mentions for the hierarchical approaches, we can now turn to the distribution of subsequent mentions across the sequential approaches. Of a possible 100 RCs with a subsequent mention by segmentation approach, approximately 50% of those RCs had at least one, if not two entities mentioned in subsequent discourse in all three sequential approaches in both languages. In English, subsequent mentions ranged from 54 (out of 100) in Sequential Embedded-Clause Centering and 55 (out of 100) in Sequential Hypotactic-Clause Centering to 59 (out of 100) in Paratactic-Clause Centering. In Spanish, the range was 46 (out of 100) in

Sequential Embedded-Clause Centering, to 50 (out of 100) in Paratactic-Clause Centering and 52 (out of 100) in Sequential Hypotactic-Clause Centering. In other words, about half of the RCs in our data had entities that continued to be mentioned in the discourse. Moreover, these entities were salient enough that when the discourse was segmented into smaller units, the cohesive link was not lost (as no significant differences were observed between Paratactic-Clause and Sequential Hypotactic-Clause Centering or between Paratactic-Clause and Sequential Embedded-Clause Centering, with respect to estimated marginal mean values for Subsequent Mention, see discussion below Table 29).

In our discussion of Subsequent Mention as a measure of cohesion, we further differentiated between subsequent mentions of head entities versus subsequent mentions of non-head RC entities. Crucially, we referred to the difference in topicality between these two types of entities, as discussed by Smith (2003): Because the head noun belongs to both the RC and its superordinate clause (matrix or dominant clause), it is more topical than other entities in the RC. Table 30 also shows the frequency of subsequent mentions by type of RC entity, namely, RC heads and RC non-head entities. Regarding the type of entity that gets subsequently mentioned, just over half of the subsequent mentions observed in English (121 out of 221) and two thirds of the subsequent mentions observed in Spanish (114 out of 183) co-referred with the head noun of the RC. This is expected, given the more topical status of RC heads. The observation that almost half of subsequent mentions co-refer with non-head entities is unexpected and it suggests that the contribution of RCs to textual cohesion is not only limited to the head of the RC.

### 6.2.2 Interpreting statistical differences in measures of cohesion

The results of the factorial mixed-design ANOVA reported above show that breaking the discourse into different types of units of analysis yields statistically significant differences between the three measures of cohesion: Pre-RC Degree of Cohesion, Post-RC Degree of Cohesion and Subsequent Mention. We have accounted for the statistical differences in Subsequent Mention between sequential and hierarchical segmentation approaches in 6.2.1 above. Here we correlate the ANOVA findings with frequency counts performed on the data. Figure 9 displays the number of times a segmentation approach outperformed the other segmentation approaches in the three measures of cohesion. For each RC (N=200), we compared the performance of the five segmentation approaches in the three measures and identified the winner approach (or approaches, if several approaches obtained the same score).

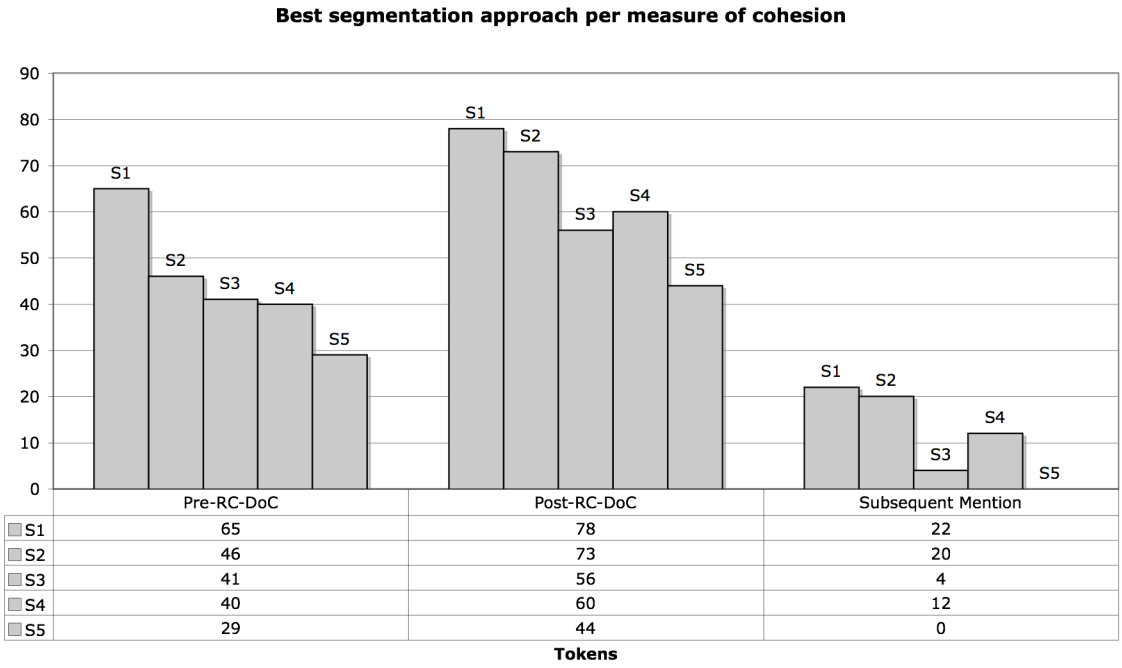


**Figure 9. Distribution of best segmentation approach by measure of cohesion** [45]

The results presented in Figure 9 show the same patterns revealed by the ANOVA calculations. For the measure Pre-RC Degree of Cohesion, Paratactic-Clause Centering (S1) was more often identified as the best segmentation approach (65 out of 200 RCs). For the measure Post-RC Degree of Cohesion, Paratactic-Clause Centering (S1) and Sequential Hypotactic-Clause Centering (S2) outperformed the remaining segmentation approaches: They more frequently obtained the best scores in transitions to the utterance following the RC (in 78 and 73 RCs, respectively). Finally, for the measure Subsequent Mention, the sequential approaches Paratactic-Clause Centering (S1), Sequential Hypotactic-Clause Centering (S2) and to a lesser extent, Sequential Embedded-

[45] Note: S1 corresponds to Paratactic-Clause Centering, S2 to Sequential Hypotactic-Clause Centering, S3 to Hierarchical Hypotactic-Clause Centering, S4 to Sequential Embedded-Clause Centering and S5 to Hierarchical Embedded-Clause Centering.

Clause Centering (S4) were found to outperform the hierarchical approaches more often (in 22, 20 and 12 RCs vs. 4 and 0 RCs for the hierarchical approaches).

In view of the results of the frequency count, the significantly higher estimated marginal means reported for Paratactic-Clause Centering for Pre-RC Degree of Cohesion and for Paratactic- and Sequential Hypotactic-Clause Centering for Post-RC Degree of Cohesion can be interpreted as better performance in the segmentation of discourse.

## 6.3 Language

Returning to our discussion of our ANOVA results, Table 31 below shows the estimated marginal means for Pre-RC Degree of Cohesion, Post-RC Degree of Cohesion and Subsequent Mention for RCs in English and Spanish.

**Table 31. Estimated marginal means by language**

| | Dependent variable | | |
|---|---|---|---|
| | Pre-RC Degree of Cohesion | Post-RC Degree of Cohesion | Subsequent Mention |
| Language | Mean (95% CI) | Mean (95% CI) | Mean (95% CI) |
| English | 3.00 (2.60 – 3.39) | 2.90 (2.50 – 3.29) | 0.374 (0.311 – 0.437) |
| Spanish | 3.12 (2.73 – 3.52) | 2.58 (2.19 – 2.98) | 0.328 (0.265 – 0.391) |

ANOVA tests of between-subjects effects revealed that the factor language was not significant for the dependent variables Pre- and Post-RC Degree of Cohesion ($F=0.2$, $df=1$, $p=.658$ and $F=1.2$, $df=1$, $p=.267$, respectively), nor for the dependent variable Subsequent Mention ($F=1.0$, $df=1$, $p=.310$). See also Table 28. We can therefore conclude that the degree of cohesion and the proportion of

subsequent mentions of RC entities do not differ across language. These findings are not unexpected, given (a) the similarity in the structural and discourse properties of RCs discussed in Chapter 4 and (b) previous research that has shown similar distributions of Centering transitions for Spanish and English conversations (Taboada & Hadic Zabala, 2008; Taboada & Wiesemann, in press).

## 6.4 Genre

Table 32 shows the estimated marginal mean values for Pre-RC Degree of Cohesion, Post-RC Degree of Cohesion and Subsequent Mention for RCs in the genres of blog, broadcast news, conversation and newspaper article.

**Table 32. Estimated marginal means by genre**

| | Dependent Variable | | |
|---|---|---|---|
| | Pre-RC Degree of Cohesion | Post-RC Degree of Cohesion | Subsequent Mention |
| Genre | Mean (95% CI)* | Mean (95% CI) | Mean (95% CI) |
| Blog | 3.15 (2.59 – 3.71) | 2.58 (2.02 – 3.13) | 0.368 (0.279 – 0.457) |
| Broadcast | 3.52 (2.96 – 4.08) | 2.83 (2.27 - 3.39) | 0.320 (0.231 – 0.409) |
| Conversation | 3.27 (2.71 – 3.83) | 2.98 (2.42 – 3.54) | 0.444 (0.355 – 0.533) |
| Newspaper | 2.29 (1.73 – 2.85) | 2.58 (2.02 – 3.14) | 0.272 (0.183 – 0.361) |

ANOVA tests of between-subjects effects revealed that the factor genre was significant for the dependent variable Pre-RC Degree of Cohesion ($F=3.6$, $df=3$, $p<.05$) and almost significant for the dependent variable Subsequent Mention ($F=2.6$, $df=3$, $p=.051$), and not significant for the dependent variable Post-RC Degree of Cohesion ($F=0.5$, $df=3$, $p=.687$). See also Table 28. This means that the degree of cohesion of transitions to the utterance containing the RC varied across

genres. In particular, post-hoc multiple comparisons tests revealed that transitions to the utterance containing the RC were significantly less cohesive for RCs in the newspaper article genre than for RCs in the broadcast news genre ($p<.01$), the conversation genre ($p<.05$) and the blog genre ($p<.05$). This is displayed in Figure 10. Post-hoc multiple comparisons tests showed similar findings for the dependent variable Subsequent Mention. The almost significant differences for that variable could be attributed to markedly lower Subsequent Mention scores for the newspaper article genre in relation to the conversation genre.

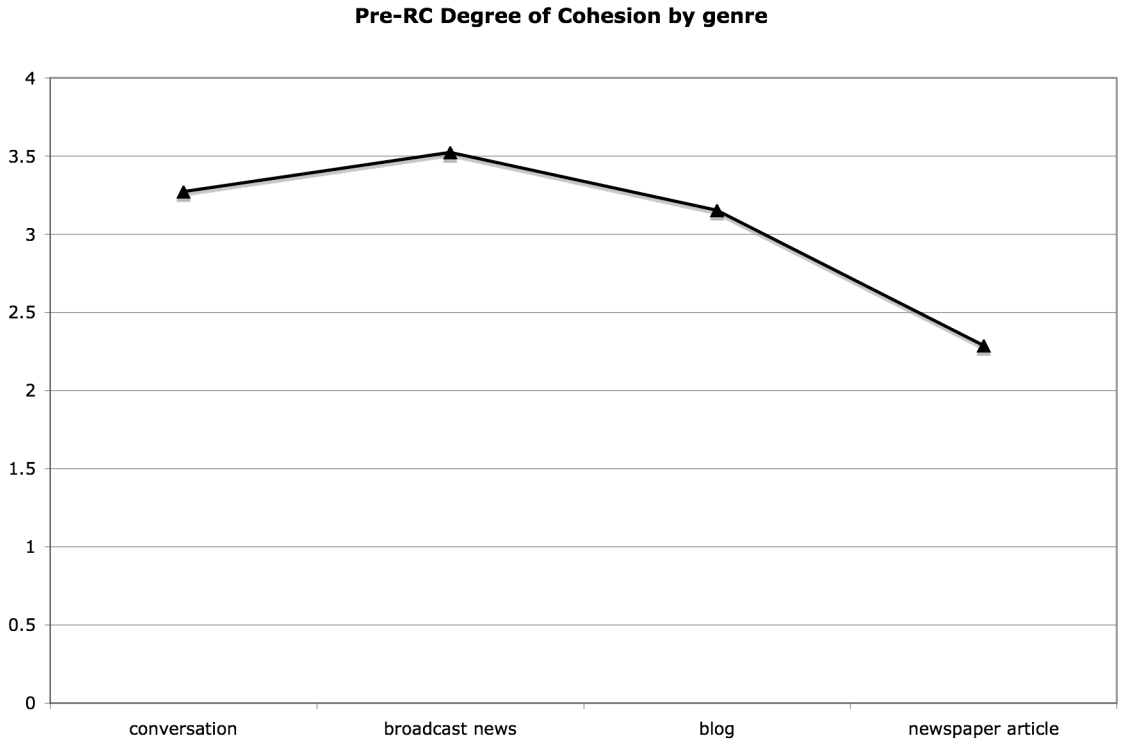


**Figure 10. Plot of estimated marginal means of Pre-RC Degree of Cohesion by genre**

Differences in cohesion across genres are not unexpected and have been otherwise reported in the literature. For instance, Taboada and Wiesemann (in press) indicated that differences in genre and mode (spoken vs. written) may account for the different distribution of Centering transitions obtained for conversations (Taboada & Wiesemann, in press) in comparison to museum descriptions and pharmaceutical leaflets (Poesio, et al., 2004a, 2004b).

A closer look at the data revealed that the lower cohesion scores for the newspaper genre may be related to the generic structure of newspaper articles. Recall from our discussion of the newspaper article genre in Chapter 5 (section 5.4.1.1), that the relation between subcomponents in the body of the news story is not linear, but orbital. White (1998) argued that the second stage in the news story genre, the body of the story, further develops the information presented in the opening stage. Depending on how the information is further developed, the body of the story may be subdivided in the subcomponents of elaboration (providing detailed information), cause-and-effect (providing causes or reasons), contextualization (setting the event in a temporal, spatial or social context) and appraisal (providing a judgment). White argued that these subcomponents do not stand in relation to each other but to the opening stage.

We reviewed the cases in which a transition to the utterance containing the RC had obtained a low score in Paratactic-Clause Centering and observed that the type of orbital organization White refers accounts for 39% of the less cohesive transitions observed in the newspaper genre (9 out of 23 transitions with a score of 1 or 0). An example is provided in (66) below.

**66**

NYT20020703-21 (Newspaper)

**Headline**

(a) VIVENDI EXPECTED TO SELL ASSETS AND RETURN TO COHERENT BUSINESS CORE

**Body of the story**

(b) Vivendi Universal, the troubled French media company, is expected to put parts of the company up for sale,

Cf: VIVENDI (MEDIA CO)> PARTS OF CO.

Cp: VIVENDI Cb: VIVENDI Ct: EST-CONTINUE DoC: 5

(c) but any auction is likely to be a tamer affair that it would have been just a few years ago.

Cf: AUCTION (OF PARTS OF CO)> AFFAIR> AFFAIR> FEW YEARS AGO

Cp: AUCTION Cb: AUCTION Ct: SMOOTH DoC: 2

(d) The corporate ambitions and market forces [$_{\text{RC-id}}$ that spurred Vivendi and other media companies like AOL Time Warner to collect company after company, in the pursuit of a digital future], have been transformed.

Cf: AMBITIONS+FORCES> AMBITIONS+FORCES> VIVENDI> OTHER MEDIA CO'S> AOLTW> COMPANIES> DIGITAL FUTURE

Cp: AMBITIONS Cb: 0 Ct: NOCB DoC: 0

(e) The media conglomerates are now loaded with debt.

Cf: MEDIA CO'S> DEBT

Cp: MEDIA Cb: MEDIA Ct: EST-CONTINUE DoC: 5

The utterance in (66d), which contains the RC, does not elaborate on the auction of parts of the media company but rather provides background on and further develops the headline by spelling out the reasons for the sale and the return to the coherent business core. In other words, the development of the

subcomponents is not strictly linear so that topics are not maintained from one utterance to the next, leading to NOCB (or ROUGH SHIFT) transitions.

Interestingly, orbital organization was also found to account for the absence of subsequent mentions of RC entities in 52% of the cases in which RC entities in the newspaper genre were not subsequently mentioned in the discourse, again in Paratactic-Clause Centering (14 out of 27 RCs with a score of 0 in Subsequent Mention). This is illustrated in Example (67) below.

**67**

NYT20020701-11 (Newspaper)

(a) Later, Simon intensified his outreach to conservative voters, particularly evangelicals.

(b) The candidate paid $30,000 to Phil Sheldon, the son of Traditional Values Coalition founder the Rev. Lou Sheldon, to serve as a consultant to reach Christian and conservative voters on the Internet.

(c) He also hired political director Steve Frank, [$_{\text{RC-rel}}$ who has worked closely with the religious right on issues ranging from support of school vouchers to opposing abortion]**.**

(d) Just days before the primary, Simon appeared on "Behind the Scenes," expressing gratitude for the opportunity to appear on TBN.

The utterances in (67 b,c and d) elaborate on how Simon has intensified his outreach to conservative voters: by hiring Phil Sheldon, by hiring Steve Frank and by appearing on a TV show on a religious network. The relevance RC in (67c) makes the hiring of Steve Frank relevant in the context. However, Steve Frank is not referred to in the following utterance. This is because utterance (67d) elaborates (67a) and not (67c).

White's orbital principle is not the only explanation for the lower Degree of Cohesion scores observed for transitions to the utterance containing the RC. 39%

(or 9 out of 23) of the ROUGH SHIFT and NOCB transitions computed in Paratactic-Clause Centering were related to a different property of the newspaper genre: reported and quoted speech and source attribution, which have been also found to affect Centering transitions in work by Mitkov and Orasan (2004). An example is provided in (68).

**68**

NYT20020703-21 (Newspaper)

(a) Yet the prospective media buyers are carrying a lot of debt themselves.

Cf: MEDIA CO. (BUYERS)> DEBT

Cp: BUYERS Cb: DEBT Ct: ROUGH DoC: 1

(b) So, analysts say, some of the potential purchasers might balk at the estimated $5 billion to $7 billion [$_{\text{RC-id}}$ that the cable television network, USA Networks, the most attractive single media property, would cost]**.**

Cf: ANALYSTS> MEDIA CO. (BUYERS)> 5-7BILLION> CABLE TV NETWORK USA.N> MEDIA PROPERTY> 5-7BILLION

Cp: ANALYSTS Cb: BUYERS Ct: ROUGH DoC: 1

The ROUGH SHIFT transition to the utterance containing the RC in (68b) is related to the structure of the utterance. The dominant clause *analysts say* reveals the source of the information whereas the dependent clause *some of the potential purchasers might balk at the estimated $5 billion to $7 billion that the cable television network, USA Networks, the most attractive single media property, would cost* further develops the information in (68a). In Paratactic-Clause Centering, entities in the dominant clause were ranked higher than entities in the dependent clause. As a result, ANALYSTS is ranked higher than

BUYERS. Had this not been the case, the transition to (68b) would have been a SMOOTH SHIFT.

White's orbital principle is not the only possible explanation for the absence of subsequent mentions of RC entities either. 33% (9 out of 27) of the RCs whose entities were not subsequently mentioned in the discourse in Paratactic-Clause Centering were found in clause complexes that contained multiple clauses, as the one shown in Example (69).

**69**

AF940517 (Newspaper)

(a) Las víctimas, dos mujeres y un hombre, viajaban en un automóvil cerca del asentamiento de Hagai, | cuando fueron atacadas con disparos de armas automáticas por palestinos [$_{\text{RC-cl}}$ que se encontraban en otro vehículo].

'The victims, two women and a man, were traveling by car near the Hagai settlement, when they were hit with fire from automatic weapons from Palestinians who were in another vehicle.'

(b) La mujer herida fue alcanzada en la cabeza.

'The woman was hit in the head.'

The classifying RC in (69a) is embedded within the dependent clause *cuando fueron atacadas con disparos de armas automáticas por palestinos* which is in a hypotactic relation to the dominant clause *Las víctimas, dos mujeres y un hombre, viajaban en un automóvil cerca del asentamiento de Hagai*. In both the dependent as well as the dominant clause, the victims are the grammatical subjects and therefore the more salient entities in those clauses. The RC modifies an oblique and therefore the head of the RC is low in salience. This means that the entities in the RC are low in salience. Given the ranking of clauses

adopted for the study (paratactic>hypotactic>embedded), we would not expect the entities in the RC to be subsequently mentioned, and they are not. In fact, one of the victims is the subject and topic of the next utterance.

In short, inherent properties of the newspaper article genre such as orbital organization, quoted and reported speech, attributions and clause complexes with multiple clauses appear to have played a role in the computation of two of the measures of cohesion: Pre-RC Degree of Cohesion and Subsequent Mention. Interestingly, RCs in newspaper articles did not perform significantly worse than their counterparts in blogs, broadcast news and conversations for the dependent variable Post-RC Degree of Cohesion (Table 32). It would seem then that those properties of the newspaper genre that affected the computation of the transition to the utterance containing the RC did not affect the computation of the transition following the RC. One possible explanation for this unexpected finding may be found in Examples (67) and (69), which illustrated cases in which RC entities were not subsequently mentioned in the discourse. In both examples, other—more salient—entities are maintained in the focus of attention in subsequent discourse. In (67), the utterance following the RC is about Simon, who was the subject of the previous utterance and also the topic of the news story. In (69), the utterance following the RC is about *la mujer* ('the woman'), one of the victims of the attack. *Las víctimas* ('the victims') was the subject of the previous utterance and also the topic of the news story. This means that the focus of attention between the utterances was maintained irrespective of the RC. Thus, properties of the newspaper genre that were said to negatively affect the dependent variable Subsequent Mention did not seem to have an effect on Post-

RC Degree of Cohesion, as more salient entities were responsible for maintaining the focus of attention across utterances. We will return to this in our discussion of the main findings in Chapter 7.

## 6.5 RC type

Table 33 shows the estimated marginal mean values for Pre-RC Degree of Cohesion, Post-RC Degree of Cohesion and Subsequent Mention for the different types of RCs.

**Table 33. Estimated marginal means by RC type**

| | Dependent Variable | | |
|---|---|---|---|
| | Pre-RC Degree of Cohesion | Post-RC Degree of Cohesion | Subsequent Mention |
| RC Type | Mean (95% CI) | Mean (95% CI) | Mean (95% CI) |
| Classifying | 2.66 (2.03 – 3.29) | 3.19 (2.57 – 3.81) | 0.390 (0.290 – 0.490) |
| Identifying | 2.72 (2.09 – 3.35) | 2.72 (2.09 – 3.34) | 0.375 (0.275 – 0.475) |
| Narrative | 3.31 (2.68 – 3.93) | 2.81 (2.19 – 3.43) | 0.340 (0.240 – 0.440) |
| Relevance | 3.46 (2.83 – 4.09) | 2.61 (1.98 – 3.23) | 0.280 (0.180 – 0.380) |
| Subjectivity | 3.15 (2.52 – 3.78) | 2.39 (1.76 – 3.01) | 0.370 (0.270 – 0.470) |

ANOVA tests of between-subjects effects revealed that the factor RC type was not significant for the dependent variables Pre-RC Degree of Cohesion (F=1.3, df=4, p=.292), Post-RC Degree of Cohesion (F=0.9, df=4, p=.474) and Subsequent Mention (F=0.7, df=4, p=.561). See also Table 28. In other words, RC type does not affect the degree of cohesion significantly. An examination of the results showed that the estimated marginal means obtained for RC type are highly heterogeneous, particularly for the dependent variable Pre-RC Degree of Cohesion. Here we can observe an absolute differences ranging from 0.06

(between identifying and classifying RCs) to 0.80 (between relevance and classifying RCs). These differences are larger than the ones observed for segmentation approaches (see Table 29), where the largest absolute difference in estimated marginal means was 0.58 between Paratactic-Clause Centering and Sequential Embedded-Clause Centering. Despite having larger absolute differences, the variance for between-subject factors (such as RC type) remains higher than for the within-subjects factor (segmentation approach). In order to evaluate whether differences between RC types are significant, a much larger sample size would be necessary.

## 6.6 Summary

In this chapter, we have reported the results of a factorial mixed-design ANOVA that was performed on our data. The most important and significant results of our study show clear differences in the degree of cohesion of RCs when the discourse is broken into different units of analysis. In particular, the segmentation approaches Paratactic-Clause Centering and Sequential Hypotactic-Clause Centering were found to yield more cohesive transitions than the embedded approaches.

With respect to the other factors in our study, language, genre and RC type, a significant main effect was only observed for the dependent variable Pre-RC Degree of Cohesion for genre. In particular, transitions to the utterance containing the RC were found to be less cohesive in the newspaper genre in comparison with the other three genres. A similar trend was observed for the

dependent variable Subsequent Mention. This was tentatively attributed to unique properties of the generic structure of newspaper articles.

In the next chapter, we focus our discussion on the major results of our study in relation to our research questions.

# CHAPTER 7: DISCUSSION

In Chapter 6, we reported on the results of a factorial mixed-design ANOVA performed on the data. The results showed significant main effects for segmentation approach and genre as well as an interaction effect between segmentation approach and RC type. We accounted for the main effect of genre and the segmentation × RC type interaction when those results were presented. Here we interpret the main effect of segmentation approach in relation to the research questions formulated in Chapter 1.

## 7.1 Degree of cohesion of RCs

### 7.1.1 Degree of cohesion and segmentation approach

Our first research question asked generally about the contribution of RCs to textual cohesion and was formulated in two sub-questions, the first one of which is reproduced below:

> Q1a: If RCs are treated as separate utterances, is the model of discourse that results from a Centering analysis more or less cohesive? Do we get more or less violations of Constraint 1 and Rule 2 of Centering? Does the Centering algorithm give us a model of discourse that is more cohesive (or less cohesive) for some types of RCs than other types of RCs?

In answering question Q1a, we refer to the major findings of our statistical analysis: In transitions to the utterance containing the RC, the segmentation of discourse in terms of independent clauses and clauses in paratactic relations (Paratactic-Clause Centering) led to higher cohesion scores than all other

segmentation approaches; in transitions to the utterance following the RC, the segmentation of discourse in terms of independent clauses, clauses in paratactic relations and clauses in hypotactic relations (Paratactic-Clause Centering and Sequential Hypotactic-Clause Centering) obtained the highest cohesion scores.

In relation to the first part of Q1a, the higher scores obtained for transitions to the utterance containing the RC in Paratactic-Clause Centering indicate that treating RCs as separate utterances does not result in models of discourse structure that are more cohesive. This is illustrated in Examples (70) and (71). In Example (70), the transition to the utterance containing the narrative NRRC in Paratactic-Clause Centering is a RETAIN (the symbol ⇒ points to the relevant utterance). Separating the NRRC from the dominant clause yields a ROUGH transition in Sequential Hypotactic-Clause Centering and an EST-RETAIN transition in Hierarchical Hypotactic-Clause Centering. The most salient entity in the previous utterance (MEXICO) is realized in the NRRC. Separating the NRRC from the dominant clause removes the linking entity thus yielding less cohesive transitions.

**70** Se96C19 (Broadcast News)

**Paratactic-Clause Centering**

En la clasificación mundial de la FIFA ((breath)) México aparece en décimo primer lugar.

'Mexico appears in 11th place in FIFA's world ranking.' '

Cf: MEXICO> 11LUGAR> CLASIFICACION> FIFA

Cp: MEXICO Cb: MEXICO Ct: SMOOTH DoC: 2

((breath)) En primera posición está Brasil, campeón del mundo, [$_{\text{RC-narr}}$ que será segundo
⇒ rival de México en la Copa América de Bolivia el año que viene. ((breath)).

'In first place is Brazil, the world champion, which will be Mexico's second opponent in
⇒ the America Cup in Bolivia next year.'

Cf: BRASIL (CAMPEON MUNDO) > 1ER LUGAR> BRASIL> RIVAL> MEXICO> COPA> BOLIVIA>
⇒ AÑO-QUE-VIENE

⇒ Cp: BRASIL Cb: MEXICO Ct: RETAIN DoC: 4

Casualmente, los tres sectores de la copa sorteados ayer quedaron agrupados de acuerdo a sus posiciones.

'Incidentally, the three groups of the cup that were drawn yesterday were grouped according to their placing.'

Cf: 3SECTORES> COPA> AYER> POSICION

Cp: SECTORES Cb: POSICION Ct: ROUGH DoC: 1

**Sequential Hypotactic-Clause Centering**

En la clasificación mundial de la FIFA ((breath)) México aparece en décimo primer lugar.

Cf: MEXICO> 11LUGAR> CLASIFICACION> FIFA

Cp: MEXICO Cb: MEXICO Ct: EST-CONTINUE DoC: 5

⇒ ((breath)) En primera posición está Brasil, campeón del mundo,

⇒ Cf: BRASIL (CAMPEON MUNDO)> 1ER LUGAR

⇒ Cp: BRASIL Cb: CLASIFICACION Ct: ROUGH DoC: 1

[RC-narr que sera segundo rival de México en la Copa América de Bolivia el año que viene]. ((breath))

Cf: BRASIL> RIVAL> MEXICO> COPA> BOLIVIA> AÑO-QUE-VIENE

Cp: BRASIL Cb: BRASIL Ct: SMOOTH DoC: 2

Casualmente, los tres sectores de la copa sorteados ayer quedaron agrupados de acuerdo a sus posiciones.

Cf: 3SECTORES> COPA> AYER> POSICION

Cp: SECTORES Cb: COPA Ct: ROUGH DoC: 1

**Hierarchical Hypotactic-Clause Centering**

En la clasificación mundial de la FIFA ((breath)) México aparece en décimo primer lugar.

Cf: MEXICO> 11LUGAR> CLASIFICACION> FIFA

Cp: MEXICO Cb: 0 Ct: NOCB DoC: 0

⇒ ((breath)) En primera posición está Brasil, campeón del mundo,

⇒ Cf: BRASIL (CAMPEON MUNDO)> 1ER LUGAR

⇒ Cp: BRASIL Cb: CLASIFICACION Ct: EST-RETAIN DoC: 3

[RC-narr que sera segundo rival de México en la Copa América de Bolivia el año que viene]. ((breath))

Cf: BRASIL> RIVAL> MEXICO> COPA> BOLIVIA> AÑO-QUE-VIENE

Cp: BRASIL Cb: BRASIL Ct: SMOOTH DoC: 2

Casualmente, los tres sectores de la copa sorteados ayer quedaron agrupados de acuerdo a sus posiciones.

Cf: 3SECTORES> COPA> AYER> POSICION

Cp: SECTORES Cb: POSICION Ct: RETAIN DoC: 4

Similarly, in Example (71), the linking entity between the previous utterance and the utterance containing the RRC is realized in the RRC (and in the hypotactic clause). While the transition to the utterance containing the identifying RRC in Paratactic-Clause Centering is a ROUGH transition, separating the RRC from its matrix clause results in NOCB transitions in both Sequential Embedded-Clause Centering and Hierarchical Embedded-Clause Centering. According to our revision of Rule 2 of Centering (provided in 65 in section 5.5.1), a RETAIN transition is preferred over an EST-RETAIN transition and a ROUGH transition; a ROUGH transition is preferred over a NOCB transition. Thus, in Examples (70) and (71), Paratactic-Clause Centering is preferred over the other approaches as a model of discourse cohesion.

**71** http://es.movies.yahoo.com/blog/ (Blog)

**Paratactic-Clause Centering**

El festival de cine más importante del mundo era el sitio ideal para montar el numerito de 'aquí no ha pasado nada',

'The most important film festival in the world was the ideal setting to set up the show that 'nothing has happened here,'

de modo que allí reaparecieron ambos.

'so that they both reappeared there.'

Cf: AJ+BP> FESTIVAL

Cp: AJ+BP Cb: FESTIVAL Ct: EST-RETAIN DoC: 3

Sin embargo, ni la alfombra roja ni la cena de gala [$_{\text{RC-id}}$ a la que acudieron juntos] ha
⇒ permitido aclarar a los medios presentes si Brad y Angelina siguen realmente juntos.

'However, neither the red carpet nor the dinner gala (they) attended together has allowed
⇒ the media to clarify if Brad and Angelina are really still together.'

⇒ Cf: ALFOMBRA ROJA> CENA> MEDIOS> AJ+BP> AJ+BP> CENA

⇒ Cp: ALFOMBRA ROJA Cb: AJ+BP Ct: ROUGH DoC: 1

La bella actriz estadounidense viajó el jueves desde La Haya (Holanda) a las Antibes, situada a siete millas de la costa de Cannes (Francia).

'The beautiful American actress traveled Thursday from The Hague (Netherlands) to Antibes, situated seven miles of the coast of Cannes (France).'

Cf: AJ> JUEVES> LAHAYA> ANTIBES> 7MILLAS> CANNES

Cp: AJ Cb: AJ Ct: CONTINUE DoC: 6

**Sequential Embedded-Clause Centering**

de modo que allí reaparecieron ambos.

Cf: AJ+BP> FESTIVAL

Cp: AJ+BP Cb: FESTIVAL Ct: EST-RETAIN DoC: 3

Sin embargo, ni la alfombra roja ni la cena de gala [emb] ha permitido aclarar a los medios
⇒ presentes

⇒ Cf: ALFOMBRA ROJA> CENA> MEDIOS

⇒ Cp: ALFOMBRA ROJA Cb: 0 Ct: NOCB DoC: 0

[RC-id a la que acudieron juntos]

Cf: AJ+BP> CENA

Cp: AJ+BP Cb: CENA Ct: EST-RETAIN DoC: 3

si Brad y Angelina siguen realmente juntos.

Cf: AJ+BP

Cp: AJ+BP Cb: AJ+BP Ct: SMOOTH DoC: 2

La bella actriz estadounidense viajó el jueves desde La Haya (Holanda) a las Antibes, situada a siete millas de la costa de Cannes (Francia).

Cf: AJ> JUEVES> LAHAYA> ANTIBES> 7MILLAS> CANNES

Cp: AJ Cb: AJ Ct: CONTINUE DoC: 6

**Hierarchical Embedded-Clause Centering**

de modo que allí reaparecieron ambos.

Cf: AJ+BP> FESTIVAL

Cp: AJ+BP Cb: FESTIVAL Ct: EST-RETAIN DoC: 3

Sin embargo, ni la alfombra roja ni la cena de gala [emb] ha permitido aclarar a los medios
⇒ presentes

⇒ Cf: ALFOMBRA ROJA> CENA> MEDIOS

⇒ Cp: ALFOMBRA ROJA Cb: 0 Ct: NOCB DoC: 0

[RC-id a la que acudieron juntos]

| | | | |
|---|---|---|---|
| Cf: AJ+BP> CENA | | | |
| Cp: AJ+BP | Cb: CENA | Ct: EST-RETAIN | DoC: 3 |
| si Brad y Angelina siguen realmente juntos. | | | |
| Cf: AJ+BP | | | |
| Cp: AJ+BP | Cb: 0 | Ct: NOCB | DoC: 0 |

La bella actriz estadounidense viajó el jueves desde La Haya (Holanda) a las Antibes, situada a siete millas de la costa de Cannes (Francia).

| | | | |
|---|---|---|---|
| Cf: AJ> JUEVES> LAHAYA> ANTIBES> 7MILLAS> CANNES | | | |
| Cp: AJ | Cb: 0 | Ct: NOCB | DoC: 0 |

While Paratactic-Clause Centering outperformed Sequential Hypotactic-Clause Centering in transitions to the utterance containing the RC, no significant differences between the two approaches were observed in transitions to the utterance following the RC. This means that in some cases, separating NRRCs from their dominant clause leads to better transitions to the following utterance, as shown in Example (72). The transition to the utterance following the relevance NRRC is a SMOOTH transition in Paratactic-Clause Centering. In the preceding utterances, the speaker (B1) has asked the listener (A) to call Delia, to tell her that the package has been received and to also ask her about the child, who is sick. The preceding utterances are then about A performing all those tasks. The utterance following the RC is about Delia (and presumably her spouse) having a sick child at home. The focus of attention has shifted. Given that Delia was talked about in the previous utterances (she's the one A is supposed to call, tell and ask), the shift is smooth.

When the NRRC is separated from the dominant clause in Sequential Hypotactic-Clause Centering, the transition to the utterance following the RC becomes a RETAIN. The topic (Cb) of the previous utterance (the NRRC) is

maintained in the current utterance: The NRRC is about *el niño*; the utterance following the NRRC is about Delia having a sick child at home. Separating the NRRC from its dominant clause thus yields a more cohesive transition between the RC and the following utterance.

**72** SP0053 (Conversation)

**Paratactic-Clause Centering**

B1: Ajá. Tal vez llamás a la Delia

'B1: uh-huh. Maybe (you) call Delia'

Cf: A> DELIA

Cp: A Cb: A Ct: CONTINUE DoC: 6

y decile que ya recibimos una caja con ropa --

'and (you) tell her that we have already received a box with clothes –'

Cf: A> DELIA> B1+> CAJA> ROPA

Cp: A Cb: A Ct: CONTINUE DoC: 6

A: Okey.

B1: y le preguntás por el niño [$_{\text{RC-rel}}$ que estaba bien enfermo].

'B1: and (you) ask her about the kid who was very sick.'

Cf: A> DELIA> NIÑO> NIÑO

Cp: A Cb: A Ct: CONTINUE DoC: 6

⇒ A: Sí. Lo tenían.

⇒ 'A: Yes. (they) had him.'

⇒ Cf: DELIA+> NIÑO

⇒ Cp: DELIA Cb: DELIA Ct: SMOOTH DoC: 2

**Sequential Hypotactic-Clause Centering**

B1: Ajá. Tal vez llamás a la Delia

Cf: A> DELIA

Cp: A Cb: A Ct: CONTINUE DoC: 6

y decile

Cf: A> DELIA

Cp: A Cb: A Ct: CONTINUE DoC: 6

que ya recibimos una caja con ropa --

Cf: B1+> CAJA> ROPA

| | | | | |
|---|---|---|---|---|
| | Cp: B1+ | Cb: 0 | Ct: NOCB | DoC: 0 |
| | A: Okey. | | | |
| | B1: y le preguntás por el niño | | | |
| | Cf: A> DELIA> NIÑO | | | |
| | Cp: A | Cb: 0 | Ct: NOCB | DoC: 0 |
| | [RC-rel que estaba bien enfermo]. | | | |
| | Cf: NIÑO | | | |
| | Cp: NIÑO | Cb: NIÑO | Ct: EST-CONTINUE | DoC: 5 |
| ⇒ | A: Sí. Lo tenían. | | | |
| ⇒ | Cf: DELIA+> NIÑO | | | |
| ⇒ | Cp: DELIA | Cb: NIÑO | Ct: RETAIN | DoC: 4 |

In relation to the contribution of RCs to textual cohesion, the two measures of cohesion discussed above present us with two different results: While Pre-RC Degree of Cohesion favours keeping the RC with its dominant or matrix clause, Post-RC Degree of Cohesion favours keeping RRCs with their matrix clauses and shows no significant difference between keeping NRRCs with their dominant clauses and separating them from their dominant clauses. In other words, RRCs appear to constitute a unit with their matrix clause whereas NRRCs seem to have some degree of independence from their dominant clauses. On the basis of these findings, a differential treatment between RRCs and NRRCs in discourse processing is warranted. This differential treatment is also theoretically justified, given their structural differences. In section 4.3.1, we identified a fundamental difference between RRCs and NRRCs in terms of their structure that has informed our analysis of RCs throughout this dissertation: According to SFL, NRRCs are clauses in dependent relations to other clauses whereas RRCs are clauses embedded in and constituent of groups in other clauses. In other words, RRCs are more intrinsic to their matrix clauses than

NRRCs to their dominant clauses. The results of our study in terms of Post-RC Degree of Cohesion support this distinction.

### 7.1.2 Degree of cohesion and RC type

The results of our statistical analysis also allow us to address the second part of Q1a, which asked whether different scores of Degree of Cohesion were obtained across functional RC type. As mentioned in section 6.5, although absolute differences in Pre-RC and Post-RC Degree of Cohesion across RC type were high, they did not reach statistical significance given the variance observed for between-subject factors. The fact that no significant differences in Degree of Cohesion were observed for the different types of RC may be seen to question the validity and necessity of a functional analysis of RC types. However, it is important to bear in mind that the corpus examined in this study consisted of a total of 200 RCs in two languages, four genres and of five different functional types. Given the number of factors and their respective levels, only five tokens of a RC type were analyzed per genre and per language. The variance observed for between-subject factors could in part be attributed to the different number of factors at play and therefore a larger sample (with more tokens per type, per genre and per language) is required in order to assess whether functional type of RC plays a role in textual cohesion. In sum, the size of our sample could be a reason why our results do not show different degrees of cohesion for different types of RCs.

## 7.2 Subsequent mention of RC entities

The second sub-question of our first research question inquired about the likelihood of subsequent mention of RC entities:

> Q1b: Are entities in RCs mentioned in subsequent discourse? Are subsequent mentions co-referential with the antecedent of the RC or with an entity inside the RC? Are there differences among the different types of RCs, so that entities in some types of RCs are more likely to be subsequently mentioned than entities in other types of RCs?

The results of both the factorial mixed-design ANOVA as well as the frequency count reported in sections 6.2 and 6.2.1 respectively showed that RC entities were subsequently mentioned approximately 50% of the time. Furthermore, the results of the frequency count presented in section 6.2.1 revealed that subsequent mentions were co-referential not only with the head noun, but also with other RC entities. We noted then, that this finding was unexpected, given the difference in topicality between head and non-head entities of RCs, as (a) head noun entities are shared with the dominant or superodinate clause and (b) they constitute the topic of the RC (i.e., what the RC is about) (e.g., Smith, 2003).

A closer look at the cases in which non-head RC entities were subsequently mentioned in the discourse revealed that the subsequent mention of non-head entities was associated with transitions (i.e., shifts) to a new topic or with the grounding of entities in the discourse.

Example (73) illustrates the first phenomenon. Father and son are discussing when the son will fly to visit the father. The son does not know when

he will be able to fly, as he has not been able to put aside enough money for the trip. The trip is nominalized for the first time in the identifying RC and it is realized as a medial demonstrative pronoun (*eso*). It then becomes the subject and topic of the following utterance. The RC can therefore be seen to facilitate the transition from the current topic (A) to a new topic (the trip).

**73** SP-0082 (Conversation)

B1: Ahora si vienes este año siquiera, pues.

B1: 'Now if (you) are coming this year at least, well.'

A: Vamos a tratar, papito,

A: '(We) will try, daddy,'

A: pero es que no he podido juntar la plata [$_{\text{RC-id}}$ que necesito para poder hacer **eso**], pues.

A: 'but (I) have not been able to put aside the money that (I) need to do that, well.'

A: **Eso** es caro, papá.

A: 'That is expensive, dad.'

Subsequent mentions of non-head RC entities were also observed in cases like Example (74), where the non-head entity grounds the head entity in the discourse. This anchoring function of RCs was identified by Fox and Thompson (1990) and previously discussed in section 4.2. According to Fox and Thompson, non-human subject head nouns (such as *la cena de gala* in 74) tend to occur with object-gap RCs (such as *a la que acudieron juntos*) because the object-gap RCs tend to have pronominal human subjects (∅ = Brad and Angelina) that make the introduction of the head noun relevant to the discourse. In other words, in

Example (74), the introduction of the dinner gala is relevant because of Brad and Angelina's attendance.

**74** http://es.movies.yahoo.com/blog/ (Blog)

Sin embargo, ni la alfombra roja ni la cena de gala [$_{\text{RC-id}}$ a la que ∅ acudieron juntos] ha permitido aclarar a los medios presentes si **Brad y Angelina** siguen realmente juntos.

'However, neither the red carpet nor the dinner gala (they) attended together has allowed the media to clarify if Brad and Angelina are really still together.'

**La bella actriz estadounidense** viajó el jueves desde La Haya (Holanda) a las Antibes, situada a siete millas de la costa de Cannes (Francia).

'The beautiful American actress traveled Thursday from The Hague (Netherlands) to Antibes, situated seven miles of the coast of Cannes (France).'

In Chapter 3 (section 3.2.1), when we reviewed the arguments for sentence-based Centering, we mentioned that the findings of Miltsakaki (2003, 2005) and Cooreman and Sanford (1996) showed entities in matrix clauses to be more topical than entities in subordinate clauses and therefore more likely to be continued in subsequent discourse. Our findings do not touch upon the topicality of entities in matrix clauses. They do however indicate that entities in subordinate clauses are likely to be mentioned in subsequent discourse. While the subsequent mention of head nouns may be related to their status in the superordinate clause, the subsequent mention of non-head RC entities appears to be linked to other discourse phenomena such as topic shifts and referent grounding. Without denying the salience of main clause entities, our results highlight the role of subordinate clause entities in the weaving of texture.

As was the case with Degree of Cohesion (Pre-RC and Post-RC), the results of the factorial mixed-design ANOVA did not yield significant differences in

Subsequent Mention across functional RC type. The table of estimated marginal means of Subsequent Mention across RC type (Table 33) showed rather similar scores for four of the five functional types of RCs: classifying, identifying, narrative and subjectivity RCs. While the score for relevance RCs was slightly lower than for the four types, no significant differences were observed. We have indicated before, that the statistical results may have been affected by sample size, so that a larger study would be required to confirm that RC type is not a significant factor.

## 7.3 An approach to discourse segmentation

Our second research question asked for the best approach to discourse segmentation given the discourse properties of RCs:

> Q2: Given the answers to Q1a and Q1b, what is the approach to discourse segmentation that best captures the functional properties of RCs?

The results of our data analysis favour first Paratactic-Clause Centering and then Sequential Hypotactic-Clause Centering over the remaining three approaches to segmentation. In particular, our findings show a clear disadvantage for the embedded approaches Sequential Embedded-Clause Centering and Hierarchical Embedded-Clause Centering. In other words, when the discourse was segmented in independent clauses and clauses in paratactic relations, the transitions obtained between utterances were more cohesive and therefore obtained higher Degree of Cohesion scores. The segmentation of discourse into smaller units, that is, separating dependent clauses (including NRRCs) from their dominant clauses affected the degree of cohesion between

utterances negatively, leading to lower scores but only in the transitions to the utterance containing the RC. Finally, the separation of embedded clauses from their matrix clauses led to even less cohesive transitions. The effect of segmentation approach on Pre-RC Degree of Cohesion was illustrated in Examples (70) and (71), for NRRCs and RRCs respectively. The effect of segmentation approach on Post-RC Degree of Cohesion is illustrated in Examples (75) and (76). The transition to the utterance following the relevance NRRC in Example (75) is an EST-RETAIN when the NRRC is processed with its dominant clause. The segmentation of discourse following Hypotactic-Clause Centering yields a NOCB transition to the utterance following the NRRC in both the sequential and the hierarchical approaches. Interestingly, this is due not only to the segmentation of the NRRC, but also to the segmentation of the utterance following the NRRC.

**75** NYT20020701-11 (Newspaper)

**Paratactic-Clause Centering**

"They support every wacky concept [$_{RC\text{-}cl}$ that comes down the line]."

Cf: CROUCHES> CONCEPT> CONCEPT> LINE

Cp: CROUCHES Cb: 0 Ct: NOCB DoC: 0

The network's lineup includes faith healer Benny Hinn,[$_{RC\text{-}rel}$ who once predicted the date [$_{RC\text{-}id}$ God would destroy all homosexuals by fire]].

Cf: NETWORK> LINEUP> BH> BH> DATE> GOD> HOMSEXUALS> FIRE> DATE

Cp: NETWORK Cb: 0 Ct: NOCB DoC: 0

⇒ In one appearance, Jan and Paul Crouch are hosts as Hinn urges viewers to put the coffins of dead loved ones near the television.

⇒ Cf: CROUCHES> APPEARANCE> BH> VIEWERS> COFFINS> DEAD> TV

⇒ Cp: CROUCHES Cb: BH Ct: EST-RETAIN DoC: 3

**Sequential Hypotactic-Clause Centering**

"They support every wacky concept [$_{RC\text{-}cl}$ that comes down the line]."

Cf: CROUCHES> CONCEPT> CONCEPT> LINE

Cp: CROUCHES Cb: 0 Ct: NOCB DoC: 0

The network's lineup includes faith healer Benny Hinn,

Cf: NETWORK> LINEUP> BH

Cp: NETWORK Cb: 0 Ct: NOCB DoC: 0

[$_{RC\text{-}rel}$ who once predicted the date [$_{RC\text{-}id}$ God would destroy all homosexuals by fire]].

Cf: BH> DATE> GOD> HOMSEXUALS> FIRE> DATE

Cp: BH Cb: BH Ct: EST-CONTINUE DoC: 5

⇒ In one appearance, Jan and Paul Crouch are hosts

⇒ Cf: CROUCHES> APPEARANCE

⇒ Cp: CROUCHES Cb: 0 Ct: NOCB DoC: 0

as Hinn urges viewers to put the coffins of dead loved ones near the television.

Cf: BH> VIEWERS> COFFINS> DEAD> TV

Cp: BH Cb: 0 Ct: NOCB DoC: 0

**Hierarchical Hypotactic-Clause Centering**

"They support every wacky concept [$_{RC\text{-}cl}$ that comes down the line]."

Cf: CROUCHES> CONCEPT> CONCEPT> LINE

Cp: CROUCHES Cb: 0 Ct: NOCB DoC: 0

The network's lineup includes faith healer Benny Hinn,

Cf: NETWORK> LINEUP> BH

Cp: NETWORK Cb: 0 Ct: NOCB DoC: 0

[$_{RC\text{-}rel}$ who once predicted the date [$_{RC\text{-}id}$ God would destroy all homosexuals by fire]].

Cf: BH> DATE> GOD> HOMSEXUALS> FIRE> DATE

Cp: BH Cb: BH Ct: EST-CONTINUE DoC: 5

⇒ In one appearance, Jan and Paul Crouch are hosts

⇒ Cf: CROUCHES> APPEARANCE

⇒ Cp: CROUCHES Cb: 0 Ct: NOCB DoC: 0

as Hinn urges viewers to put the coffins of dead loved ones near the television.

Cf: BH> VIEWERS> COFFINS> DEAD> TV

Cp: BH Cb: 0 Ct: NOCB DoC: 0

In Example (76), keeping the identifying RRC with its matrix clause yields a CONTINUE transition to the utterance following the RRC. Separating the RRC

from its matrix clause results in a NOCB transition in Sequential Embedded-Clause Centering but in an EST-CONTINUE in Hierarchical Embedded-Clause Centering. While the RRC is a statement about the moment in which the aircraft hit the Everglades, the topic of the co-text is the presence of fire in the aircraft. Elevating the RRC to the status of utterance deviates the topic from fire to aircraft, thus causing a breach in cohesion. Skipping the RRC, as is the case in Hierarchical Embedded-Clause Centering, averts this breach in cohesion.

**76** h960514 (Broadcast News)

**Paratactic-Clause Centering**

um my own feeling is (( )) listening to the eyewitness accounts ((breath)) is that there was no external fire in the aircraft

Cf: SPKR> FEELING> EYEWITNESS> FIRE> AIRCRAFT

Cp: SPKR Cb: AIRCRAFT Ct: ROUGH DoC: 1

in other words smoke was not coming from the aircraft (( )) ((breath)) until the moment [$_{RC\text{-}id}$ it hit the everglades]

Cf: SMOKE> AIRCRAFT> MOMENT> AIRCRAFT> EVERGLADES> MOMENT

Cp: SMOKE Cb: SMOKE Ct: SMOOTH DoC: 2

⇒ so that would kind of lead ((breath)) me to believe that ah there could be internal fire ((breath))

⇒ Cf: SMOKE> SPKER> FIRE

⇒ Cp: SMOKE Cb: SMOKE Ct: CONTINUE DoC: 6

**Sequential Embedded-Clause Centering**

um my own feeling is (( )) listening to the eyewitness accounts ((breath)) is [emb]

Cf: SPKR> FEELING> EYEWITNESS

Cp: SPKR Cb: 0 Ct: NOCB DoC: 0

that there was no external fire in the aircraft

Cf: FIRE> AIRCRAFT

Cp: FIRE Cb: 0 Ct: NOCB DoC: 0

in other words smoke was not coming from the aircraft (( )) ((breath)) until the moment [emb]

Cf: SMOKE> AIRCRAFT> MOMENT

Cp: SMOKE Cb: SMOKE Ct: EST-CONTINUE DoC: 5

[$_{RC\text{-}id}$ it hit the everglades]

Cf: AIRCRAFT> EVERGLADES> MOMENT

Cp: AIRCRAFT Cb: AIRCRAFT Ct: SMOOTH DoC: 2

⇒ so that would kind of lead ((breath)) me to believe

⇒ Cf: SMOKE> SPKER

⇒ Cp: SMOKE Cb: 0 Ct: NOCB DoC: 0

that ah there could be internal fire ((breath))

Cf: FIRE

Cp: FIRE Cb: FIRE Ct: EST-CONTINUE DoC: 3

**Hierarchical Embedded-Clause Centering**

um my own feeling is (( )) listening to the eyewitness accounts ((breath)) is [emb]

Cf: SPKR> FEELING> EYEWITNESS

Cp: SPKR Cb: 0 Ct: NOCB DoC: 0

that there was no external fire in the aircraft

Cf: FIRE> AIRCRAFT

Cp: FIRE Cb: 0 Ct: NOCB DoC: 0

in other words smoke was not coming from the aircraft (( )) ((breath)) until the moment [emb]

Cf: SMOKE> AIRCRAFT> MOMENT

Cp: SMOKE Cb: 0 Ct: NOCB DoC: 0

[$_{RC\text{-}id}$ it hit the everglades]

Cf: AIRCRAFT> EVERGLADES> MOMENT

Cp: AIRCRAFT Cb: AIRCRAFT Ct: EST-CONTINUE DoC: 5

⇒ so that would kind of lead ((breath)) me to believe

⇒ Cf: SMOKE> SPKER

⇒ Cp: SMOKE Cb: SMOKE Ct: EST-CONTINUE DoC: 5

that ah there could be internal fire ((breath))

Cf: FIRE

Cp: FIRE Cb: FIRE Ct: CONTINUE DoC: 6

Interestingly, the results of our study showed no significant differences between hierarchical and sequential approaches in terms of Degree of Cohesion (Pre-RC and Post-RC). In Chapter 3 (section 3.2.2), we introduced Kameyama's

(1998) distinction between sequential intrasentential Centering and hierarchical intrasentential Centering. We explained that while sequential intrasentential Centering results in a flat sequence in which the output of the last Centering unit is the input for the following Centering unit, hierarchical intrasentential Centering results in a tree structure where the output of certain units is not accessible to following discourse. We included Kameyama's sequential/hierarchical distinction in our analysis in order to assess the contribution of RRCs and NRRCs to the cohesion of a text: If, as proposed by Miltsakaki (2003, 2005), RCs do not contribute to textual cohesion, the fact that the entities in RCs are not accessible to the following utterance should not result in less cohesive transitions. This is because in the hierarchical condition, NRRCs in Hypotactic-Clause Centering, and RRCs and NRRCs in Embedded-Clause Centering are not accessible to the following utterance, meaning that the referring expressions in the following utterance are not allowed to search for an antecedent in the RC.

The results of our study then, appear to suggest that the contribution of RCs to textual cohesion is rather minimal: Skipping the RC in the computation of the transition to the following utterance does not yield less cohesive transitions. In some cases, as illustrated in Example (77), the transition to the utterance following the RC is more cohesive in Hierarchical Hypotactic-Clause Centering than in Sequential Hypotactic-Clause Centering.

77 http://new.ca.music.yahoo.com/blogs/realityrocks/ (Blog)

**Paratactic-Clause Centering**

So now Taylor Hicks is indeed a country boy.

Cf: TAYLOR> COUNTRY

Cp: TAYLOR Cb: TAYLOR Ct: CONTINUE DoC: 6

Below is his official country music debut, "Seven Mile Breakdown," [$_{RC\text{-}subj}$ which to be honest isn't that big a departure from his previous work].

Cf: TAYLOR> COUNTRY MUSIC> DEBUT> SMB> BELOW> SMB> TAYLOR> WORK

Cp: TAYLOR Cb: TAYLOR Ct: CONTINUE DoC: 6

⇒ Really, it's all about marketing these days:

⇒ Cf: IT=MUSIC> MARKETING> THESE DAYS

⇒ Cp: MUSIC Cb: MUSIC Ct: SMOOTH DoC: 2

**Sequential Hypotactic-Clause Centering**

So now Taylor Hicks is indeed a country boy.

Cf: TAYLOR> COUNTRY

Cp: TAYLOR Cb: TAYLOR Ct: CONTINUE DoC: 6

Below is his official country music debut, "Seven Mile Breakdown,"

Cf: TAYLOR> COUNTRY MUSIC> DEBUT> SMB> BELOW

Cp: TAYLOR Cb: TAYLOR Ct: CONTINUE DoC: 6

[$_{RC\text{-}subj}$ which to be honest isn't that big a departure from his previous work].

Cf: SMB> TAYLOR> WORK

Cp: SMB Cb: TAYLOR Ct: RETAIN DoC: 4

⇒ Really, it's all about marketing these days:

⇒ Cf: IT=MUSIC> MARKETING> THESE DAYS

⇒ Cp: MUSIC Cb: 0 Ct: NOCB DoC: 0

**Hierarchical Hypotactic-Clause Centering**

So now Taylor Hicks is indeed a country boy.

Cf: TAYLOR> COUNTRY

Cp: TAYLOR Cb: TAYLOR Ct: CONTINUE DoC: 6

Below is his official country music debut, "Seven Mile Breakdown,"

Cf: TAYLOR> COUNTRY MUSIC> DEBUT> SMB> BELOW

Cp: TAYLOR Cb: TAYLOR Ct: CONTINUE DoC: 6

[$_{RC\text{-}subj}$ which to be honest isn't that big a departure from his previous work].

Cf: SMB> TAYLOR> WORK

Cp: SMB Cb: TAYLOR Ct: RETAIN DoC: 4

⇒ Really, it's all about marketing these days:

⇒ Cf: IT=MUSIC> MARKETING> THESE DAYS

⇒ Cp: MUSIC Cb: MUSIC Ct: SMOOTH DoC: 2

The fact that the computation of the Centering transition is not negatively affected if the RC is not taken into account does not necessarily obliterate the role of RCs in maintaining cohesive ties with their environment. In Example (77), the blog writer makes reference to discourse-old entities in the RC: SMB (the album) and TAYLOR. In our discussion of subsequent mention, we noted that reference to RC entities is maintained in subsequent discourse. From this point of view, the role of RCs in textual cohesion is not insignificant. Why is then that the computation of Centering transitions is not negatively affected when RCs are ignored? One possibility concerns the theoretical model of discourse cohesion, Centering Theory. When we introduced the Centering algorithm in section 2.2, we identified three types of centers as the starting point of Centering Theory: the set of forward-looking centers (Cf), the preferred center (Cp) and the backward-looking center (Cb). We explained that transitions between two utterances are computed based on the realization of the Cp and the Cb. In the best-case scenario, the Cp and the Cb of the current utterance coincide with the Cb of the previous utterance (CONTINUE transition). This means that Centering gives us the best model of discourse cohesion when the discourse is about one thing at a time, what Kameyama (1998) identified as monadic tendency. The findings of Miltsakaki (2003, 2005) and Cooreman and Sanford (1996) showed entities in matrix clauses to be more salient than entities in subordinate clauses. As a result, we would expect the best model of discourse cohesion to reflect the connection between main clauses through the most salient entities. In terms of modelling discourse cohesion, then, Centering Theory can be said to model what is salient.

Entities in RCs, while they contribute to the weaving of discourse texture, they are not salient in the discourse by virtue of their realization in dependent and embedded clauses and therefore appear to be not adequately modelled by Centering Theory. Tofiloski (2009) recently proposed a revision of the Centering algorithm that would allow us to keep track of all entities in a text and may, as a result, be more adequate to model the contribution to cohesion of entities in subordinate clauses.

The fact that the computation of Centering transitions was not negatively affected when RCs are ignored may also be related to cohesive properties of RCs. Underlying this study was the assumption that different functional types of RCs would make significantly different contributions to textual cohesion. It is important to consider the possibility that this assumption may be flawed and that the major contribution of RCs may be to other aspects of discourse structure. NRRCs may be of particular relevance in the study of relational coherence, as suggested by Ferrari (2005). In discussing the differences between RRCs and NRRCs in Italian, Ferrari emphasized the role of NRRCs in relational coherence arguing that NRRCs realize minimal textual units that are in pragmatic relations with their main clauses as well as with other co-textual units. The fact that NRRCs can be substituted by other types of clauses without major changes in meaning is presented in support for the treatment of NRRCs as minimal textual

units (or units of local discourse structure). This is shown in Example (78) (from Ferrari, 2005, p. 18)[46].

**78**

a. // Maria, / che è generosa,/ mi darà certamente una mano//

'//Maria, / who is generous,/ will certainly lend me a hand//'

b. // Maria è generosa:// mi darà certamente una mano//

'//Maria is generous:// (she) will certainly lend me a hand//'

c. // Maria è generosa// (e) mi darà certamente una mano//

'//Maria is genereous// (and) (she) will certainly lend me a hand//'

d. // Poiché Maria è generosa/ mi darà certamente una mano//

'//Because Maria is generous/ (she) will certainly lend me a hand//'

e. // Maria è generosa// perciò mi darà certamente una mano//

'//Maria is generous// therefore (she) will lend me a hand//'

f. // Maria mi darà certamente una mano/ perché è generosa//

'//Maria will certainly lend me a hand/ because (she) is generous//'

g. // Maria/ data la sua generosità/ mi darà certamente una mano//

'//Maria/ given her generosity/ will certainly lend me a hand//'

The utterances in Example (78 b-g) illustrate the variety of relational meanings that may hold between a NRRC and its dominant clause. In section 2.1, we introduced a theoretical framework that focuses on coherence relations between discourse utterances: Rhetorical Structure Theory (RST) (Mann & Thompson, 1988). There we explained that RST identifies the relations that hold

[46] While the Italian original clauses are Ferrari's, the English free translations are my own. I have preserved the original word order and structure whenever possible.

between spans of text on the basis of functional criteria: The main component in a rhetorical relation is the effect, which is a plausibility judgment the analyst makes about the goals the author may have had in producing the text. Following the definitions of RST relations provided in Mann and Thompson (1988), we can identify the following relations in Example (78). In both (78b) and (78c), the first clause in these two bi-clausal units (i.e., the satellite) appears to set the framework in which the second clause (i.e., the nucleus) is to be interpreted. This corresponds to the definition of the relation of CIRCUMSTANCE. According to Mann and Thompson (1988, p. 272), in a relation of CIRCUMSTANCE, the intention of the writer is that "R [reader] recognizes that the situation presented in S [satellite] provides the framework for interpreting N [nucleus]." The relation between the satellite and the nucleus in utterances (78d-f) may be better described with the relation of VOLITIONAL CAUSE. The discourse markers *poiché* ('because' or 'since'), *perciò* ('therefore') and *perché* ('because') make the causal relation between the two clauses explicit: Maria's generosity is the cause for her lending a hand to the writer. The writer explicitly acknowledges Maria's generosity as the cause for her action and intends the reader to recognize it as the cause for her action. In (78g), the information about Maria's generosity is presented in a parenthetical non-finite (participial) clause. The relation of BACKGROUND appears to hold between the parenthetical and the main clause. Mann and Thompson (1988) explain that the presentation of background information in the satellite (the parenthetical in this case) increases the reader ability to comprehend the nucleus (Maria's lending a hand).

Now, this discussion of possible rhetorical relations between the clauses in utterances (78b-g) is not comprehensive and is only meant to illustrate Ferrari's (2005) point that NRRCs are subordinate textual units that enrich the content of the unit to which they attach. They may contribute to their dominant clause by specifying and confirming presuppositions linked to the content of the dominant clause or by confirming or denying implicatures associated with it. Crucially, the choice of a NRRC over an adverbial clause is in itself revealing: Choosing a NRRC instead of an adverbial clause makes the logical relation between the main clause and the relative clause rather vague, suggesting that this vagueness may in fact be sought and consequently, one of the meanings the speaker or writer wishes to express. In RST, NRRCs have been identified as syntactic markers of the rhetorical relation of ELABORATION (Scott & de Souza, 1990). In other words, NRRCs are seen as satellites that provide additional information about the nucleus (Mann & Thompson, 1988, p. 273). This classification is consistent with SFL's analysis of NRRCs, whereby NRRCs are seen as dependent clauses in a hypotactic relation typically expressing the logico-semantic relation of elaboration, as discussed in section 4.3.1, following Halliday and Matthiessen (2004, pp. 399-400). Stating that all NRRCs are in a relation of elaboration with their dominant clause however, may be an oversimplification: Scott and de Souza (1990, pp. 56-7) do not distinguish between embedded and hypotactic RCs in their discussion of RCs; Halliday and Matthiessen (2004) note that other types of logico-semantic relations (i.e., extension) can also be expressed with NRRCs. In light of these observations and given our analysis of Ferrari's examples in terms of RST relations, we believe that a more in-depth analysis of the types of

rhetorical relations expressed by NRRCs is justified and may be an area of further research.

Our examination of the research findings so far indicates that when it comes to the best segmentation approach, no absolute consensus among the three measures of evaluation may be found. There is however, a clear tendency in one direction: The results obtained for Pre-RC Degree of Cohesion favoured Paratactic-Clause Centering, those obtained for Post-RC Degree of Cohesion favoured Paratactic-Clause Centering and Sequential Hypotactic-Clause Centering, while the ones obtained for Subsequent Mention favoured Paratactic-Clause Centering, Sequential Hypotactic-Clause Centering and to a lesser extent, Sequential Embedded-Clause Centering. These preferences are presented in tabulated form in Table 34.

**Table 34. Summary of segmentation approach by evaluation measure**

| | **Measures of Evaluation** | | |
|---|---|---|---|
| | Pre-RC Degree of Cohesion | Post-RC Degree of Cohesion | Subsequent Mention |
| Paratactic | ✓ | ✓ | ✓ |
| Sequential Hypotactic | × | ✓ | ✓ |
| Hierarchical Hypotactic | × | × | × |
| Sequential Embedded | × | × | ✓ |
| Hierarchical Embedded | × | × | × |

As shown in Table 34, the results of our study, given our three measures of evaluation, Pre-RC and Post-RC Degree of Cohesion and Subsequent Mention, suggest a ranking of segmentation approaches according to which Paratactic-

Clause Centering is preferred over Sequential Hypotactic-Clause Centering and this in turn over Sequential Embedded-Clause Centering. This is to say, that the segmentation of discourse in terms of independent clauses and clauses in paratactic relations best captures the contribution of RCs to textual cohesion, as measured by Centering Theory.

When we first introduced our segmentation approaches in Chapter 3, we explained that Paratactic-Clause Centering was a reformulation (and standardization) of sentence-based Centering. Consequently, the results of this study are consistent with previous studies in Centering Theory that have obtained better models of textual cohesion when the "sentence" is the unit of analysis (Miltsakaki, 2002, 2003, 2005; Poesio, et al., 2000; Poesio, et al., 2004a, 2004b; Taboada & Hadic Zabala, 2008). Theoretical considerations, in particular the compatibility of Centering Theory with other models of discourse structure (e.g., RST), led researchers to express a preference for clause-based Centering (Poesio, et al., 2004a; Taboada & Hadic Zabala, 2008). While the results of our study clearly point to Paratactic-Clause Centering as the best instantiation for utterance segmentation, our ranking of segmentation approaches in terms of the statistical findings of our study allows for different implementations of utterance segmentation, when these are theoretically motivated. In other words, identifying clauses in hypotactic relations as units of local discourse structure will not yield significantly different results in two of the measures of cohesion and may be required so as to best model other aspects of discourse structure, such as relational coherence.

## 7.4 Summary

In this chapter, we have discussed the results of the statistical analyses in relation to the research questions of the study. We have seen that the segmentation of discourse into different units of analysis yields significantly different models of discourse cohesion. Identifying independent clauses and clauses in paratactic relations as units of discourse structure results in the most cohesive discourse models. The contribution of embedded clauses (RRCs) to textual cohesion was best modelled when these clauses were processed with their matrix clauses. For hypotactic clauses (NRRCs), while Pre-RC Degree of Cohesion favoured their segmentation with their dominant clause, Post-RC Degree of Cohesion favoured segmenting them on their own. These findings are noteworthy not only in that they validate structural differences between RRCs and NRRCs concerning their relationship to the matrix or dominant clause, but also in that they validate structural differences between dominant and dependent clauses in hypotactic relations on one side, and independent clauses and clauses in paratactic relations on the other, by highlighting the difference in the strength of the interdependency relation.

The discussion of results has also revealed the contribution of RCs to textual cohesion to be rather minimal: Skipping RRCs and NRRCs from the Centering analysis in the hierarchical segmentation approaches did not yield significant differences to the sequential segmentation approaches. While we acknowledged the shortcomings of Centering Theory as a model of cohesion, we also contemplated the possibility that the major contribution of RCs, in particular NRRCs, may be found in other levels of discourse structure.

Taking the different findings obtained for RRCs, NRRCs and independent clauses and clauses in paratactic relations, we proposed a ranking of segmentation approaches according to which the preferred unit of analysis at the local level of discourse structure is the independent clause and the paratactic clause. The segmentation of utterances into smaller units of analysis may yield lower levels of cohesion and is therefore dispreferred.

# CHAPTER 8: CONCLUSION

## 8.1 Summary

This study set out to explore the processing of English and Spanish RCs in discourse, so that an understanding of the contribution of RCs to the textuality of a text would inform the selection of the most adequate method for RC segmentation in Centering Theory. The choice of Centering Theory as a theoretical framework for this study was rooted in a view of discourse that distinguished among different aspects of textuality: intentionality, coherence and cohesion. In our review of discourse structure, we showed that these different aspects of textuality can be modelled theoretically if one adopts a theory of discourse that distinguishes between a global and a local level of structure. Centering Theory is a theory of local discourse structure that allows us to model cohesion in discourse.

The implementation of Centering Theory however, required the identification of the local unit of discourse structure. Drawing from work in Systemic Functional Linguistics, we formulated three possible candidates for the local unit of analysis, depending on their level of interdependency with other clauses. They were the paratactic clause, the hypotactic clause and the embedded clause. The reformulation of the notion of utterance in terms of paratactic, hypotactic and embedded clauses allowed us to capture one of the fundamental distinctions within RCs: RRCs are embedded clauses whereas NRRCs are clauses

in hypotactic relations. This basic distinction was enhanced with the findings of RC studies that have differentiated between identifying and classifying RRCs, and narrative, relevance and subjectivity NRRCs.

We selected RCs of these five different functional types from texts that realized four different genres in English and Spanish. The total 200 RCs and their immediate environment were segmented following five approaches to discourse segmentation that were derived from SFL's clausal taxonomy and Kameyama's (1998) sequential/hierarchical distinction. Centering transitions were computed and the models of discourse cohesion provided by the Centering algorithm were evaluated with the measures Degree of Cohesion (Pre-RC and Post-RC) and Subsequent Mention. A factorial mixed-design ANOVA showed clear differences in the degree of cohesion of RCs when the discourse was broken into different units of analysis: When the discourse was segmented in independent clauses and clauses in paratactic relations, the transitions obtained between utterances were more cohesive and therefore obtained higher Degree of Cohesion scores. The segmentation of discourse into smaller units, that is, separating dependent clauses (including NRRCs) from their dominant clauses affected the Degree of Cohesion between utterances negatively, leading to lower scores but only in the transitions to the utterance containing the RC. Finally, the separation of embedded clauses from their matrix clauses led to even less cohesive transitions.

Based on the results of the statistical analyses, we proposed a ranking of segmentation approaches, according to which independent clauses and clauses in paratactic relations are the preferred unit of local discourse structure, over

clauses in hypotactic relations and embedded clauses. This ranking of candidates for unit of analysis mirrors the ranking of clauses in our clause taxonomy (paratactic>hypotactic>embedded) and reflects the degree of interdependency between clauses: Clauses in paratactic relations are clauses of equal status; clauses in hypotactic relations are dependent on a dominant clause; embedded clauses are constituents of a group (or phrase) within a clause, their relation to another clause is mediated through a group (Halliday & Matthiessen, 2004).

Finally, with regard to the contribution of RCs to textual cohesion, the results of this study suggested that the role of RCs in textual cohesion as modelled by Centering Theory is not prominent enough to warrant identifying them as units of local discourse structure.

## 8.2 Implications

In our view, the major implication of this study is the definition of the unit of analysis or utterance in Centering Theory. Drawing from Systemic Functional Linguistics, we formulated Miltsakaki's sentence-based Centering and Kameyama's clause-based Centering in terms of paratactic, hypotactic and embedded clauses. Sentence in sentence-based Centering was specified as independent clauses and clauses in paratactic relations. Clause-based Centering was further subdivided to capture the distinction between clauses in hypotactic relation to other clauses, and clauses embedded in other clauses. This distinction between paratactic, hypotactic and embedded clauses allowed us to provide as systematic account of certain types of clauses that have proved problematic for Centering analysis, such as reported speech, quoted speech and non-report

complements (Kameyama, 1998; Taboada & Hadic Zabala, 2008). The distinction between hypotactic and embedded clauses was also instrumental to our analysis of the contribution of RCs to textual cohesion, as it allowed us to formalize the distinction between restrictive and non-restrictive RCs. Finally, our reformulation of sentence-based and clause-based Centering in terms of Paratactic-Clause Centering, Hypotactic-Clause Centering (sequential and hierarchical) and Embedded-Clause Centering (sequential and hierarchical) enabled us to evaluate the adequacy of the different segmentation approaches. As a result, we were able to identify Paratactic-Clause Centering as the best approach to the segmentation of RCs in discourse. This, in our opinion, constitutes a significant step towards a systematic approach to the segmentation of discourse into local units of analysis.

## 8.3 Limitations

While the results of this study suggest that the contribution of RCs to textual cohesion is such that they are best processed with their superordinate clause, it is important to bear in mind that these results were obtained on the analysis of 200 RCs. This limitation in sample size was particularly felt in the analysis of RC type effects, where, as reported in Chapter 6, considerable variance was observed for between-subject factors. Expanding the sample size is a possible direction for further research.

The generalizability of the findings would also benefit from research on other languages. This study examined the contribution of RCs to cohesion in English and Spanish texts. English and Spanish are both Indo-European

languages, and following our discussion of RCs in Chapter 4, RCs in these two languages have similar structural properties as well as discourse functions. Research on languages in which RCs perform different discourse functions and have different structural properties (e.g., OV languages) would be a welcomed addition.

## 8.4 Further research

Expanding this corpus study, either by increasing its sample size or by including different languages, is a possible direction for further research. So is complementing this corpus study on discourse processing of RCs with an experimental study of psycholinguistic processing of RCs. Indeed, the results of this corpus study may be best seen as the beginning and not the end of an enquiry into the processing of RCs in discourse. Sanders and Gernsbacher (2004), in their introduction to a special issue on the interaction of linguistic theory and psycholinguists, speak of a division of labour between discourse linguists and psycholinguists:

> There is a logical division of labor: (text and discourse) linguistics identify the relevant signals that guide the interpretation, develop theories on how the linguistic realization of information systematically varies as to 'instruct' the interpreters, and—ideally – check the validity of their theoretical work in natural language corpora. Psycholinguists develop cognitive theories on how the actual processing occurs and test these theories in psycholinguistic experiments (pp. 85-6).

In this study, we have identified "the relevant signals that guide interpretation" and have posited a paratactic-clause-based approach to segmentation. The

necessary next step would be to test the validity of our findings in psycholinguistic experiments.

There is a long tradition of psycholinguistic experiments on RCs. Frequently, studies on RC processing have focused on one of the following areas of research: (a) RC complexity, (b) site of RC attachment, (c) resolution of temporary ambiguity, and (d) syntactic priming.

Studies of RC complexity have examined properties of RCs that facilitate processing. For instance, researchers have investigated the different cues speakers use to guide their processing of RCs, that is, the types of linguistic information that allows them to comprehend RCs. Their studies have shown that the same cues that guide the interpretation of simple sentences also guide the interpretation of complex sentences containing RCs: English speakers rely on word order information for sentence and RC interpretation (Bates, Devescovi, & D'Amico, 1999), whereas Italian and Spanish speakers rely on inflectional information (Bates, et al., 1999; Hoover, 1992). This difference in processing strategy accounts for RC complexity differences between English and Spanish: Hoover (1992) shows that the reliance on inflectional information allows Spanish speakers to process double-centre-embedded RCs, whereas English speakers' reliance on word order prevents them from doing so. While research on processing cues is concerned with more general processing strategies, most research on RC complexity focuses on local properties of the NP that contains the RC. For example, researchers have found animacy of the antecedent NP and the NP in the RC to affect RC comprehension in English, German and Dutch, so that

the difficulty normally associated with object gap RCs is reduced or eliminated if the object gap RC modifies an inanimate head noun and has an animate subject (e.g., Kidd, et al., 2007; Mak, Vonk, & Schriefers, 2002, 2006; Traxler, Morris, & Seely, 2002; Traxler, Williams, Blozis, & Morris, 2005). The form of expression of the RC NP[47] and the accessibility of the RC NP were also found to play a role in RC processing. Warren and Gibson (2002) found that center-embedded (subject head) RCs with pronouns are easier to process than center-embedded RCs with lexical NPs and that the processing of center-embedded RCs is affected by the accessibility status of the RC NP so that more given entities facilitate RC processing. Reali and Christiansen (2007) found that object gap RCs with pronominal subjects are more frequent in English discourse and are more easily processed by English speakers than subject gap RCs with pronominal objects. Finally, Gordon, Hendrick and Johnson (2004) found that the semantic content of the antecedent NP facilitated the processing of object gap RCs in comparison to subject gap RCs. All of these studies point to a variety of factors that contribute to RC complexity.

Another aspect of RC processing involves the choice of a site for RC attachment. In structures of the form NP1-Prep-NP2 RC, both NP1 (high attachment) and NP2 (low attachment) are possible antecedents for the RC. Attachment preferences have been the focus of extensive research in English, Spanish, Dutch, Italian, Greek, French, Farsi, Korean and Japanese. The results of these studies have shown lexical properties of the NP such as animacy, concreteness (Desmet, De Baecke, Drieghe, Brysbaert, & Vonk, 2006)

[47] These RC NPs were NPs in subject position in object-extracted RCs.

referentiality (Gilboy, Sopena, Clifton Jr., & Frazier, 1995), lexical properties of the preposition connecting both NPs (De Vincenzi & Job, 1993, 1995), the placement of focus accent (Schafer, 1996) and prosodic phrasing (Jun, 2003) to affect RC attachment. The immediately preceding discourse context, however, was not found to affect RC attachment (Desmet, De Baecke, & Brysbaert, 2002).

English sentences that are temporarily ambiguous between a main clause and a relative clause reading have also been the subject of psycholinguistic experiments (e.g., Filik, Paterson, & Liversedge, 2005; Phillips & Gibson, 1997; Sedivy, 2002). As indicated in Chapter 3, these studies suggest that discourse context and the lexical properties of discourse connectives result in a preferred RC reading.

Discourse context can also have an effect on production. Cleland and Pickering (2003) show that for English speakers, the presence of RCs in the immediately previous discourse (i.e., in the prime) leads to RC production in the target[48]. An effect of syntactic priming has also been found for RC attachment for speakers of Dutch, German, English and Dutch-English bilinguals, whereby RCs with high attachment in the prime prompted participants to produce RCs with high attachment in the target and RCs with low attachment in the prime prompted participants to produce RCs with low attachment in the target (e.g., Desmet & Declercq, 2006; Scheepers, 2003).

[48] The effects of syntactic priming were observed when the prime noun, the antecedent of the RC in the prime, and the target noun, the noun which is a possible antecedent for an RC in the target, were the same or semantically related.

While the studies briefly reviewed here cover many aspects of RC processing, they do not examine how RCs contribute to the unfolding of discourse, namely, whether they are processed as single units or in conjunction with their main clauses, whether the information they contain is accessible in the processing of the next unit or not. Throughout our discussion of segmentation approaches in Chapter 3, we reviewed psycholinguistic studies that provided a point of comparison between RC processing and main clause processing, but their results are not conclusive enough to settle this issue once and for all. In light of the findings of this study, a pyscholinguistic study that investigates the on-line processing of RCs is desirable. Without going into a lot of detail, we speculate that such a study could consist of a self-paced reading time experiment (an on-line measure) in which mode of presentation and RC structural type would be the factors to be manipulated. Mode of presentation would include paratactic-clause-based, hypotactic-clause-based and embedded-clause-based and RC structural type would include restrictive (embedded) RCs and non-restrictive (hypotactic) RCs. Participants would be instructed to read the information presented on the screen and then asked a comprehension question referring to an entity either in the main clause or the RC. Response time and accuracy would be measured in order to show accessibility differences between entities in main (paratactic) clauses, hypotactic clauses and embedded clauses.

Echoing Sanders and Gernsbacher (2004), we believe that the results of linguistic analyses on natural corpora should inform theories of cognitive processing and the experiments that test them. The contribution of this study is a ranked set of possible units of local discourse structure. It remains then to be

explored whether the preference for independent clauses and paratactic clauses over hypotactic clauses and over embedded clauses observed in discourse processing is also found in how fast and accurately speakers recall entities. Should the findings of such an investigation confirm the results of this study, it would constitute strong evidence for the status of independent and paratactic clauses as the basic unit of local discourse structure.

# APPENDICES

## Appendix A: Coding manual

This dissertation examines the function of Spanish and English relative clauses (RCs) in discourse and the consequences of this discourse function for text segmentation. The RCs investigated in this project are headed RCs, that is, RCs that modify a head noun in a matrix clause, as in Example (1a-b), and **not** headless RCs that constitute themselves an NP that functions as an argument in a matrix clause, as in Example (2a-b).

**Example (1)**

(a) I know there's a lady in our office [$_{RC}$ who was married and got divorced]

(b) Se quejan de los terribles cambios [$_{RC}$ que ha sufrido para mal la serie]

'They complain about the terrible changes [that the series has gone through].'

The headed RCs in Example (1a-b) are clausal nominal postmodifiers, consisting of a relativizer (e.g., *that*, *que*) that co-refers with a gap in the RC and a head noun in the matrix clause. In (1a), the RC *who was married and got divorced* modifies the head noun *lady*, which functions as a predicate noun in the matrix clause. The relativizer *who* co-refers both with the head noun *lady* and with the gap in the RC, the subject of *was married and got divorced*. In (1b), the RC *que ha sufrido para mal la serie* modifies the head noun *cambios*, which functions as a direct object in the matrix clause. The relativizer *que* co-refers with the head noun and with the gap in the RC, the direct object of *ha sufrido para mal la serie*.

**Example (2)**

(a) I still appreciate [what you did with the computer].

(b) Estudiar eso, y una actualización para estudiar secretariado computadorizado, eso es [lo que piden].
'That you study that and that you are up-to-date in order to study computer-based secretarial studies, that is [what they require].

In (2a), *what you did with the computer* is a headless RC: The RC functions as the direct object of the matrix clause. In (2b), *lo que piden* is a headless RC: The RC functions as a predicate noun phrase in the matrix clause.

Once they are identified, headed RCs are coded following the steps in (1) – (4). Details for each step are provided below.

(1) Identify the discourse function of the RC.
(2) Segment the RC and its adjacent context following the five different approaches to segmentation.
(3) List entities in Cf-lists, identify Cp and Cb, compute Centering transitions.
(4) Check if entities in the RC are mentioned in the next utterance. If so, identify whether it is the head noun of the RC or an NP inside the RC that is subsequently mentioned.

**1. Identify the discourse function of the RC.**
Each RC should be coded following this taxonomy:

| Type | Function | Description |
|---|---|---|
| Restrictive | Identifying | RC adds information necessary for referent identification. |
| | Classifying | RC creates a new subclass within a reference class. |
| Non-restrictive | Narrative | RC moves the narrative time forward. |
| | Relevance | RC makes an antecedent relevant in context. |
| | Subjectivity | RC expresses opinion. |

**Table A1. Discourse functions of RCs**

The following tests may assist in the identification of the different types of RCs:

*Restrictive vs. Non-restrictive:* Restrictive RCs (RRCs) allow one-substitution (Lambrecht, 1988), whereas non-restrictive RCs (NRRCs) do not (as in Examples 3a-b).

**Example (3)**

(a) The cockroach [$_{RC}$ who lived in the paper bag] was very arrogant.

The **one** [$_{RC}$ who lived in the paper bag] was very arrogant.

(b) The cockroach, [$_{RC}$ who was very arrogant], was hated by his neighbours.

*The **one**, [$_{RC}$ who was very arrogant], was hated by his neighbours.

*Identifying RRCs* have a definite head (i.e., definite articles, possessive determiners, demonstrative determiners, quantifiers and proper nouns).

| Language | Definite articles | Possessive determiners | Demonstrative determiners | Quantifiers |
|---|---|---|---|---|
| English | the | my, your… | this, that… | all of, both of… |
| Spanish | el, la, los… | mi, tu, su… | este, ese, aquel… | todos los… |

**Table A2. Examples of definite determiners**

**Example (4)**

Gracias a BlogdeCine os traigo **las** escenas [$_{RC}$ en las que suenan estas canciones]**.**

'Courtesy of BlogdeCine I bring you the scenes [in which you can her these songs].'

*Classifying RRCs* have an indefinite head (i.e., indefinite articles, zero articles, numeral determiners (unless they are preceded by a definite determiner), quantifiers and indefinite pronouns).

| Language | Indefinite articles | Numeral determiners | Quantifiers | Indefinite pronouns |
|---|---|---|---|---|
| English | a, an | one, two… | some, much, many | everybody, anyone, someone |
| Spanish | un, una, unos… | un, dos, tres… | mucho, algo de, algunos, muchos, | alguien, todos |

**Table A3. Examples of indefinite determiners**

**Example (5)**

Además hay más puntos a favor, ya que la gente ha ido paulatinamente dejando de apoyar a la huelga, porque **muchísima** gente [$_{RC}$ que trabaja en los medios audiovisuales (técincos, maquilladores, cámaras, etc.)] ha sido despedida de sus puestos de trabajo por la finalización del rodaje de capítulos de series, programas de televisión o películas.

'In addition, there are some favourable points, especially since the public has slowly been withdrawing support for the strike, because a lot of people [who work in the media (technicians, make-up artists, cameramen, etc.,)] were fired from their jobs when they stopped shooting episodes for TV shows or movies.'

*Narrative NRRCs:* NRRCs belong to the narrative type if it is possible to insert an adverbial such as *later*.

**Example (6)**

(a) La cancillería de Bogotá inició este jueves gestiones tendientes a dar nacionalidad colombiana al hijo de los esposos Ames-Casas por petición de la abuela del menor Paul Ames, la colombiana Cecilia Dupuy de Casas, [$_{RC}$ quien se encargará del menor]**.**

'The consular office in Bogota started procedures this Thursday to give Colombian citizenship to the son of Mr & Mrs Ames-Casas, as requested by the grandmother of Paul Ames, Colombian

Cecilia Dupuy de Casas, [who will look after the minor].'

(b) La cancillería de Bogotá inició este jueves gestiones tendientes a dar nacionalidad colombiana al hijo de los esposos Ames-Casas por petición de la abuela del menor Paul Ames, la colombiana Cecilia Dupuy de Casas, [$_{RC}$ quien **más tarde** se encargará del menor]

'The consular office in Bogota started procedures this Thursday to give Colombian citizenship to the son of Mr & Mrs Ames-Casas, as requested by the grandmother of Paul Ames, Colombian Cecilia Dupuy de Casas, [who will **later on** look after the minor].

*Relevance NRRCs:* NRRCs belong to the relevance type if the relativizer can be replaced by a connector that makes the relation between the two clauses explicit. Sometimes it is hard to find a connector that specifies this relation, mainly, because NRRCs are used when the relation is not particularly clear. As a general rule, relevance NRRCs tend to provide background information that is seen as relevant to the discourse.

**Example (7)**

(a) Una mujer radicada en la Argentina, [$_{RC}$ que dice haber sido testigo de un atentado cometido en Roma hace cincuenta años], que costó la vida de 33 militares alemanes, pidió el jueves que "perdonen" al ex oficial nazi Erich Priebke, arrestado en San Carlos de Bariloche, al sur.

'A woman residing in Argentina, [who claims to have been witness of an attack that took place in Rome 50 years ago], which claimed the lives of 33 German soldiers, asked on Thursday that the former Nazi officer Erich Priebke recently arrested in San Carlos de Bariloche be pardoned.'

(b) Una mujer radicada en la Argentina, [$_{RC}$ **visto que fue** testigo de un atentado cometido en Roma hace cincuenta años], que costó la vida de 33 militares alemanes, pidió el jueves que "perdonen" al ex oficial nazi Erich Priebke, arrestado en San Carlos de Bariloche, al sur

'A woman residing in Argentina, [given that she claims to have been witness of an attack that took place in Rome 50 years ago], which claimed the lives of 33 German soldiers, asked on Thursday that the former Nazi officer Erich Priebke recently arrested in San Carlos de Bariloche be pardoned.'

*Subjectivity NRRCs:* NRRCs belong to the subjectivity type if it is possible to insert parentheticals such as *in my opinion.*

**Example (8)**

(a) I had a Reuben [$_{RC}$ which was a little bit greasy]

(b) I had a Reuben [$_{RC}$ which, **in my opinion**, was a little bit greasy]

## 2. Segment the RC and its adjacent context following the five different approaches to segmentation.

The five approaches to segmentation investigated in the study are: paratactic-clause centering, hypotactic-clause centering (sequential and hierarchical) and

embedded-clause centering (sequential and hierarchical). The concepts of paratactic, hypotactic and embedded clause are defined below. They are followed by a description and examples for all segmentation approaches.

According to Halliday and Matthiessen (2004, p. 374-5), "**Hypotaxis** is the relation between a dependent element and its dominant, the element on which it is dependent. Contrasting with this is **parataxis**, which is the relation between two like elements of equal status, one initiating and the other continuing" (pp. 374 – 5). Examples are provided in Table 4.

| **paratactic** | **hypotactic** |
|---|---|
| 1 John didn't wait;<br>=2 he ran away. | α John ran away,<br>=β which surprised everyone. |
| 1 John ran away,<br>+2 and Fred stayed behind. | α John ran away,<br>+β whereas Fred stayed behind. |
| 1 John was scared<br>x2 so he ran away. | α John ran away,<br>xβ because he was scared. |
| 1 John said:<br>"2 "I'm running away" | α John said<br>"β he was running away. |
| 1 John thought to himself:<br>'2 'I'll run away' | α John thought<br>'β he would run away. |

**Table A4. Examples of paratactic and hypotactic clauses**

*Note:* Greek notation is used to represent hypotactic structures, whereas numerals are used to represent paratactic structures: 1 identifies an initiating clause and 2 a continuing clause in a paratactic relation; α signals a dominant clause and β a dependent clause in a hypotactic relation. (The other symbols concern the type of logico-semantic relation: = signals a relation of elaboration; + one of extension, x one of enhancement; " one of locution and ' one of idea. We don't use this taxonomy of logico-semantic relations in our analysis.)

Halliday & Matthiessen (2004, p. 426) define embedding as "a semogenic mechanism whereby a clause or phrase comes to function as a constituent within the structure of a group, which itself is a constituent of a clause, e.g. *who came to dinner* in *the man who came to dinner*. Hence there is no direct relationship between an embedded clause and the clause within which it is embedded; the relationship of an embedded clause to the 'outer' clause is an indirect one, with a group as intermediary. The embedded clause functions in the structure of the group, and the group functions in the structure of the clause."

Given this distinction among paratactic, hypotactic and embedded clauses, **restrictive RCs are embedded clauses, non-restrictive RCs are hypotactic clauses**.

RCs and their adjacent contexts are segmented following these 5 approaches.

**Paratactic-Clause Centering**

A center-updating unit is an independent clause or a clause that is in a paratactic relation with another clause. Clauses that are in a hypotactic relation with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded.

| Utterance | Cf | Cp | Cb | Transition |
|---|---|---|---|---|
| B: and uh we're going to try for Germany | B+>GERMANY | B+ | B+ | EST-CONTINUE |
| B: If Germany doesn't happen we're going to try for ((Beal)) possibly ((Nollis)) Peterson [$_{RC}$ which is in Colorado] | B+>BEAL> NOLLIS PETERSON> GERMANY> PETERSON> COLORADO | B+ | B+ | CONTINUE |
| A: ooh Colorado would be neat | COLORADO | COLORADO | COLORADO | SMOOTH |
| A: I don't know | A | A | A | CONTINUE |
| A: I'm trying to find an American university [$_{RC}$ that has master's programs] | A> UNIVERSITY> UNIVERSITY> MASTERS | A | A | CONTINUE |
| B: ((breath)) NYU has one for one year ((breath)) | NYU> MASTERS> ONE YEAR | NYU | UNIV. | SMOOTH |

**Sequential Hypotactic-Clause Centering**

A center-updating unit is an independent clause or a clause that is either in a paratactic relation or a hypotactic relation with another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is sequential: The output of one center-updating unit constitutes the input for the next center-updating unit.

| Utterance | Cf | Cp | Cb | Transition |
|---|---|---|---|---|
| B: If Germany doesn't happen we're going to try for ((Beal)) possibly ((Nollis)) Peterson | GERMANY<br>B+> BEAL> NOLLIS PETERSON | GERMANY<br>B+ | GERMANY<br>0 | SMOOTH<br>NOCB |
| B: [$_{RC}$ which is in Colorado] | PETERSON> COLORADO | PETERSON | PETERSON | EST-CONTINUE |
| A: ooh Colorado would be neat | COLORADO | COLORADO | COLORADO | SMOOTH |

**Hierarchical Hypotactic-Clause Centering**

A center-updating unit is an independent clause or a clause that is either in a paratactic relation or a hypotactic relation with another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is hierarchical: Hypotactic clauses constitute embedded Centering units. Their output is not accessible to the following center-updating unit. The output of their dominant clause is used to compute the transition to the following center-updating unit.

| Utterance | Cf | Cp | Cb | Transition |
|---|---|---|---|---|
| B: and uh we're going to try for Germany | B+> GERMANY | B+ | B+ | EST-CONTINUE |
| ⇒B: If Germany doesn't happen | GERMANY | GERMANY | GERMANY | SMOOTH |
| we're going to try for ((Beal)) possibly ((Nollis)) Peterson | B+> BEAL> NOLLIS PETERSON | B+ | B+ | CONTINUE |
| ⇒B: [$_{RC}$ which is in Colorado] | PETERSON> COLORADO | PETERSON | PETERSON | SMOOTH |
| A: ooh Colorado would be neat | COLORADO | COLORADO | 0 | NOCB |

**Sequential Embedded-Clause Centering**

A center-updating unit is an independent clause or a clause that is either in a paratactic relation, in a hypotactic relation with another clause or embedded within another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Non-finite embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is sequential: The output of one center-updating unit constitutes the input for the next center-updating unit.

| Utterance | Cf | Cp | Cb | Transition |
|---|---|---|---|---|
| A:I'm trying to find an American university | A> UNIVERSITY | A | A | CONTINUE |
| A: [$_{RC}$ that has master's programs] | UNIVERSITY> MASTERS | UNIV. | UNIV. | SMOOTH |
| B: ((breath)) NYU has one for one year ((breath)) | NYU> MASTERS> ONE YEAR | NYU | UNIV. | CONTINUE |

**Hierarchical Embedded-Clause Centering**

A center-updating unit is an independent clause or a clause that is either in a paratactic relation, in a hypotactic relation with another clause or embedded within another clause. Non-finite clauses in hypotactic relations with another clause belong to the center-updating unit of their dominant clause. Non-finite embedded clauses belong to the center-updating unit of the clause in which they are embedded. The segmentation is hierarchical: Hypotactic and embedded clauses constitute embedded Centering units. Their output is not accessible to the next center-updating unit. The output of their dominant clause (for hypotactic clauses) or their matrix clause (embedded clauses) is used to compute the transition to the following center-updating unit.

| Utterance | Cf | Cp | Cb | Transition |
|---|---|---|---|---|
| A: I'm trying to find an American university | A> UNIVERSITY | A | A | CONTINUE |
| ⇒A: [RC that has master's programs] | UNIVERSITY> MASTERS | UNIV. | UNIV. | SMOOTH |
| B: ((breath)) NYU has one for one year ((breath)) | NYU> MASTERS> ONE YEAR | NYU | UNIV. | SMOOTH |

In the segmentation of RCs and their immediate context, please take into account the following special cases:

## Adverbial clauses modifying embedded clauses

Adverbial clauses modifying embedded clauses (as in Example 9) are processed with the embedded clause in the paratactic and in both hypotactic approaches and they are only separated in clauses in the embedded-clause segmentation approaches. This is because these adverbial clauses are lower than the embedded clauses in the clause hierarchy and if we were to treat them as hypotactic clauses in relation to the super-super-ordinate clause, we would make them more salient than the embedded clause (which is the clause they are in a dependent relation with).

**Example (9)**

A: That's that was like the last thing [RC she said to me [ADV before I left]] ((laugh))

## Complement-taking-predicates (CTP)

Following Hadic Zabala and Taboada (2006) and references therein, certain complement-taking-predicates signal the epistemic, evidential or evaluative stance of the speaker and are therefore treated as monoclausal utterances. These sentence-initial CTPs occur mostly with first person singular subjects (*I bet, I guess, I think, I mean*) and occasionally with second person singular subjects (*you know*).

In order to facilitate the identification of these CTPs, the following guidelines are proposed:

We differentiate between sentence-initial and sentence-final (or medial) expressions. Sentence-final expressions, like the one in Example (10), most often express epistemic, evidential and evaluative stance. (Do not include subject of CTP in Cf-ranking).

**Example (10)**

Nettie would hate that I think.

When these expressions occur sentence-initially, note the presence or absence of the complementizer ‘that’. If the complementizer 'that' is present (Example 11), treat CTP and its predicate clause as two clauses.

**Example (11)**

I know that she won't live with me

Clause 1: I know

Clause 2: that she won't live with me

If the complementizer is not present and the CTP can be replaced with an epistemic, evidential or possibility adverb ( 'perhaps', 'in my opinion', 'apparently'), as in Example (12), then treat CTP and predicate as a monoclausal unit (Do not include subject of CTP in Cf-ranking).

**Example (12)**

I guess little Felipe is staying at our house.

Apparently little Felipe is staying at our house.

If the complementizer is not present but it is not possible to replace the CTP with an adverb, treat the CTP and its predicate as two clauses (no example in the data so far).

These guidelines apply only to first person singular subjects (I), when they occur with verbs like *guess*, *think*, *mean*, *bet*, etc., and second person singular subjects (you) when they occur with the verb *know*.

## Attributions

Both in the news broadcast and the newspaper genre, it is common for speakers and writers to reveal the source of information in clauses of the type shown in Example (13).

**Example (13)**

Un sobrino del ex Presidente chileno Augusto Pinochet, el comerciante Santiago Townsend Pinochet, fue detenido en Santiago en un proceso por estafa, **informaron el jueves fuentes policiales.**

'A nephew of former Chilean president Augusto Pinochet, businessman Santiago Townsend Pinochet, was arrested in Santiago in a case of fraud, police sources announced on Thursday.'

Our treatment of attributions depends on their form: We treat them as dependent clauses in hypotactic relations when they are introduced by subordinators such as *según* ('as'); we treat them as matrix clauses or dominant clauses when they introduce report complements.

## 3. List entities in Cf-lists, identify Cp and Cb, compute Centering transitions.

*Realization:* All entities realized in an utterance are included in the Cf-list. Entity realization includes indirect realization: substitution, synonymy, superordinate and general word.

*Ranking:* Entities in the Cf –list are ranked following the English and Spanish templates:

**English Cf Template (**Walker et al., 1998)

Subject > Indirect Object > Direct Object > Other

**Spanish Cf Template (Taboada, 2002; 2008)**

Experiencer > Subject > Animate Indirect Object > Direct Object > Other > Impersonal pronoun

The Cf-list for each clause is populated following these two rankings. For the segmentation approaches that allow more than one clause, the ranking of entities in the dominant clause precedes the ranking of entities in the dependent clause (following Miltsakaki, 2002; 2003). So, in the paratactic segmentation approach, if there is a hypotactic clause in relation with a paratactic clause, we rank entities in the following order: **paratactic clause > hypotactic clause**, as in Example (14a). If there are more than one hypotactic clause, we follow linear order. When hypotactic and paratactic clauses are separated (in the hypotactic-clause segmentation), the order in which they were produced is preserved. In Example (14b), the hypotactic clause precedes the paratactic clause.

**Example (14)**

Because John was sick, Mary stayed home.

**(a) Paratactic segmentation**:

Because John was sick, Mary stayed home.

*Cf: Mary>home>John*

**(b) Hypotactic segmentation**

Because John was sick,

*Cf: John*

Mary stayed home.

*Cf: Mary>home*

In the paratactic and hypotactic segmentation approaches, the integrity of paratactic and hypotactic clauses will be preserved and all embedded clauses (RCs, sentence complements, etc.) will be ranked with their matrix clauses, as in Example (15a-b). In these cases, entities in the matrix clause are ranked before entities in the embedded clause. The embedded-clause segmentation approaches (sequential and hierarchical) are the ones that test embedded clauses as units of processing, so embedded clauses are separated from their matrix clauses only in these approaches, as in Example (15c).

**Example (15)**

Because John was sick, the woman who usually looks after his kids had to look after him.

**(a) Paratactic-clause segmentation**:

Because John was sick, the woman who usually looks after his kids had to look after him.

*Cf: woman>John>woman>John>kids>John*

**(b) Hypotactic-clause segmentation**

Because John was sick,

*Cf: John*

the woman who usually looks after his kids had to look after him.

*Cf: woman>John>woman>John>kids*

**(c) Embedded-clause segmentation**

Because John was sick,

*Cf: John*

the woman [rc] had to look after him.

*Cf: woman>John*

who usually looks after his kids

*Cf: woman>John>kids*

*Possessive NPs*: Following Byron and Stent (1998), we will rank possessive NPs in linear order (see Example 15).

*Cp and Cb:* Following Centering's rules and principles:

- The Cp is the highest element of the current Cf-list.
- The Cb is the highest element of the Cf-list of the previous utterance realized in the current utterance.

*Centering transitions*: Transitions are computed following Wiesemann's revision of Centering Transitions (based on Grosz et al., 1995, Kameyama, 1985 and Poesio et al., 2004a,b).

| | $Cb(U_{i-1}) = Cb(U_i)$ | $Cb(U_{i-1}) = 0$ | $Cb(U_{i-1}) \neq Cb(U_i)$ |
|---|---|---|---|
| $Cb(U_i) = Cp(U_i)$ | CONTINUE | EST-CONTINUE | SMOOTH SHIFT |
| $Cb(U_i) \neq Cp(U_i)$ | RETAIN | EST-RETAIN | ROUGH SHIFT |

**Table A5. Centering Transitions**

### 4. Check if entities in the RC are mentioned in the next utterance. If so, identify whether it is the head noun of the RC or an NP inside the RC that is subsequently mentioned.

Once the Centering analysis has been computed, check if any of the entities in the RC are mentioned in the following utterance (the following utterance may differ from one segmentation approach to the next). If any of the entities are mentioned in the following utterance, please note whether this subsequent mention co-refers with the head noun of the RC (the noun that the RC modifies) or with an entity inside the RC (i.e. head or non-head).

Note that in the hierarchical segmentation approaches, embedded Centering units and are not accessible to the following utterance. These utterances have no next utterance.

## Appendix B: Normalized frequencies

In order to be able to compare the frequency of the different types of RCs across genres in both languages, we normalized their raw frequency to occurrence per 1,000 words, following Biber, Conrad and Reppen (1994) and Biber, Conrad and Reppen (1998). Biber et al. (1998, p. 263) describe the procedure for normalizing frequency counts: "The raw frequency should be divided by the number of words in the text, and then multiplied by whatever basis is chosen for normalizing." The normalized frequencies are provided in Table B1.

**Table B1. Normalized frequency of RCs in our data**

| Genre | Words | Identifying (n-freq)[a] | Classifying (n-freq) | Narrative (n-freq) | Relevance (n-freq) | Subjectivity (n-freq) |
|---|---|---|---|---|---|---|
| English | | | | | | |
| Conv. | 21660 | 31 (1.43) | 41 (1.89) | 5 (0.23) | 17 (0.78) | 15 (0.69) |
| Blog | 14845 | 37 (2.49) | 34 (2.29) | 7 (0.47) | 13 (0.87) | 9 (0.60) |
| Broad. | 13250 | 50 (3.77) | 30 (2.26) | 8 (0.60) | 11 (0.83) | 6 (0.45) |
| Newsp. | 5875 | 12(2.04) | 20 (3.40) | 11 (1.87) | 26 (4.42) | 5 (0.85) |
| Spanish | | | | | | |
| Conv. | 23418 | 43 (1.83) | 33 (1.40) | 11 (0.46) | 14 (0.59) | 8 (0.34) |
| Blog | 4477 | 16 (3.57) | 21 (4.69) | 12 (2.68) | 10 (2.23) | 11 (2.45) |
| Broad. | 13794 | 59 (4.27) | 32 (2.31) | 21 (1.52) | 35 (2.53) | 8 (0.57) |
| Newsp. | 4918 | 14 (2.84) | 15 (3.05) | 22 (4.47) | 14 (2.84) | 8 (1.62) |

[a] n-freq refers to normalized frequency, RCs per 1000 words.

# REFERENCE LIST


Aarts, B. (2000). Corpus linguistics, Chomsky and fuzzy tree fragments. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory. Papers from the 20th International Conference on English Language Research and Computerized Corpora (ICAME 20)* (pp. 5-13). Amsterdam: Rodopi.

Aarts, J. (2002). Does corpus linguistics exist? Some old and new issues. In L. E. Breivik & A. Hasselgren (Eds.), *From the colt's mouth...and others. Language corpora studies in honour of Anna-Brita Stenström* (pp. 1-17). Amsterdam: Rodopi.

Asher, N. (2004). Troubles with topics: Comments on Kehler, Oberlander, Stede and Zeevat. *Theoretical Linguistics, 30*(2-3), 255-262.

Baldwin, F. B. (1995). *CogNIAC: A discourse processing engine*. Unpublished Dissertation, University of Pennsylvania.

Ballantyne, K. G. (2004). Givenness as a ranking criterion in Centering Theory: Evidence from Yapese. *Oceanic Linguistics, 43*(1), 49-72.

Barzilay, R., & Lapata, M. (2005). Modeling local coherence: An entity-based approach. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 141-148.

Bates, E., Devescovi, A., & D'Amico, S. (1999). Processing complex sentences: A cross-linguistic study. *Language and Cognitive Processes, 14*(1), 69-123.

Beaver, D. (2000). The optimization of discourse. Unpublished Manuscript. Stanford University.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in Applied Linguistics. *Applied Linguistics, 15*(2), 169-189.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Essex, UK: Pearson Education Limited.

Bolinger, D. (1977). *Meaning and form*. London: Longman.

Breivik, L. E. (1999). On the pragmatic function of relative clauses and locative expressions in existential sentences in the LOB Copus. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 121-136). Amsterdam: Rodopi.

Brennan, S. E., Walker Friedman, M., & Pollard, C. J. (1987). A Centering approach to pronouns. *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL)*, 155-162.

Brucart, J. M. (1999). La estructura del sintagma nominal: Las oraciones de relativo. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 395-522). Madrid: Espasa Calpe.

Butler, C. S. (2004). Corpus studies and functional linguistic theories. *Functions of Language, 11*(2), 147-186.

Butt, J., & Benjamin, C. (2000). *A new reference grammar for modern Spanish* (3rd ed.). Lincolnwood, IL: NTC Publishing Group.

Byron, D. K., & Stent, A. (1998). A preliminary model of Centering in dialog. *Proceedings of the 36th Annual Meeting of the Association for Coomputational Linguistics (ACL 98)*, 1475-1477.

Callaway, C. B., & Lester, J. C. (2002). Pronominalization in generated discourse and dialogue. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 88-95.

Canac Marquis, R., & Tremblay, M. (1997). The wh-feature and the syntax of restrictive and non-restrictive relatives in French and English. In J. Lema & E. Treviño (Eds.), *Theoretical analyses on Romance languages. Selected papers from the 26th linguistic symposium on Romance languages (LSRL XXVI)* (pp. 127-141). Amsterdam: John Benjamins Publishing Company.

Chafe, W. (1988). Linking intonation units in spoken English. In J. Haiman & S. A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 1-27). Amsterdam: John Benjamins Publishing Company.

Chafe, W. (1992). The importance of corpus linguistics to understanding the nature of language. In J. Svartvik (Ed.), *Directions in corpus linguistics. Proceedings of the Nobel Symposium* (pp. 79-97). Berlin: Mouton de Gruyter.

Chambers, C. G., & Smyth, R. (1998). Structural parallelism and discourse coherence: A test of Centering Theory. *Journal of Memory and Language, 39*(4), 593-608.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin and use*. New York: Praeger.

Clarke, J., & Lapata, M. (2007). Modeling compression with discourse constraints. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1-11.

Clayman, S., & Heritage, J. (2002). *The news interview: Journalists and public figures on the air*. New York: Cambridge University Press.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language, 49*(2), 214-230.

Collier-Sanuki, Y. M. F. (1993). *Word order and discourse grammar: A contrastive analysis of Japanese and English relative clauses in written narratives*. Unpublished Dissertation, University of California, Los Angeles.

Cooreman, A., & Sanford, A. (1996). *Focus and syntactic subordination in discourse* (Technical Report): Human Communication Research Centre, Glasgow University.

Coupland, N. (2001). Stylization, authenticity and TV news review. *Discourse Studies, 3*(4), 413-442.

Cristea, D., Ide, N., & Romary, L. (1998). Veins Theory: A model of global discourse cohesion and coherence. *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 98)*, 281-285.

De Haan, P. (1987). Relative clauses in indefinite noun phrases. *English Studies, 68*(2), 171-190.

De Vincenzi, M., & Job, R. (1993). Some observations on the universality of the late-closure strategy. *Journal of Psycholinguistic Research, 22*(2), 189-206.

De Vincenzi, M., & Job, R. (1995). An investigation of late closure: The role of syntax, thematic structure, and pragmatics in initial interpretation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1303-1321.

Desmet, T., De Baecke, C., & Brysbaert, M. (2002). The influence of referential discourse context on modifier attachment in Dutch. *Memory & Cognition, 30*(1), 150-157.

Desmet, T., De Baecke, C., Drieghe, D., Brysbaert, M., & Vonk, W. (2006). Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes, 21*(4), 453-485.

Desmet, T., & Declercq, M. (2006). Cross-linguistic priming of syntactic hierarchical configuration information. *Journal of Memory and Language, 54*(4), 610-632.

Di Eugenio, B. (1996). The discourse functions of Italian subjects: A Centering approach. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, 352-357.

Di Eugenio, B. (1998). Centering in Italian. In M. Walker, A. Joshi & E. F. Prince (Eds.), *Centering Theory in discourse* (pp. 115-137). Oxford: Clarendon Press.

Diessel, H., & Tomasello, M. (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics, 11*(1-2), 131-151.

Dimitriadis, A. (1995). When pro-drop languages don't: On overt pronominal subjects in Greek. *University of Pennsylvania Working Papers in Linguistics, 2*(2), 45-60.

Dimitriadis, A. (1996). When pro-drop languages don't: Overt pronominal subjects and pragmatic inference. *Papers from the Regional Meetings, Chicago Linguistic Society, 32*(1), 33-47.

Downing, B. T. (1978). Some universals of relative clause structure. In J. H. Greenberg (Ed.), *Universals of human language* (Vol. 4 Syntax, pp. 375-418). Stanford, CA: Stanford University Press.

EAGLES (1996). Expert Advisory Group on Language Engineering Standards, 2008, from http://www.ilc.cnr.it/EAGLES96/

Eggins, S. (1994). *An introduction to Systemic Functional Linguistics*. London: Pinter Publishers.

Eggins, S. (2004). *An introduction to Systemic Functional Linguistics* (2nd ed.). New York: Continuum.

Eggins, S., & Slade, D. (1997). *Analysing casual conversation*. London: Cassell.

Fais, L. (2004). Inferable centres, Centering transitions and the notion of coherence. *Computational Linguistics, 30*(2), 119-150.

Fais, L., & Yamura-Takei, M. (2003). The nature of referent resolution in Japanese e-mail. *Discourse Processes, 36*(3), 167-204.

Ferrari, A. (2005). Appositive relatives in text construction. *Cuadernos de Filologia Italiana, 12*, 9-32.

Filik, R., Paterson, K. B., & Liversedge, S. P. (2005). Parsing with focus particles in context: Eye movements during the processing of relative clause ambiguities. *Journal of Memory and Language, 53*(4), 473-495.

Fillmore, C. J. (1992). "Corpus linguistics" or "computer-aided armchair linguistics". In J. Svartvik (Ed.), *Directions in corpus linguistics. Proceedings of the Nobel Symposium 82* (pp. 35-60). Berlin: Mouton de Gruyter.

Forbes, K., & Miltsakaki, E. (2002). Empirical studies of Centering shifts and cue phrases as embedded segment boundary markers. *University of Pennsylvania Working Papers in Linguistics, 7*(2), 39-57.

Fox, B. A., & Thompson, S. A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language, 66*(2), 297-316.

Frossard, M., Cardebat, D., & Nespoulous, J. L. (2001). Anaphoric resolution and discourse focus: The case of the French "hybrid" demonstrative pronoun [celui-ci]. *Psicologia: Reflexao e Critica, 14*(2), 429-437.

Gervasi, K. L. (2000). *A variationist study of relative clauses in Spanish*. Unpublished Dissertation, University of Southern California.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1-76.

Gilboy, E., Sopena, J. M., Clifton Jr., C., & Frazier, L. (1995). Argument structure and association preferences in Spanish and English complex NPs. *Cognition, 54*(2), 131-167.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in Education and Psychology*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gordon, P. C., & Chan, D. (1995). Pronouns, passives, and discourse coherence. *Journal of Memory and Language, 34*(2), 216-231.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science, 17*(3), 311-347.

Gordon, P. C., & Hendrick, R. (1997). Intuitive knowledge of linguistic co-reference. *Cognition, 62*, 325-370.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language, 51*(1), 97-114.

Graff, D., & Alabiso, J. (1997). 1996 English broadcast news transcripts (HUB4) LDC97T22 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Greenbaum, S. (1996). *The Oxford English grammar*. Oxford: Oxford University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). New York: Academic Press.

Griffiths, D., Stirling, W. D., & Weldon, K. L. (1998). *Understanding data. Principles & practice of statistics*. Brisbane, Australia: John Wiley & Sons.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL)*, 44-50.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203-225.

Grosz, B. J., & Sidner, C. L. (1986). Attentions, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175-204.

Grosz, B. J., & Sidner, C. L. (1998). Lost intuitions and forgotten intentions. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering Theory in discourse* (pp. 39-54). Oxford: Clarendon Press.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*(2), 274-307.

Hadic Zabala, L. (2004). *Complex noun phrases in ESL narratives: Structural and discourse properties*. Unpublished M.A. Thesis, Simon Fraser University, Burnaby.

Hadic Zabala, L. (2007). The genre of on-line personal ads. *Proceedings of the 22nd Northwest Linguistics Conference (NWLC 22)*, 133-144.

Hadic Zabala, L., & Taboada, M. (2006). Centering Theory in Spanish: A coding manual. Unpublished Manuscript. Simon Fraser University.

Halliday, M. A. K. (2004). The spoken language corpus: A foundation for grammatical theory. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)* (pp. 11-38). Amsterdam: Rodopi.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to Functional Grammar* (3rd ed.). London: Arnold.

Hatch, E., & Lazaraton, A. (1991). *The research manual. Design and statistics for Applied Linguistics*. Boston, MA: Heinle & Heinle Publishers.

Hedberg, N. (in press). Centering and noun phrase realization in Kaqchikel Mayan. *Journal of Pragmatics*.

Hedberg, N., & Dueck, S. (1999). Cakchiquel reference and Centering Theory. *Proceedings of the Workshop on Structure and Constituency in the Languages of the Americas, University of British Columbia Working Papers in Linguistics*, 59-74.

Herring, S. C., Scheidt, L. A., Wright, E., & Bonus, S. (2005). Weblogs as a bridging genre. *Information Technology & People, 18*(2), 142-171.

Hitzeman, J., & Poesio, M. (1998). Long distance pronominalization and global focus. *Proceedings of the 36th annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics (ACL/COLING 1998)*, 550-556.

Hoffman, B. (1996). Translating into free word order languages. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, 556-561.

Hoover, M. L. (1992). Sentence processing strategies in Spanish and English. *Journal of Psycholinguistic Research, 21*(4), 275-299.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.

Iida, M. (1998). Discourse coherence and shifting centers in Japanese texts. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering Theory in discourse* (pp. 161-182). Oxford: Clarendon Press.

Johansson, S. (1995). ICAME - Quo Vadis? Reflections on the use of computer corpora in linguistics. *Computers and the Humanities, 28*, 243-252.

Johnstone, B. (2002). *Discourse analysis*. Malden, MA: Blackwell Publishers.

Jun, S.-A. (2003). Prosodic phrasing and attachment preferences. *Journal of Psycholinguistic Research, 32*(2), 219-249.

Kameyama, M. (1985). *Zero anaphora: The case of Japanese*. Unpublished Dissertation, Stanford University.

Kameyama, M. (1998). Intrasentential Centering: A case study. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering Theory in discourse* (pp. 89-114). Oxford: Clarendon Press.

Karamanis, N. (2006). Evaluating Centering for sentence ordering in two new domains. *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, 65-68.

Karamanis, N. (2007). Supplementing entity coherence with local rhetorical relations for information ordering. *Journal of Logic, Language and Information, 16*, 445-464.

Karamanis, N., Mellish, C., Poesio, M., & Oberlander, J. (2009). Evaluating Centering for information ordering using corpora. *Computational Linguistics, 35*(1), 29-46.

Karamanis, N., Poesio, M., Mellish, C., & Oberlander, J. (2004). Evaluating Centering-based metrics of coherence for text structuring using a reliably annotated corpus. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 04)*, 391-398.

Keenan, E. L., & Comrie, B. (1977). Noun Phrase accessibility and universal grammar. *Linguistic Inquiry, 8*(1), 63-99.

Kehler, A. (1993). Intrasentential constraints on intersentential anaphora in Centering Theory *Proceedings of the Workshop on Centering Theory in Naturally Occurring Discourse*: University of Pennsylvania.

Kehler, A. (1997). Current theories of Centering for pronoun interpretation: A critical evaluation. *Computational Linguistics, 23*(3), 467-475.

Kibble, R. (1999). *Cb or not Cb? Centering Theory applied to NLG* (Report No. ITRI-99-17). Brighton, UK: University of Brighton.

Kibble, R. (2001). A reformulation of Rule 2 of Centering Theory. *Computational Linguistics, 27*(4), 579-587.

Kibble, R., & Power, R. (1999). *Using Centering Theory to plan coherent texts* (Report No. ITRI-99-19). Brighton, UK: University of Brighton.

Kibble, R., & Power, R. (2004). Optimizing referential coherence in text generation. *Computational Linguistics, 30*(4), 401-416.

Kidd, E., & Bavin, E. L. (2002). English-speaking children's comprehension of relative clauses: Evidence for general-cognitive and language-specific constraints on development. *Journal of Psycholinguistic Research, 31*(6), 599-617.

Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes, 22*, 860-897.

Kim, H., Cho, J. M., & Seo, J. (1999). Anaphora resolution using an extended Centering algorithm in a multi-modal dialogue system. *Proceedings of the ACL Workshop on the Relation of Discourse/Dialogue Structure and Reference*, 21-28.

Kingsbury, P., Strassel, S., McLemore, C., & McIntyre, R. (1997). CALLHOME American English transcripts LDC97T14 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Knox, J. (2007). Visual-verbal communication on online newspaper home pages. *Visual Communication, 6*(1), 19-53.

Kruijff-Korbayova, I., & Hajicova, E. (1997). Topics and centers: A comparison of the salience-based approach and the Centering Theory. *Prague Bulletin of Mathematical Linguistics, 67*, 25-50.

Lambrecht, K. (1988). There was a farmer had a dog: Syntactic amalgams revisited. In S. Axmaker, A. Jaisser & H. Singmaster (Eds.), *Berkeley Linguistics Society Proceedings of the 14th Annual Meeting* (pp. 319-339). Berkeley, CA: Berkeley Linguistics Society.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics. Proceedings of the Nobel Symposium 82* (pp. 105-122). Berlin: Mouton de Gruyter.

Leech, G. (2004). Recent grammatical change in English: Data, description, theory. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research and Computerized Corpora (ICAME 23)* (pp. 61-81). Amsterdam: Rodopi.

Loock, R. (2007). Appositive relative clauses and their functions in discourse. *Journal of Pragmatics, 39*(2), 336-362.

Maes, A. (1997). Referent ontology and Centering in discourse. *Journal of Semantics, 14*(3), 207-235.

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of Memory and Language, 47*(1), 50-68.

Mak, W. M., Vonk, W., & Schriefers, H. (2006). Animacy in processing relative clauses: The hikers that rocks crush. *Journal of Memory and Language, 54*(4), 466-490.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.

Martin, J. R. (1992). *English text: System and structure*. Philadelphia, PA: John Benjamins Publishing Company.

Matsumoto, K. (2003). *Intonation units in Japanese conversation: Syntactic, informational, and functional structures*. Amsterdam: John Benjamins.

McCawley, J. D. (1981). The syntax and semantics of English relative clauses. *Lingua, 53*(2-3), 99-149.

Miltsakaki, E. (2001). Centering in Greek *Proceedings of the 15th International Symposium on Theoretical and Applied Lingusitics*. Thessaloniki, Greece.

Miltsakaki, E. (2002). Toward an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics, 28*(3), 319-355.

Miltsakaki, E. (2003). *The syntax-discourse interface: Effects of the main-subordinate distinction on attention structure*. Unpublished Dissertation, University of Pennsylvania.

Miltsakaki, E. (2005). A Centering analysis of relative clauses in English and Greek. *University of Pennsylvania Working Papers in Linguistics, 11*(1), 183-197.

Mitkov, R., & Orasan, C. (2004). Discourse and coherence: Revisiting specific conventions of the Centering Theory. *Proceedings of DAARC2004*, 109-114.

Mosegaard Hansen, M. B. (1998). *The function of discourse particles: A study with special reference to spoken standard French*. Amsterdam: Rodopi.

Munoz, E., Alabiso, J., & Graff, D. (1998). 1997 Spanish broadcast news transcripts (HUB4-NE) LDC98T29 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Okumura, M., & Tamura, K. (1996). Zero pronoun resolution in Japanese discourse based on Centering Theory. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, 871-876.

Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics, 23*(1), 103-139.

Pérez González, L. (2006). Interpreting strategic recontextualization cues in the courtroom: Corpus-based insights into the pragmatic force of non-restrictive relative clauses. *Journal of Pragmatics, 38*(3), 390-417.

Phillips, C., & Gibson, E. (1997). On the strength of the local attachment preference. *Journal of Psycholinguistic Research, 26*(3), 323-346.

Poesio, M., Cheng, H., Henschel, R., Hitzeman, J., Kibble, R., & Stevenson, R. (2000). Specifying the parameters of Centering Theory: A corpus-based evaluation using text from application-oriented domains. *Proceedings of*

*the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, 400-407.

Poesio, M., Stevenson, R., Di Eugenio, B., & Hitzeman, J. (2004a). Centering: A parametric theory and its instantiations. *Computational Linguistics, 30*(3), 309-363.

Poesio, M., Stevenson, R., Di Eugenio, B., & Hitzeman, J. (2004b). *Centering: A parametric theory and its instantiations* (Report No. CSM-369). Colchester, UK: Natural Language Engineering Group, Univeristy of Essex.

Prasad, R., & Strube, M. (2000). Discourse salience and pronoun resolution in Hindi. *University of Pennyslvania Working Papers in Linguistics, 6*(3), 189-208.

Prince, E. F. (1990). Syntax and discourse: A look at resumptive pronouns. *Proceedings of the Berkeley Linguistics Society, 16*, 482-497.

Prince, E. F. (1998). Subject-prodrop in Yiddish. In P. Bosch & R. van der Sandt (Eds.), *Focus: Linguistic, cognitive, and computational perspectives* (pp. 82-104). Cambridge, UK: Cambridge University Press.

Rambow, O. (1993). Pragmatic aspects of scrambling and topicalization in German: A Centering approach *Proceedings of the Workshop on Centering Theory in Naturally-Occurring Discourse*. Institute of Research in Cognitive Science, University of Pennsylvania.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language, 57*(1), 1-23.

Renkema, J. (1993). *Discourse studies: An introductory textbook*. Amsterdam: John Benjamins.

Renkema, J. (2004). *Introduction to discourse studies*. Amsterdam: John Benjamins.

Reppen, R., Ide, N., & Suderman, K. (2005). American National Corpus (ANC) Second Release LDC2005T35 [Corpus]. Philadelphia, PA: Linguisitc Data Consortium.

Roberts, R., & Gibson, E. (2002). Individual differences in sentence memory. *Journal of Psycholinguistic Research, 31*(6), 573-598.

Rogers, W. (2000). TREC Spanish LDC2000T51 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Roh, J. E., & Lee, J.-H. (2006). Examining Cf-ranking methods for text structuring and pronominalization in Korean. *International Journal of Computer Processing of Oriental Languages, 19*(1), 39-61.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Sag, I. (1997). English relative clauses. *Journal of Linguistics, 33*, 431-483.

Sanders, T. J. M., & Gernsbacher, M. A. (2004). Accessibility in text and discourse processing. *Discourse Processes, 37*(2), 79-89.

Schafer, A. (1996). Focus in relative clause construal. *Language and Cognitive Processes, 11*(1), 135.

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition, 89*(3), 179-205.

Schiffrin, D. (1994). *Approaches to discourse*. Malden, MA: Blackwell Publishers.

Scott, D., & de Souza, C. S. (1990). Getting the message across in RST-based text generation. In R. Dale, C. Mellish & M. Zock (Eds.), *Current research in natural language generation* (pp. 47-73). London: Academic Press.

Sedivy, J. C. (2002). Invoking discourse-based contrast sets and resolving syntactic ambiguities. *Journal of Memory and Language, 46*(2), 341-370.

Sheldon, A. (1977). On strategies for processing relative clauses: A comparison of children and adults. *Journal of Psycholinguistic Research, 6*(4), 305-318.

Sinclair, J. (Ed.). (1998). *Collins Cobuild English grammar*. London: Harper Collins.

Smith, C. S. (1964). Determiners and relative clauses in a Generative Grammer of English. *Language, 40*(1), 37-52.

Smith, C. S. (2003). *Modes of discourse: The local structure of texts*. Cambridge, UK: Cambridge University Press.

SPSS (1998). SPSS Base 8.0 for Windows user's guide. Chicago, IL: SPSS Inc.

Strube, M., & Hahn, U. (1996). Functional Centering. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, 270-277.

Strube, M., & Hahn, U. (1999). Functional Centering - Grounding referential coherence in information structure. *Computational Linguistics, 25*(3), 309-344.

Stys, M. E., & Zemke, S. (1995). *Incorporating discourse aspects in English-Polish MT: Towards robust implementation* (Report No. LiTH-IDA-R-95-18). Linköping, Sweden: Department of Computer and Information Science, Linköping University.

Suri, L. Z., & McCoy, K. F. (1994). RAFT/RAPR and Centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics, 20*(2), 301-317.

Suri, L. Z., McCoy, K. F., & DeCristofaro, J. D. (1999). A methodology for extending focusing frameworks. *Computational Linguistics, 25*(2), 173-194.

Taboada, M. (2002). Centering and pronominal reference: In dialogue, in Spanish. *Proceedings of the 6th Workshop on the Semantics and Pragmatics of Dialogue (EDILOG 2002)*, 177-184.

Taboada, M. (2004). *Building coherence and cohesion: Task-oriented dialogue in English and Spanish*. Philadelphia, PA: John Benjamins Publishing.

Taboada, M. (2008). Reference, centers and transitions in spoken Spanish. In J. Gundel & N. Hedberg (Eds.), *Reference and reference processing*. Oxford: Oxford University Press.

Taboada, M., & Hadic Zabala, L. (2008). Deciding on units of analysis within Centering Theory. *Corpus Linguistics and Linguistic Theory, 4*(1), 63-108.

Taboada, M., & Wiesemann, L. (in press). Subjects and topics in conversation. *Journal of Pragmatics*.

Tao, H., & McCarthy, M. J. (2001). Understanding non-restrictive which-clauses in spoken English, which is not an easy thing. *Language Sciences, 23*(6), 651-677.

Tetreault, J. R. (1999). Analysis of syntax based pronoun resolution methods. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, 602-605.

Tetreault, J. R. (2001). A corpus-based evaluation of Centering and pronoun resolution. *Computational Linguistics, 27*(4), 507-520.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10*(1), 1-13.

Thompson, S. A., & Couper-Kuhlen, E. (2005). The clause as a locus of grammar and interaction. *Discourse Studies, 7*(4-5), 481-505.

Tofiloski, M. (2009). *Extending Centering Theory for the measure of entitiy coherence*. Unpublished MSc Thesis, Simon Fraser University.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins Publishing Company.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language, 47*(1), 69-90.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language, 53*(2), 204-224.

Turan, U. D. (1996). *Null vs. overt subjects in Turkish discourse: A Centering analysis*. Unpublished Dissertation, University of Pennsylvania.

Ungerer, F. (2004). Ads as news stories, news stories as ads: The interaction of advertisements and editorial texts in newspapers. *Text, 24*(3), 307-328.

Ventola, E. (1987). *The structure of social interaction: A systemic approach to the semiotics of service encounters*. London: Frances Pinter.

Walker, M., Iida, M., & Cote, S. (1994). Japanese discourse and the process of Centering. *Computational Linguistics, 20*(2), 1-37.

Walker, M., Joshi, A., & Prince, E. F. (1998). Centering in naturally-occurring discourse: An overview. In M. Walker, A. Joshi & E. Prince (Eds.), *Centering Theory in discourse* (pp. 1-30). Oxford: Clarendon Press.

Walker, M., & Prince, E. F. (1996). A bilateral approach to Givenness: A hearer-status algorithm and a Centering algorithm. In T. Fretheim & J. Gundel (Eds.), *Reference and referent accessibility* (pp. 291-306). Amsterdam: John Benjamins.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition, 85*(1), 79-112.

Weinert, R. (2004). Relative clauses in spoken English and German: Their structure and function. *Linguistische Berichte, 197*(Feb), 3-51.

Wheatley, B. (1996). CALLHOME Spanish transcritps LDC96T17 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

White, P. R. (1998). *Telling media tales: The news story as rhetoric.* Unpublished Dissertation, University of Sidney.

Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies.* Cambridge, UK: Cambridge University Press.

Yeh, C. L., & Chen, Y. C. (2004). Topic identification in Chinese based on Centering model. *iProceedings of the Conference on Reference Resolution and its Applications*, 103-109.

Yuksel, O., & Bozsahin, C. (2002). Contextually appropriate reference generation. *Natural Language Engineering, 8*(1), 69-89.