Entity Coherence for Descriptive Text Structuring

Nikiforos Karamanis

Doctor of Philosophy

Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh

2004

Abstract

Although entity coherence, i.e. the coherence that arises from certain patterns of references to entities, is of attested importance for characterising a descriptive text structure, whether and how current formal models of entity coherence such as Centering Theory can be used for the purposes of natural language generation remains unclear. This thesis investigates this issue and sets out to explore which of the many formulations of Centering best suits text structuring. In doing this, we assume text structuring to be a search task where different orderings of propositions are evaluated according to scores assigned by a metric.

The main question behind this study is how to choose a metric of entity coherence among many alternatives as the only guidance to the text structuring component of a system that produces descriptions of objects. Different ways of defining metrics of entity coherence using Centering's notions are discussed and a general corpus-based methodology is introduced to identify which of these metrics constitute the most promising candidates for search-based text structuring before the actual generation of the descriptive structure takes place.

The performance of a large set of metrics is estimated empirically in a series of computational experiments using two kinds of data: (i) a reliably annotated corpus representing the genre of interest and (ii) data derived from an existing natural language generation system and ordered according to the instructions of a domain expert. A final experiment supplements our main methodology by automatically evaluating the best scoring orderings of some of the best performing metrics in comparison to an upper bound defined by orderings produced by multiple experts on additional application-specific data and a lower bound defined by a random baseline.

The main findings are summarised as follows: In general, the simplest metric of entity coherence constitutes a very robust baseline for both datasets. However, when the metrics are modified according to an additional constraint on entity coherence, then the baseline is beaten in domain (ii). The employed modification is supported by the subsidiary evaluation which renders all employed metrics superior to the random baseline and helps identify the metric which overall constitutes the most suitable candidate (among the ones investigated) for search-based descriptive text structuring in domain (ii).

This thesis provides substantial insight into the role of entity coherence as a descriptive text structuring constraint. Viewing Centering from an NLG perspective raises a series of interesting challenges that the thesis identifies and attempts to investigate to a certain extent. The general evaluation methodology and the results of the empirical studies are useful for any subsequent attempt to generate a descriptive text structure in the context of an application that makes use of the notion of entity coherence as modelled by Centering.

Acknowledgements

I am deeply grateful to Chris Mellish and Jon Oberlander, my supervisors, not only for their careful guidance and continuous commitment to my work, but primarily for their enormous support every time things did not move all that smoothly and quickly. My views on Centering have benefited immensely from the insights of Massimo Poesio who made the GNOME corpus available to me and commented on subsequent drafts of the thesis. I am also thankful to Marilyn Walker and Frank Keller, my examiners, for a thoroughly critical, yet enjoyably rewarding, defence.

Special thanks to the participants in the studies of chapters 4 and 9, to Aggeliki Dimitromanolaki for entrusting me with her data and for her assistance in the preparation of the materials in chapter 9 and to Maria Lapata for her prompt advice and for providing me with her computational tools. Katerina Kolotourou's assistance in recruiting the archaeologists for chapter 9 was invaluable, while Amy Isard and Tracy Markusic spent precious time checking the materials for chapter 4. Ruli Manurung and David Schlangen have been great friends and extremely knowledgeable colleagues who never minded me asking silly questions about, frequently trivial, matters. It is because of the help from all aforementioned that this thesis has been completed.

More occasional discussions in person or by email with Rodger Kibble, Eleni Miltsakaki, Dora Alexopoulou, Natasha Mangana, Colin Matheson, David Beaver, Alistair Knott, Johanna Moore, Donia Scott, Barbara Di Eugenio and Ellen Bard improved my knowledge and the clarity of my views substantially. The comments of the audiences of INLG-02, IGK-02, EGK-01, CLUK-4, the IDT seminar and other fora within and outside the University of Edinburgh, where different aspects of my work were presented, were particularly useful as well.

Betty Hughes, David Dougal, Melissa Davies, Eva Barrett and the rest of the people in the administration section were always extremely helpful. Many thanks to Andrew Woods, Julieta Albertina Pineda, Roger Burroughes and everyone else in the Informatics and the Linguistics technical support teams who had to deal with my frequent, but often poorly expressed, questions. They all made my everyday life easier, helping me devote more time to my actual research.

Surely, my most substantial gain from the last five years in Edinburgh was many new friends: Eva Zemlickova, Atif Suleman, Antonis Rokas, Yiannos Tolias, Vaso Sapountzi, Emilio Perez, Maria Navarrete Lopez, Ivan Yuen, Malte Gabsdil, Peter Dienes, Karen Wahl, Hara Klassina, Marta Corsin Jimenez, Vicky Kynourgiopoulou, Laura Serlenga, Tassos Christianopoulos, Dimitris Konstantinou, Alexander Melengoglou, the residents of Churchill House and my officemates were always tremendous company. Antonis Kariniotakis, Maria Melissari and Effi Georgala remained in touch regularly although we were scattered in different parts of the world. Yufrita and David Skyner were ideal landlords while Antoon Goderis and Craig Morris impeccable flatmates. It is friends like these that make one feel very lucky and special.

Last but not least, I am grateful to my family, especially my eldest sister Elina, for always being close to me despite the distance between us (and for their financial support throughout the years). And to Franziska Kindervater for her attention, understanding and patience at very stressful times. This thesis is devoted to my grandmother Amalia for always prioritising παιδεία over εκπαίδευση.

Submission: October 3, 2003
Defence: November 7, 2003
Binding: December 1, 2003

****

My PhD study was mainly supported by a generous scholarship from the Greek Scholarships Foundation (IKY), while most of my travel expenses for talks and conferences outside Edinburgh were covered by the organisers or by the Informatics and ICCS travel grants.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Nikiforos Karamanis*)

Hope your road is a long one.
Ithaka, C.P. Cavafy (1863-1933),
translation by Edmund Keeley & Philip Sherrard

Table of Contents

1 Introduction **1**

1.1 Text structuring in NLG . . . 2

1.2 Text structuring and entity coherence . . . 3

1.3 Chapter-by-Chapter overview . . . 5

2 Motivation **9**

2.1 Major approaches to text structuring . . . 9

2.1.1 Schemata . . . 9

2.1.2 RST-based planning . . . 11

2.1.3 An ELABORATIONless framework of descriptive text structure . . . 14

2.1.4 Structuring descriptive text . . . 16

2.1.5 A search-based approach to descriptive text structuring . . . 18

2.1.6 Centering: the missing entity link? . . . 19

2.2 Centering Theory . . . 20

2.2.1 Basic terminology . . . 20

2.2.2 Formalisation . . . 21

2.2.3 Example discourses . . . 23

2.2.4 Underlying notions of Centering . . . 25

2.2.5 More Centering . . . 28

2.3 Corpus-based evaluations of Centering . . . 29

2.3.1 Underspecification of Centering . . . 29

2.3.2 Using a reliably annotated corpus for evaluation . . . 32

2.3.3 Remaining issues in corpus-based evaluation . . . 34

2.4 Centering and natural language generation . . . 34

2.4.1 Centering and focusing in TEXT . . . 35

2.4.2 Centering and the generation of referring expressions . . . 36

2.4.3 Integrating Centering with text generation . . . 37

2.4.4 Centering beyond ELABORATION 38
2.5 Applying Centering to longer spans of text 41
2.5.1 RETAIN as a prediction for a SHIFT 42
2.5.2 Estimating the coherence of the whole text 44
2.5.3 Looking at alternative sequences of utterances 45
2.5.4 Estimating the coherence of the whole text structure requires search 47
2.6 A research question for search-based text structuring 51

3 Defining metrics of entity coherence **53**
3.1 Existing metrics of entity coherence 53
3.1.1 Metrics in stochastic ILEX 54
3.1.2 Summing up the underlying notions 55
3.1.3 Isolating the effects of entity coherence 57
3.2 An OT ranking of the underlying principles 58
3.2.1 Resolving RSI 59
3.2.2 Alternative POT rankings 62
3.3 Simpler metrics of entity coherence 63
3.3.1 Computing only the violations of CHEAPNESS 63
3.3.2 Computing only the violations of CONTINUITY 64
3.3.3 ROUGH-SHIFT as a source of incoherence 65
3.3.4 What is a SHIFT? 66
3.4 Transition-based metrics of entity coherence 67
3.4.1 The BFP metric 68
3.5 Examining the relation between principles and transitions 70
3.5.1 Extending the table of FC transitions 70
3.5.2 A new set of transitions 72
3.6 The proliferation of CT-based metrics 74
3.7 The second research question for text structuring 75

4 A preliminary study on text acceptability **77**
4.1 Introduction 77
4.2 Magnitude estimation 77
4.3 Experimental conditions 79
4.3.1 Predictions 79
4.3.2 Testing for significance 81
4.4 Method 82

4.4.1 Participants . . . 82
4.4.2 Materials . . . 82
4.4.3 Procedure . . . 84
4.5 Results . . . 85
4.5.1 Outliers . . . 85
4.5.2 Results for strict optimality . . . 87
4.5.3 Computing the acceptability order . . . 88
4.5.4 Discussion . . . 90
4.6 The cost of human-based evaluation . . . 91

5 Corpus-based evaluation: Methodology 93
5.1 Motivation . . . 93
5.2 Issues in corpus-based evaluation . . . 94
5.2.1 The assumption of the gold standard . . . 94
5.3 The Basis for Comparison . . . 96
5.4 Exploring the search space . . . 97
5.4.1 Comparing permutations using SEEC . . . 98
5.5 Computing a performance measure . . . 100
5.5.1 Calculating the classification rate . . . 100
5.5.2 Testing for significance . . . 104
5.5.3 Comparing several metrics . . . 104
5.6 Dealing with factorial complexity . . . 105
5.6.1 Sample size . . . 106
5.6.2 Implementing random sampling . . . 107
5.6.3 Empirical studies on random sampling . . . 108

6 Experiments on the GNOME corpus 111
6.1 Motivation and aims . . . 111
6.2 The GNOME corpus . . . 112
6.2.1 GNOME-LAB: Museum labels in GNOME . . . 114
6.3 Computing the inputs to SEEC . . . 115
6.3.1 Definition of utterance . . . 115
6.3.2 Computing CF lists for local rhetorical relations . . . 117
6.3.3 CF ranking . . . 118
6.4 Experimental questions . . . 120
6.4.1 Rhetorical compensation . . . 120

6.5 Results . . . 121
6.5.1 Average classification rate . . . 122
6.5.2 Pairwise comparisons with M.NOCB . . . 123
6.5.3 Differences between Finite and Finite-RR . . . 125
6.5.4 Discussion . . . 127
6.6 Future work . . . 128
6.7 Summary . . . 129

7 Initial experiments on the MPIRO domain 131
7.1 Motivation and aims . . . 131
7.2 Database facts in MPIRO . . . 132
7.3 Computing the inputs to SEEC . . . 133
7.4 Results . . . 134
7.4.1 Average classification rate . . . 134
7.4.2 Pairwise comparisons with M.NOCB . . . 135
7.4.3 Examining the number of ROUGH-SHIFTs in the BfC . . . 137
7.4.4 Discussion . . . 140
7.5 Summary of chapters 6 and 7 . . . 141

8 The role of PageFocus 143
8.1 Motivation . . . 143
8.1.1 Computing the BestTable . . . 143
8.1.2 The PageFocus constraint on entity coherence . . . 144
8.2 Experimental questions . . . 147
8.3 Results from MPIRO-PROP . . . 148
8.3.1 Average classification rate . . . 148
8.3.2 Differences between PF-modified and non-PF metrics . . . 149
8.3.3 Pairwise comparisons with PF.NOCB . . . 149
8.4 Results from GNOME-LAB . . . 152
8.4.1 Average classification rate . . . 152
8.4.2 Differences between PF-modified and non-PF metrics . . . 154
8.4.3 Pairwise comparisons with M.NOCB . . . 155
8.5 Summary and discussion . . . 157

9 Using data from more than one expert 159
9.1 Motivation and aims . . . 160

9.1.1 Secondary experiment . . . 161
9.2 Gathering additional data . . . 162
9.2.1 Examining the BestOrders . . . 162
9.2.2 Implementing a random baseline . . . 163
9.2.3 Realising facts as sentences . . . 163
9.3 Interviews with experts . . . 165
9.4 Dependent variable . . . 165
9.4.1 Calculating significance . . . 166
9.4.2 Computing the distance between the experts . . . 167
9.4.3 Computing the distance between the experts and a metric . . . 169
9.5 Main experiment: Predictions . . . 169
9.6 Results of main experiment . . . 171
9.6.1 Distances between the experts and each other . . . 171
9.6.2 Distances between the experts and RB . . . 173
9.6.3 Distances between the experts and each metric . . . 174
9.6.4 Concluding analysis . . . 176
9.6.5 PF.BFP: A case for variability in text structuring? . . . 178
9.7 Differences between MPIRO-PROP and AllEq . . . 180
9.8 Summary and conclusion . . . 181

10 Concluding remarks **183**
10.1 Contributions . . . 183
10.2 Possible extensions . . . 185
10.2.1 Experimenting with more metrics . . . 185
10.2.2 Extending GNOME-LAB . . . 185
10.2.3 Future work in the MPIRO domain . . . 186
10.2.4 Choosing between more complex metrics . . . 187
10.2.5 Computing input characteristics . . . 187
10.2.6 Psycholinguistic plausibility . . . 187

A Examples of basic CT and extended PT transitions **189**

B Instructions to participants in ME experiment **193**

C Weighting Equal for the classification rate **199**

D Instructions to experts **201**

Bibliography **203**

List of Figures

4.1 Experimental conditions and their NOCB transitions and COHERENCE violations . . . 80
4.2 Boxplots of conditions with and without outliers 86
4.3 Experimental modulus and lexicalisation in condition (4.4) 91

5.1 Algorithm for calculating how many permutations of the elements of the set SC_B score Better, Equal and Worse than the permutation represented by B according to metric M 99
5.2 Distributions of the search space of possible permutations for metrics M_x and M_y . . 101
5.3 Spliting the set of Equal for metric M_x in the distribution of the search space for M_y . 103

9.1 Differences in BestOrders for PF.NOCB, PF.BFP and PF.KP 163
9.2 Schema employed by expert E3 for the ordering task 173
9.3 Example BestOrders for metric PF.BFP 178

List of Tables

2.1 Centering transitions 22
2.2 COHERENCE and SALIENCE in the table of Centering transitions 27
2.3 Transitions in Functional Centering 28

3.1 Violations of Centering notions and scores for examples (3.1) and (3.2) according to the scoring function S.KP 57
3.2 Violations of Centering principles in the utterances of example (3.3) 61
3.3 Some alternative rankings of the violations of Centering notions 63
3.4 Scores for examples (3.1) and (3.2) according to metric M.MIL 66
3.5 Scores for examples (3.1) and (3.2) according to metrics M.SH and M.SHOT1 67
3.6 Standard Centering transitions in examples (3.1) and (3.2) 69
3.7 Basic Centering transitions using all combinations of the three Centering principles . 71
3.8 The various types of the ESTABLISHMENT transition 71
3.9 Basic transitions in the Principles and Transitions (PT) formulation of Centering . . . 72
3.10 Extending the transitions in the Principles and Transitions (PT) formulation of Centering with ESTABLISHMENTS 74
3.11 The metrics used in our experiments 75

4.1 Rankings of conditions under the Strict Optimality Hypothesis for each metric 80
4.2 Rankings of conditions under the Relative Suboptimality Hypothesis for each metric 81
4.3 Mean acceptability and standard error (SE) for the experimental conditions 87
4.4 Mean acceptability and standard error (SE) for merged conditions 88
4.5 Determining the acceptability order of the experimental conditions 89

5.1 Approximate computation time for $n!$ permutations 105
5.2 Average number of replicated permutations within and between samples 107

6.1 Average classification rate (Y=Better+Equal/2) in Finite and Finite-RR 122
6.2 Winners of pairwise comparisons with M.NOCB in Finite and Finite-RR 123

6.3 Details of pairwise comparisons with M.NOCB in Finite 124
6.4 Details of pairwise comparisons with M.NOCB in Finite-RR 125
6.5 Details of the comparisons of Finite-RR versus Finite for each metric 126
6.6 Distribution of BfCs according to the difference in a) the percentage of NOCBs and b) the classification rate (υ) of M.NOCB between Finite-RR and Finite 127

7.1 Average classification rate (Y=Better+Equal/2) in MPIRO-PROP and GNOME-LAB 135
7.2 Winners of pairwise comparisons with M.NOCB in the two datasets (MPIRO-PROP and GNOME-LAB) 136
7.3 Details of pairwise comparisons with M.NOCB in MPIRO-PROP 136
7.4 Distribution of the BfCs according to a) their number of ROUGH-SHIFTs and b) the result of M.NOCB versus M.MIL in MPIRO-PROP, Finite and Finite-RR 139

8.1 Scoring functions of the modified metrics which compute the sum of NOCBPF∗s independently from other scores 147
8.2 Average classification rate (Y=Better+Equal/2) for PF-modified and non-PF metrics in MPIRO-PROP 148
8.3 Details of comparisons of each PF-modified metric with its non-PF counterpart in MPIRO-PROP 149
8.4 Details of pairwise comparisons with metric PF.NOCB in MPIRO-PROP 150
8.5 Details of comparisons of PF-modified metrics overtaking the baseline with each other 151
8.6 Average classification rate (Y=Better+Equal/2) of PF-modified metrics in GNOME-LAB 152
8.7 Average classification rate (Y) of PF-modified and non-PF metrics in GNOME-LAB 153
8.8 Details of comparisons of each PF-modified metric with its non-PF counterpart in Finite 153
8.9 Details of comparisons of each PF-modified metric with its non-PF counterpart in Finite-RR 154
8.10 Winners of comparisons of each PF-modified metric with its non-PF counterpart in GNOME-LAB and MPIRO-PROP 155
8.11 Details of comparisons of metric M.NOCB with potential PF-modified competitors in GNOME-LAB 156
8.12 Winners of pairwise comparisons of metric M.NOCB with some PF-modified metrics and their non-PF counterparts 157

9.1 Example of orders for sentences A to F from two experts E1 and E2 166
9.2 Comparison of distances between the experts (EM, E1, E2, E3) and each other 171

9.3 Comparison of distances between experts (EM, E1, E2, E3) and the random baseline (RB) 174
9.4 Comparison of distances between experts (EM, E1, E2, E3) and each metric (PF.BFP, PF.KP, PF.NOCB, M.NOCB) 175
9.5 Results of the concluding analysis comparing the distance between the experts and each other (EXP_{EXP}) with the distance between the experts and each metric (PF.BFP, PF.NOCB, PF.KP, M.NOCB) and the random baseline (RB) 176
9.6 Testitems (%) for which the experts achieve significant τ scores 180
9.7 Average T in MPIRO-PROP and AllEq 181

Chapter 1

Introduction

This doctoral thesis is about:

a) Defining metrics of entity coherence in a text structure.

b) Exploring the usefulness of some of these metrics for the text structuring component in natural language generation.

After identifying a large set of potential answers to point (a) above, we focus on evaluating some of the possible candidates from the generation perspective presented in (b). Crucially, although the problem we are dealing with is extremely relevant to text structuring, the thesis **is not** about implementing the various possibilities in the context of a generation system and deciding which works out best on the basis of elicited human judgements. Instead, the possible solutions are investigated prior to the actual generation of a text structure. Thus, the evaluation is **automatic** and **corpus-based**, instead of relying on generally more expensive psycholinguistic techniques.

Following Mellish and Dale (1998), we distinguish between the evaluation of a **theory** for the purposes of language generation and the evaluation of the chosen **implementation** in a system. The experimental methodology described in this thesis represents a principled way of evaluating the underlying theory from a generation perspective. It is our means for choosing the most motivated candidates for generating a text structure within the premises of a specific theory before the actual generation takes place. Then, human-based evaluation on the output of an NLG system can show whether implementing such solutions in the text structuring component does indeed generate felicitous structures. In this sense, the corpus-based evaluation reported in this thesis can be seen as a test-bench that provides a subsequent human-based evaluation with some testable hypotheses.

In this chapter, we start with a brief introduction to natural language generation and the text structuring task in particular. Then, we show how text structuring relates to the notion of entity coherence. We conclude the chapter with an overview of the thesis.

1.1 Text structuring in NLG

Reiter and Dale (2000) define natural language generation (NLG) as the subfield of artificial intelligence and computational linguistics which focuses on computer systems that produce understandable texts in English or some other human language. Typically starting from some nonlinguistic representation of information as the input, NLG systems use knowledge about language and the application domain to automatically produce reports, descriptions, directions, explanations, help messages and other kinds of text.

Despite the variability in the structure of NLG systems, the process of generation appears to break down logically into at least six tasks, each of which can be described informally as follows (for more details see Reiter and Dale 2000, Chapter 3):

- *Content determination* selects which information should be communicated to the user.
- *Text structuring* organises this information. This includes decisions on how chunks of content should be related with each other, in which order they will be presented, etc.
- *Aggregation* maps the output of text structuring into linguistic constituents such as phrases, clauses and sentences, often by merging already related pieces of information into more concise structures.
- *Referring expression generation* determines the properties of the phrases used to identify and describe domain objects.
- *Lexicalisation* chooses the lexemes that will be used to express the terminal nodes of a linguistic constituent.
- *Surface realisation* converts the abstract representations of sentences into surface text.

The task that we are concerned with is text structuring, independently from its possible interactions with other tasks. Oversimplifying things even further, we view text structuring as the task of only *ordering* the output of content determination. Hence, a *text structure*, i.e. the output of text structuring, is merely an *order* in our view (as e.g. in Sibun 1992, Barzilay et al. 2002, Dimitromanolaki and Androutsopoulos 2003, Lapata 2003).

Additionally, in accordance with Mellish et al. (1998a) and Kibble and Power (2000) among others, text structuring (in our case simply ordering) is assumed to be a *search* task where different possible solutions are generated and evaluated according to scores assigned by a *metric*. The output of text structuring is the order which scores best among its alternatives.

1.2 Text structuring and entity coherence

The importance of text structuring arises from the fact that normally a text is not a randomly-ordered collection of information. As most literature in text linguistics argues (Halliday and Hasan 1976, Lyons 1981, De Beaugrande and Dressler 1981, inter alia), a text possesses *coherence* which is to say that the content is organised in a way that is easy for humans to read and understand.

Arguably, the easiest way to demonstrate this is by arbitrarily reordering the sentences that a text consists of. The result of this process will often be hard to comprehend, although the information content is the same before and after the reordering (Hovy 1988, Marcu 1997, Mellish and Dale 1998, Reiter and Dale 2000, among others). Consider for example the following potential answers to the question *how does the system enhance a program?* (from Hovy 1988):

(1.1) The system performs the enhancement. Before that, the system resolves conflicts. First, the system asks the user to tell it the characteristic of the program to be enhanced. The system applies transformations to the program. It confirms the enhancement with the user. It scans the program in order to find opportunities to apply transformations to the program.

(1.2) The system asks the user to tell it the characteristic of the program to be enhanced. Then the system applies transformations to the program. In particular, the system scans the program in order to find opportunities to apply transformations to the program. Then the system resolves conflicts. It confirms the enhancement with the user. Finally, it performs the enhancement.

According to Hovy (1988), example (1.1) is not as satisfactory as (1.2) because the reader has to work much harder to make sense of it. In contrast, paragraph (1.2) where the same propositions are rearranged (and linked with appropriate phrases such as “then” and “in particular”) is far easier to understand.

Text structuring is usually viewed as a genre, or even application, specific problem (Reiter and Dale 2000, p.80). Hence, the scope of this thesis must be restricted to investigating the properties of a certain type of text structure, the one that characterises a specific genre, namely descriptions of objects. Although these texts have already been studied for the purposes of *descriptive text generation* (e.g. McKeown 1985, Knott et al. 2001, O’Donnell et al. 2001), the thesis identifies and attempts to address a number of unresolved problems in the field of *descriptive text structuring*, i.e. the generation of the structure of a descriptive text.

These problems have to do with the fact that descriptive texts are often described as “entity coherent” which means that their coherence is based on the way *entities* (also known as domain objects or

concepts) are introduced and discussed in the discourse (Poesio et al. 2002). Evidence for the role of entity coherence in characterising a descriptive text structure comes from examples like the following:

(1.3) This exhibit is an amphora. Amphoras have an ovoid body and two looped handles, reaching from the shoulders up. They were produced in two major variations: type A and the type with a neck. This exhibit is a type A amphora. It comes from the archaic period.

The first sentence in this example introduces two entities in the discourse, namely the referents of the phrases "this exhibit" and "an amphora". The discourse continues with two sentences providing information about the characteristics of amphoras and their variations. Then, the current exhibit is identified as belonging to one of these variations and the discourse concludes with additional information about the current exhibit. Thus, the organisation of the text can be seen as evolving around some general patterns for introducing and discussing entities sentence after sentence.

Poesio et al. (2002) identify Chafe (1976), Kintsch and van Dijk (1978) and Givon (1983), among others, as the earliest attempts to account for some of these principles on the basis of empirical evidence. Important aspects of entity coherence are also discussed in theories of *topichood* (e.g. Reinhart 1981, Horn 1986), *givenness* (Gundel et al. 1993) and the computational theory of *focusing* in Sidner (1979) (see Poesio and Stevenson 2003 for more details).

Following the suggestions of Knott et al. (2001) and Kibble and Power (2000), we formalise entity coherence according to Centering Theory (Brennan et al. 1987, Grosz et al. 1995, Walker et al. 1998a, inter alia), a computational model that has been used quite extensively for the purposes of natural language understanding in the last 10 to 15 years, but has only recently started to attract attention within the generation community. However, viewing this model from an NLG perspective raises a series of interesting challenges that the thesis identifies and attempts to investigate to a certain extent. More specifically, in this thesis we discuss how Centering gives rise to many different metrics of entity coherence and how these metrics can be evaluated with respect to their potential usefulness for guiding the text structuring process (under the assumptions stated at the end of the previous section).

In general, entity coherence as modelled by Centering is the only characteristic of the genre in question which is investigated in significant depth in this thesis. Hence, entity coherence is isolated and assessed as the most relevant factor for descriptive text structuring, while additional constraints such as rhetorical relations (Mann and Thompson 1987) are considered only to the limited extent that the datasets available for the study allow us.

Although the investigation of the problem is unavoidably adapted to some specific features of the available datasets, the proposed methodology and the questions raised in the study are general enough to be useful for any subsequent attempt to generate a text structure in the context of an application which makes use of the notion of entity coherence. Further to this, a great amount of effort has been

placed on identifying which results are specific to the employed domain of application and which hold across the text genre in general.

In the next section, we present a chapter-by-chapter overview of the thesis which serves as an introduction to the main issues that each chapter is concerned with and shows how these contribute to answering the questions that motivate the thesis as a whole.

1.3 Chapter-by-Chapter overview

In the next two chapters, we motivate and define the scope of our research in more detail. After setting up the relevant background in chapter 2, Centering's potential for providing a solution to the generation of descriptive text structures is assessed. The chapter concludes with the formulation of the first research question that our work addresses:

Q1: How can Centering be used to define an evaluation metric of entity coherence for search-based descriptive text structuring?

Possible answers to (Q1) are presented in chapter 3. This chapter discusses possible ways of defining a metric of entity coherence, starting with an investigation of existing metrics of text structure that employ notions from Centering. Then, we define additional metrics of entity coherence based on the different formulations of Centering. We conclude the chapter with the following question that our empirical work investigates to a certain extent:

Q2: Which metrics of entity coherence constitute the most promising candidates for text structuring?

This is the main question that we deal with in the subsequent chapters of the thesis. Chapter 4 presents a psycholinguistic study that aims at testing the different predictions of three metrics using acceptability judgements. After reporting the problems we encountered in this study and commenting on the general cost of human-based evaluation, we conclude that an alternative methodology is desirable for deciding which metrics represent good candidates for the purposes of NLG, the results of which can be supplemented by subsequent human-based evaluation on a smaller scale.

Chapter 5 defines such a methodology. We present the basic aspects of a corpus-based, search-oriented evaluation task and describe the main features of SEEC, the system that was implemented to carry out our experiments. We show how each text in the corpus is used as the *Basis for Comparison* in a search-oriented evaluation which calculates the *classification rate* of each metric and compares their performance. We conclude the chapter with a discussion of our solution to the factorial complexity of the operation that this search entails.

The next two chapters report on a series of empirical studies that make use of SEEC and eight metrics from chapter 3 to investigate potential answers to (Q2). Each of these chapters is structured in a similar way, first by specifying the aims of the experiments, as motivated by the discussion in earlier chapters. Then, we present an overview of the data used and conclude with the discussion of the results of each study.

In chapter 6, we make use of GNOME (Poesio 2000), a corpus reliably annotated for the features that the Centering-based metrics make use of. The relevant subset of texts that represent the genre of interest is identified and experiments are conducted using the methodology of chapter 5. The main result of this study is that none of the employed metrics of entity coherence manages to return significantly better results than the baseline which in fact overtakes two of its competitors. The chapter also touches on the role of rhetorical coherence as a conflicting text structuring constraint in the investigated genre.

In chapter 7, we employ data from MPIRO (Dimitromanolaki and Androutsopoulos 2003), an existing NLG application, which manifest the superiority of the baseline even more emphatically. The baseline now beats all its competitors with only one exception. An examination of the structures which differentiate the baseline from the metric that overtakes it shows that the marginal difference against the baseline is due to a specific feature of the dataset from MPIRO which does not characterise the dataset from GNOME.

In chapter 8, the best scoring structures of the baseline and the metric that overtakes it are inspected more closely. This investigation equips the employed metrics with an additional constraint on entity coherence and motivates a new set of pairwise comparisons between the modified metrics. In these comparisons, a number of modified metrics overtake the baseline in the dataset from MPIRO, but not in the dataset from GNOME. Hence, a number of best-performing candidates for text structuring are identified in the particular application domain, although the baseline remains very robust as far as the genre of interest is concerned.

A question not addressed until this point is whether the results from the MPIRO domain are specific to EM (the expert who provided the orderings for the application-specific dataset) or whether they reflect more general strategies for ordering the information derived from the MPIRO system.

In order to answer this question in a general way, the dataset from MPIRO is enhanced with orderings provided by more than one expert. The distance between the orderings of EM and the orderings of her colleagues is computed and compared to the distance between the orderings of her colleagues and each other following the methods of Lapata (2003). The analysis in chapter 9 identifies another “stand-alone” expert but indicates that EM shares a lot of common ground with her other colleagues in the ordering task, deviating from them only as much as they deviate from each other.

Further to this, the methodology used to investigate the distance between the experts is extended in chapter 9 to automatically evaluate the best scoring orderings of some of the metrics from the previous chapter in comparison to an upper bound defined by the combined data of the human experts and a random baseline. This defines a subsidiary evaluation task that deals with potential shortcomings of the methodology specified in chapter 5.

The results in chapter 9 indicate that the distance between the orderings of the experts and the best scoring orderings of each metric is significantly smaller than the distance between the orderings of the experts and the orderings of the random baseline. Thus, all metrics score significantly better than the random baseline in this experiment.

The analysis also provides additional evidence in favour of the modification of the metrics suggested in chapter 8, as it is shown that the modified baseline metric stands much closer to the experts than the unmodified one. This in turn indicates that the additional constraint of entity coherence is not specific to EM but is shared by her colleagues to a great extent.

Only one of the modified metrics employed in chapter 9 manages to return a distance from the experts which is not significantly greater than the distance between the orderings of the experts and each other. Hence, this metric is identified as the one that performs best across all evaluation tasks and can be rendered as the most promising candidate for text structuring in the MPIRO domain among the ones investigated in the thesis.

Finally, chapter 10 summarises the primary results of the thesis, presents our main contributions and points out possible directions for future work.

Chapter 2

Motivation

This chapter provides a review of relevant approaches to text structuring and entity coherence, discusses their shortcomings and describes how previous research motivates our work in this field. Centering Theory, a current formal model of entity coherence, is introduced and its potential for providing a solution to descriptive text structuring is assessed. The chapter concludes with the formulation of the first research question that our work aims to address:

Q1: How can Centering be used to define an evaluation metric of entity coherence for search-based descriptive text structuring?

2.1 Major approaches to text structuring

In this section, we review the two main approaches to text structuring, namely *schemata* and *RST-based planning*. We discuss their appropriateness for the generation of descriptive text structures and how our work contributes to existing research in this field.

2.1.1 Schemata

The idea of using entity coherence in NLG is not entirely new. The TEXT system described in McKeown (1985) performs text structuring using a predefined representation called a *schema*. A schema portrays stereotypical patterns of expression, and can be seen as a template with slots at appropriate positions (called predicates by McKeown) that are filled with propositions during text structuring. The matching process is monitored by a focusing mechanism, based on the theory of immediate focus in Sidner (1979), which selects between alternative propositions.[1]

[1] As we mention in the introductory chapter, Sidner (1979) represents one of the earliest formalisations of the notion of entity coherence in computational linguistics. Her theory of immediate focus supported her attempt to specify algorithms for anaphora resolution. McKeown (1985) modified her model to constrain NLG as explained in the current section.

More specifically, during the process of schema filling more than one proposition from the relevant knowledge pool may match the next predicate in the schema. To choose between the alternatives in the schema, the system applies a set of immediate focus rules (McKeown 1985, pp.60-75). The proposition that satisfies the most preferred rule for immediate focus movement is chosen over the rest of the candidates for what to say next. Focus information is then available for the tactical component in order to control for pronominalisation and select the appropriate syntactic constructions that appear in the surface text (McKeown 1985, pp.77-79).

For each proposition that matches a predicate in the schema a *default focus* singles out the argument that is most likely to be focused on. The default focus is the first argument of the proposition to be tested for the application of the immediate focus rules. If it satisfies the most preferred rule, it is then established as the current focus and appears in the surface subject position as indicated by the unmarked syntax of the predicate.

The default focus is overridden if another argument within the proposition allows for the application of a more preferred immediate focus rule. This argument will then appear as the surface subject of the sentence by means of syntactic mechanisms such as passivisation or there-insertion which account for reordering the constituents of a sentence on the basis of focus information.

McKeown (1985) defined 4 schemata that were found to capture the structure of 56 paragraphs from 10 different authors. In order to define the schemata, McKeown analysed the paragraphs by hand into sequences of predicates. Except from the subjectivity that this methodology entails, pointed out by McKeown herself (McKeown 1985, p.25), the main practical problem with the schema-based approach to text structuring is that new schemata may need to be defined every time a system like TEXT has to be ported to a new domain, admittedly a laborious and time-consuming effort.

Duboue and McKeown (2002) present a recent attempt to define schemata automatically in order to improve the efficiency and reliability of schema-based text structuring. They describe an evolutionary algorithm that learns a schema from an aligned corpus of semantic inputs and corresponding human outputs. Duboue and McKeown (2002) use two evaluation functions in their genetic search. The first evaluation function is based on the ordering constraints acquired on their domain (Duboue and McKeown 2001). The second evaluation function computes the average of the alignment scores between the texts generated by their system and the corresponding human transcripts for a set of semantic inputs.

Other recent approaches to text structuring and multidocument summarisation have also addressed the problem of ordering information using machine learning techniques. Dimitromanolaki and Androutsopoulos (2003) apply standard machine learning algorithms in order to specify the most natural ordering of propositions derived from the database of the MPIRO system (Isard et al. 2003). Kan and McKeown (2002) use an *n*-gram model to infer ordering constraints between facts, whereas

Lapata (2003) presents an unsupervised probabilistic model for text structuring that learns ordering constraints from a large corpus operating on sentences rather than facts.

Finally, Barzilay et al. (2002) present an integrated strategy for content organisation derived from experiments asking humans to order information. In order to yield a coherent summary, the algorithm in Barzilay et al. (2002) combines the chronological order of events with a constraint which ensures that sets of sentences on the same topic occur together. This results in a bottom-up approach for text structuring that opportunistically groups topically related sets of sentences.

2.1.2 RST-based planning

As we saw in the previous section, schemata are an effective way to express frequently occurring, domain-dependent text structures that exhibit little variation. On the other hand, *Rhetorical Structure Theory* (RST) is an attempt to describe the structure of a wider variety of texts in terms of the combination of a more or less fixed set of rhetorical relations which are seen as the building elements from which coherent texts are composed (Mann and Thompson 1987). According to RST, a natural text can be described as a tree-like hierarchical structure with rhetorical relations applying recursively between adjacent spans of text as well as between larger text spans already related via a rhetorical relation. For each pair of text spans related via a rhetorical relation, RST distinguishes between the span which is more important to the writer's purpose (the *nucleus* of the relation) and the span that simply supports the nucleus (termed the *satellite*) which can often be deleted without severely impairing the comprehensibility of the text.[2]

RST-based approaches to text structuring have been very popular within the NLG literature.[3] The seminal work of Hovy (Hovy 1988, 1990, 1991, 1993) defines a dynamic top-down text planning strategy that formalises each rhetorical relation as an RST plan operator with preconditions on its nucleus and satellite as well as intended effects which express the goals and beliefs of the conversational participants. Each operator has growth points which are collections of additional goals. The planner starts by selecting an RST plan operator whose intended effects include achieving (one of) the systems' communicative goals; it then inspects which of the input propositions match the preconditions of the operator and adds them to the text structure. When the preconditions are fulfilled, the planner

[2] RST also recognises more complex linear combinations of nuclei and satellites, such as *multinuclear* relations, single relations with more than one satellite contributing to the same nucleus, etc. See Mann and Thompson (1987) for a discussion of these patterns and a wide range of RST analyses of natural texts. The principle of *nuclearity* is discussed in Mann and Thompson (1987, pp.30-38). A study on the psycholinguistic plausibility of the principle of *strict compositionality* which is Marcu's recursive formulation of nuclearity appears in Marcu (1997, Chapter 6).

[3] One of the problems caused by the popularity of RST for text generation is the proliferation of rhetorical relations. Mann and Thompson provide definitions for 23 relations although the authors are careful to say that "other relations might be reasonable constructs in a theory of text structure" (Mann and Thompson 1987, p.8, footnote 5). Indeed, Hovy and Maier (1995) taxonomise more than 400 relations that have been proposed by approximately 30 researchers into a hierarchy of around 70 increasingly semantic relations. Hovy and Maier (1995) argue that even though the taxonomy is open-ended in one dimension, it is bounded in the other and therefore does not give rise to anarchy.

tries to achieve each growth point goal by searching for an appropriate RST operator again. When the new operator is found, its preconditions are matched to the input propositions and added to the text structure. The text planning process finishes when either the input propositions are exhausted or no goals remain to be satisfied. Thus, top-down text planning results in a tree-like structure with non-crossing branches in which terminal nodes are elementary propositions and intermediate nodes correspond to RST plan operators representing discourse relations.

As Hovy himself points out (e.g. Hovy 1990, p.31 and Hovy 1991, p.94) treating growth points as "suggestions" to include additional material rather than "injunctions" as happens in his standard text planning approach makes the difference between an RST plan operator that shows more flexibility during the text planning process and one that acts like a schema. In an extension of Hovy's basic approach, Hovy and McCoy (1989) propose an architecture that combines a more flexible rhetorically-driven planner with a constraining mechanism that bars certain expansions of growth points using domain-specific focus information.

A further extension of Hovy's planning strategy for the purposes of dialogue generation is described by Moore and Paris (1993). Moore and Paris (1993) argue that for an expert system to be able to participate in a dialogue with the user it must have an explicit representation of the intentional structure of the conversation at each step. This is not provided by Hovy's operationalisation of RST relations that does not distinguish between *informational* and *intentional* rhetorical relations (Moore and Pollack 1992; Moser and Moore 1996).[4] According to Moore and Paris (1993), the mapping between a speaker's intention and an intentional relation is one-to-one. By contrast, the mapping between a speaker's intention and an informational relation is one-to-many. This means that a dialogue system requires an explicit representation of the intention that lies behind the introduction of an informational relation in the text structure in order to be able to participate in a dialogue efficiently.

Based on the distinction between the two types of rhetorical relations, Moore and Paris (1993) describe a system that preserves an explicit representation of both intentional and rhetorical structure. Using information of the intended effect of individual parts of a text on the hearer, Moore and Paris (1993) make their dialogue system capable of reasoning about its previous utterances, interpreting follow-up questions and generating appropriate explanations in the context of ongoing conversation.

Marcu (1997, Chapter 7) discusses the inability of both top-down planning and schema-based techniques to construct a text structure that subsumes all the information in the relevant knowledge pool and proposes an RST-based, bottom-up, data-driven text planning method to deal with this problem. His text structuring algorithms assume that global coherence is achieved by satisfying as many as

[4]Mann and Thompson (1987, pp.17-18) do distinguish between *subject-matter* (i.e. informational) and *presentational* (i.e. intentional) relations, but do not discuss the implications of this distinction. Although they recognise that RST "sometimes" triggers simultaneous analyses, they attribute this to the fact that the speaker occasionally tries to achieve more than one goal with a single utterance (Mann and Thompson 1987, pp.26-30).

possible of the local constraints on ordering and clustering of the nuclei and satellites of the rhetorical relations that hold between the pairs of units in the knowledge pool. Using corpus analysis to calculate the weights of these constraints for each rhetorical relation, Marcu (1997) is able to construct hierarchical text structures which not only subsume the whole information in the knowledge pool but also satisfy multiple high-level communicative goals.

Despite its popularity, the appropriateness of RST-based text structuring has been challenged in a number of domain-specific applications: Kittredge et al. (1991) argue in favour of representing domain-specific communication knowledge explicitly, albeit complementary to domain-independent rhetorical knowledge, for the generation of short reports such as weather forecasts and employment statistics summaries. Mooney et al. (1991) argue that the high-level structure of extended explanations is not completely recursive as RST has claimed and propose a bottom-up process for generating this sort of structure in terms of basic blocks each consisting of a domain-specific organisational focus and textual units clustered around this focus. Once the high-level structure is determined by the bottom-up strategy, it can be used to control the top-down generation of local plans within the resulting block structure. Moreover, Sibun (1992) describes a system that does not use any top-down guidance from explicit knowledge of overall hierarchical structure (be it domain or rhetorical). The approach in Sibun (1992) relies largely on local regularities in the structure of the knowledge base which makes it possible to generate the best next increment of text without any tree-like structure.

More recently, Bouayad-Agha et al. (2000) questioned the compatibility between rhetorical structure and text structure as assumed by Scott and de Souza (1990), by presenting examples of acceptable text structures which are not compatible with their underlying rhetorical structures. Bouayad-Agha et al. (2000) attribute this incompatibility to a phenomenon that they call *extraposition* and discuss how an NLG system can be configured in order to improve the quality of the generated text by allowing solutions that violate compatibility.

In conclusion, although RST appears to be a more domain-independent framework than McKeown's schemata for describing the structure of a text, neither schemata nor RST-based text planning are able to generate text structures across all genres. In fact, it has often been argued that it is difficult to deal with text structuring within a unified general model due to the inherent domain-dependence of the task (Reiter and Dale 2000). On the other hand, domain-independent approaches that can be used in as many different settings and applications as possible would be particularly appealing due to their generality and potential portability.

In the next section, we discuss how certain structural properties of descriptive texts, which is our own genre of interest, have motivated an attempt to identify **which** component of a general theory such as RST fails to make the appropriate predictions about the role of entity coherence. Further to this, an effort to replace this component with a more elegant general mechanism that accounts for the

observed phenomena is desirable. This gives rise to the question whether a general theory of entity coherence can supplement an already suggested refinement of RST.

2.1.3 An ELABORATIONless framework of descriptive text structure

In a standard RST analysis of a descriptive text, most of the material appears to be related via a specific kind of rhetorical relation called ELABORATION. In general, ELABORATION has been characterised as "the weakest of all rhetorical relations in that its semantic role is simply one of providing more detail" (Scott and de Souza 1990, p.60). Thus, it is somehow surprising that ELABORATION turns out to be the most frequent rhetorical relation in the corpus analysis of Marcu (2000, p.438). As Cheng (2002, p.157) notices, because this relation can have a very large number of possible expansions, its predominance in the descriptive genre makes it hard for a top-down planner based on growth points to determine which specific information to select from the knowledge base.

In a further analysis motivated by the structural properties of the descriptive genre, Knott et al. (2001) identified a number of additional general theoretical problems in the RST framework all related to OBJECT-ATTRIBUTE ELABORATION.[5] One of their main observations is that ELABORATION does not hold between two propositions directly in the same way as the rest of the RST relations. Rather, it holds indirectly by virtue of an identity relation between two entities that both propositions refer to. Consequently, it is better thought of as expressing constraints on how the focus of a text moves from one entity to another.

Knott et al. (2001) suggest that ELABORATION be eliminated from the group of RST relations and replaced by a theory of entity coherence. The main operational unit in the suggested framework of text structure is the *entity chain*. An entity chain consists of a sequence of local *RS-trees* connected with each other linearly via subsequent entity links. These trees can either consist of only one proposition (in which case they are called trees purely by convention), or have additional levels of hierarchy, thus being complex trees each constructed using RST's relations (minus ELABORATION). The *top nucleus* of a tree is the proposition which is reached by following a chain of nuclei from its root. In other words, the top nucleus is the nucleus of the nucleus of (...) the nucleus of the tree.

A *legal entity chain* C_n is one where the top nucleus of each local RS-tree that the chain consists of is a proposition "about" the same entity E_n.[6] Hence, each legal entity chain C_n sets a unique entity E_n as its global focus.[7] Crucially, the propositions within a single RS-tree do not all have to be about

[5] The objections of Knott et al. (2001) are specific to OBJECT-ATTRIBUTE ELABORATION and not to the other types of ELABORATION such as PROCESS-STEP ELABORATION (see Mann and Thompson 1987, p.52). OBJECT-ATTRIBUTE ELABORATION applies between two spans when the nucleus mentions an entity and the satellite subsequently "presents additional detail about" this entity. Unless otherwise stated, subsequent mentions of ELABORATION in this text should be taken to refer to OBJECT-ATTRIBUTE ELABORATION only.

[6] A working definition of what it means for a proposition to be "about" a certain entity according to Knott et al. (2001) is provided in the next section.

[7] Knott et al. (2001) take an entity chain to correspond to a *focus space* as discussed by Grosz and Sidner (1986).

the same entity. Coherence between these propositions is not determined by their having entities in common, but by the rhetorical relations between them.

At a higher level of discourse structure, a coherent descriptive text is defined as a *legal sequence of entity chains*. A legal sequence of entity chains is a sequence in which the global focus E_n of a chain C_n is mentioned in any RS-tree within the i previous chains.[8] Hence, the admissibility of C_n with focus E_n at a particular point in a text is seen as a function of the linear distance of C_n from the previous mention of E_n. Also note that the position of the RS-tree where the previous mention of E_n took place does not matter.

As a result, the constraints of global coherence in this framework are much weaker than applying ELABORATION recursively to build a tree-like structure. This accounts for a phenomenon that Knott et al. (2001) call *resumption* which violates RST's assumption of hierarchical text structure. Knott et al. (2001) present an example of resumption and define it pretheoretically as the move where an entity mentioned in the middle of a paragraph becomes the central topic in a subsequent (not necessarily adjacent) paragraph.[9]

In summary, Knott et al. (2001) are making three main claims with respect to the structure of descriptive texts:

- Rhetorical relations apply only **locally**: local RS-trees are related linearly with each other via entity links.
- Rhetorical and entity coherence are **not simultaneous** constraints on text structure (as assumed e.g. by Hovy and McCoy 1989): Two adjacent propositions are related coherently if either there is a rhetorical relation between them or they have an entity in common.
- ELABORATION **overlaps** with theories of global and local entity coherence: Entity links within a chain account for the application of ELABORATION on adjacent text spans. Subsequent entity chains are not related hierarchically but via loose constraints on their global foci (similarly to Mooney et al. 1991).

In the next two sections, we discuss how the framework of text structure in Knott et al. (2001) has been used in a system that generates text describing artefacts in a virtual museum. We conclude our review of the main approaches to text structuring with a discussion of the underspecification of local entity coherence in the ELABORATIONless model and how this motivates the investigation of Centering Theory as a potential solution.

Hitzeman and Poesio (1998) argue that relating each focus space to a single entity is needed to account for long distance pronominal anaphora, a suggestion supported by the analysis of long distance pronominalisation in Poesio et al. (2002).

[8] Knott et al. (2001) set i to 4 by convention.

[9] Another example of resumption appears in Kittredge et al. (1991). Also see example (2.1) in the next section.

2.1.4 Structuring descriptive text

Mellish et al. (1998b) argue that describing an object such as a museum exhibit lacks an explicit unitary overriding communicative goal. The meta-level intention behind the structure of this kind of text can be stated generally as "provide the hearer with suitable, unknown, interesting, and accurate information about an object in the gallery". This is better analysed as a set of non-hierarchical descriptive goals each providing an opportunity for other goals to be executed. This setting makes both a standard top-down approach to text planning based on goal decomposition and schema-driven text generation unsuitable for the descriptive domain. Mellish et al. (1998b) describe an opportunistic approach to text generation that structures important, untold and non-trivial interconnected pieces of information from a precompiled knowledge base at runtime. This strategy was implemented in the context of a generation system called ILEX that delivers hypertext descriptions of artefacts in a virtual museum (O'Donnell et al. 2001).

The domain knowledge of ILEX is organised in the form of a directed acyclic graph called the *content potential*. The content potential consists of three kinds of nodes: entity nodes (each corresponding to a generic or specific domain object), fact nodes each linked to two entity nodes (corresponding to an instantiated binary predicate with the two entities filling the argument positions)[10] and relation nodes which are linked to fact nodes (each representing a rhetorical relation between two facts). In accordance with the claims of Knott et al. (2001), the rhetorical relations in the content potential do not include ELABORATION. Given the ubiquity of ELABORATION in the descriptive domain, the relations in ILEX's content potential are a small subset of the complete set of rhetorical relations defined by RST.

When ILEX receives a request from the user to describe an object in the collection, it sets this object as the focal entity of the description which will always consist of a single page of hypertext.[11] Then, the content determination algorithm extracts a set of *relevant* facts from the content potential (see O'Donnell et al. 2001, pp.243-244 for more details on how relevance is computed). The text structuring stage starts with the system extracting a subgraph of the content potential based on the facts that were delivered from the content determination module. This subgraph contains all the entity nodes pointed to by the relevant facts together with all the relation nodes which link pairs of fact nodes. Once the subgraph has been obtained, the text structuring problem is the same as the one discussed by Marcu (1997) with the exception that the representation that ILEX strives for is a legal sequence of entity chains according to the framework in Knott et al. (2001) rather than a recursive RST tree.

[10] These entities are annotated as *Arg1* and *Arg2* of the predicate. Arg1 is used as the working definition of the entity that the fact "is primarily about" in Knott et al. (2001).

[11] In chapter 8, we shall use the term *PageFocus* to refer to this entity. A particular instantiation of the PageFocus results into the selection of the relevant information from ILEX's database.

In order to generate this representation, the system proceeds in two independent directions (O'Donnell et al. 2001, p.246). First, it builds a number of entity chains, grouping together the fact nodes that have the same Arg1. Second, it searches for the best RS-tree that can be created from the complete set of fact nodes, regardless of which entities they are about.[12] After the RS-tree is built, the fact nodes that make it up are deleted from their entity chains and the fact node that corresponds to the top nucleus of the tree is replaced by the whole tree in its entity chain. The algorithm then checks if a legal sequence of entity chains exists; if it does not, it tries the same procedure on the next best tree. The whole process is iterated and the set of trees containing facts not yet incorporated into the tree is produced. The algorithm finishes when no more trees can be added to the legal sequence of entity chains.

In short, what the text structuring component of ILEX does is to "saturate" an initially constructed set of entity chains with local RS-trees. Note that in the discussion of the ELABORATIONless framework of text structure, Knott et al. (2001) claim that either entity links or rhetorical relations are sufficient for establishing coherence between two text spans. However, the description of the text structuring algorithm in O'Donnell et al. (2001) prioritises the construction of as many "good" trees as possible which are subsequently linked to each other and to stray facts with Arg1 links, probably because rhetorical relations are seen as more interesting features of text structure than plain entity links. This procedure can give rise to the following type of text (taken from Knott et al. 2001, p.192, example 6). The example is annotated with the primary constructs of global text structure in ILEX. Satellites of local RS-trees are pointed out by additional levels of indentation:[13]

(2.1) (C_1, E_1: `J-999`)

a. This piece is a necklace.

b. It was designed by a jeweller called Jessie King.

c. It was designed in 1905.

d. It is made of silver.

(C_2, E_2: `King`)

e. Jessie King was a famous designer.

fS. She was Scottish,

f. but she worked in London.

[12] The details of this process are not documented substantially. The evaluation of an RS-tree appears to be based on intuitive preferences for some relations over others and a preference for "bushy" trees without self-expanding relations.

[13] For instance, the first RS-tree of example (2.1) consists of the satellite (fS) followed by its top nucleus (f).

g. It was in London that this piece was made.

(C_3, E_3: `Arts-and-Crafts-style`)

hS. Like the previous piece,

h. this piece is in the Arts-and-Crafts style.

iS. Although the previous piece had a simple shape,

i. Arts-and-Crafts style jewels tend to be elaborate;

iS. for instance, this piece has detailed florals.

There are three entity chains in this text according to Knott et al. (2001). The first entity chain C_1 consists of the fact nodes (2.1a) to (2.1d). Its focus E_1 is the focal object of the whole description, that is, the particular exhibit `J-999` (as recorded in the content potential). The focus of the second entity chain consisting of spans (2.1e-g) is `King`, the jewel's designer. The last entity chain C_3 sets the entity `Arts-and-Crafts-style` as its focus.

Within the chains there exist a number of local RS-trees. The first tree consists of the top nucleus (f) and its satellite (fS). The top nucleus of the second tree is fact node (h). The third tree consists of the top nucleus (i) and one satellite on each side.

According to Knott et al. (2001) there are two resumption relations in (2.1), from C_2 to C_1 and from C_3 to C_1.[14] Neither of these resumptions is to material in an adjacent text span. Nevertheless, Knott et al. (2001) are satisfied that the text in (2.1) represents a good optimisation of entity- and relation-based constraints on coherence.

2.1.5 A search-based approach to descriptive text structuring

In the seminal paper of Mellish et al. (1998a), text structuring is viewed as a formal *search* problem. Broadly speaking, in search-based text structuring a number of potential solutions are generated and emphasis is placed on defining a *scoring function* which assigns each solution with a measure of "goodness". The scoring function is an essential part of the *evaluation metric* that compares the solutions and picks the one which scores best.[15]

In the experiments of Mellish et al. (1998a), a text structure is represented as an ordered RS-tree with propositions at its leaves. A stochastic search approach to text structuring in the ILEX domain

[14] Arguably, the entity `Arts-and-Crafts-style`, which is recognised as the focus of the chain C_3 by Knott et al. (2001), is not mentioned in any of the previous two entity chains, so it is difficult to see C_3 as a legal entity chain introduced via a resumption. In addition, the Arg1 of the top nucleus (h) is clearly `J-999`, the referent of the NP "*this piece*". However, the RS-tree of (h) is included in a chain that sets the global focus to the `Arts-and-Crafts-style`.

[15] A more formal definition of the terms *scoring function* and *evaluation metric* is given in the next chapter.

is introduced that maintains a population of candidate solutions which evolves according to genetic rules of selection, recombination and mutation. At each evaluation cycle, new candidate structures are created by applying these rules, and subsequently assigned scores by an ad-hoc function which takes into account factors such as interestingness, substructure size and fulfilled preconditions for rhetorical relations as well as some intuitively specified features of entity coherence. Then, the new structures replace the least fit individuals in the population. This process is not guaranteed to find the optimal solution, but can be stopped at any point during the generation process and output the best structure "found so far". An extension of Mellish et al. (1998a) which accounts for the interaction between aggregation and text structuring is presented in Cheng (2002).

The functions in Mellish et al. (1998a) and Cheng (2002) are devised in accordance with the model of text structure in Knott et al. (2001).[16] The main difference between the approaches of Cheng (2002) and Mellish et al. (1998a) and the algorithm in O'Donnell et al. (2001) is that the former are much less deterministic in nature, taking into account a much larger space of possible, but not equally plausible, candidate structures.

Another search-based approach to text structuring is followed in ICONOCLAST, a system which generates medical leaflets (Kibble and Power 2000). Although the set of candidate solutions enumerated by Kibble and Power (2000) appears to be much more restricted than in Mellish et al. (1998a), both approaches evaluate a population of text structures according to the values of an intuitive scoring function. Because the formalisation of entity coherence in Kibble and Power (2000) is extremely relevant to our purposes, as it will soon become obvious, it is discussed in subsequent sections of the thesis in substantial detail. At this point, we would simply like to emphasise that a search-based approach as represented by the work reviewed in this section is a plausible alternative to the more deterministic methods traditionally used for descriptive text structuring.

2.1.6 Centering: the missing entity link?

To our understanding, McKeown (1985) and Knott et al. (2001) argue in favour of using **general** notions of entity coherence to guide the generation process.[17] As we mentioned earlier, although focus information was used in schema-driven text generation to constrain what to say next, schemata are not flexible enough to express the opportunism in the domain of ILEX. The same holds for top-down text planning including the variations in Hovy and McCoy (1989) and Mooney et al. (1991) which employ domain-specific notions of focus as additional constraints to rhetorical coherence. By contrast, Knott

[16] Note that the features for entity coherence carry less weight than rhetorical features in Mellish et al. (1998a), similarly to the strategy followed in O'Donnell et al. (2001). It is not clear to us how this function accounts for resumptions. The features for entity coherence employed by Mellish et al. (1998a) and Cheng (2002) are discussed in section 3.1.1 of the next chapter in more detail.

[17] The same is true for Kibble and Power (2000) as section 2.4.3 discusses.

et al. (2001) argue in favour of exploiting a general theory of entity coherence alongside independent, locally applying rhetorical relations in the generation of a descriptive text structure.

However, as Knott et al. (2001) state in their conclusion, the structure within and between entity chains is underspecified. Hence, a formally defined model of entity coherence might be necessary to further constrain a descriptive text structure. Although *Centering Theory* is identified as one of the models of entity coherence that can be possibly used in the context of descriptive text structuring, whether Centering can indeed serve this purpose remains an open question.

In order to assess whether Centering is compatible with the framework of text structure which underlies ILEX, in the next section we describe the basic aspects of the theory, the ways it was evaluated and how it was used for NLG so far. Then, the preferences which underlie Centering are compared with the framework of Knott et al. (2001) in section 2.4.4. Subsequent and more thorough investigation of these preferences leads to the conclusion that the most appropriate way to incorporate Centering into text structuring is by defining scoring functions of entity coherence which are paramount to search-based text structuring as presented in the previous section.

2.2 Centering Theory

In this section, we discuss the basic aspects of Centering Theory (henceforth CT). First, we present the general claims of the theory and its formalisation in the seminal papers of Brennan et al. (1987) and Grosz et al. (1995), and then we discuss some more recent formulations. Next, we review how CT was used by its various proponents, with particular reference to a recent corpus-based evaluation which points out and addresses CT's underspecification (Poesio et al. 2002).[18] We conclude by discussing how specific NLG applications used CT with particular reference to Kibble and Power (2000).

The material which is covered in these sections should be enough for the reader to gain an overview of the various aspects of CT in order to be able to follow our discussion of the model throughout the subsequent chapters of the thesis. For more details on CT the reader is referred to the work cited in this section, especially Brennan et al. (1987), Grosz et al. (1995), the collection of papers in Walker et al. (1998b), the evaluation of CT by Poesio et al. (2002) and the original paper of Kibble and Power (2000).

2.2.1 Basic terminology

CT is a simple entity-oriented theory of text coherence. Grosz et al. (1983, 1995) defined CT as a model of some aspects of immediate focus (Sidner 1979). It is assumed that discourses are composed of discourse segments (Grosz and Sidner 1986), each of which consists of a sequence of utterances.

[18] Clearly, this underspecification is different from the one discussed in the previous section.

Each segment is represented as a part of a discourse model. *Centers* are semantic objects that are part of the discourse model for each utterance in a discourse segment. The centers are evoked and subsequently referred to by some of the NPs in each utterance and correspond to discourse entities in the sense of Webber (1978) or Kamp and Reyle (1993).[19]

Each utterance U_n in a given discourse segment is assigned a *list of forward-looking centers*, denoted as CF(U_n), and a unique *backward-looking center*, the CB(U_n). The CF(U_n) represents a partial ranking of the discourse entities evoked or referred to by the NPs in U_n in order of prominence. The *preferred center*, CP(U_n), is the most highly ranked member of CF(U_n), whereas the CB(U_n) represents the discourse entity that U_n is most centrally concerned with. As a result, the CB(U_n) corresponds to the immediate center of attention, similar to what is elsewhere called the *topic* (e.g. Reinhart 1981, Horn 1986).

The CB(U_n) links the current utterance to the previous discourse. The ranking imposed on the elements of CF(U_n) reflects the assumption that the preferred center, CP(U_n), will most likely be the CB(U_{n+1}). The most highly ranked element of CF(U_n) that is finally realised in U_{n+1} is the actual CB(U_{n+1}). Obviously, segment-initial utterances lack a CB.

Grosz et al. (1995) define the CB(U_n) as being strictly local: The choice of a backward-looking center for an utterance U_n is from the set of forward looking centers of the previous utterance U_{n-1} while the forward-looking centers of U_{n-1} depend only on the discourse entities that constitute U_{n-1}. In other words, CB(U_n) cannot be from CF(U_{n-2}) or other prior sets of forward-looking centers.

2.2.2 Formalisation

The distinction between looking back to the previous discourse with the CB(U_n) and projecting preferences for interpretation in subsequent discourse with the CP(U_n) is a key aspect of CT. Based on this distinction, CT defines four transition relations across pairs of adjacent utterances. The typology of transitions (from Walker et al. 1998a, p.6 and Walker et al. 1994, p.200), presented in Table 2.1, is based on two factors: whether the backward-looking center, CB, is the same from U_{n-1} to U_n, and whether the CB(U_n) is the same as the CP(U_n).[20]

The formal system of constraints and rules in CT (as they appear in Brennan et al. 1987 and Walker et al. 1998a, pp.3-4) is as follows:

[19] Although Grosz et al. (1995, p.209, footnote 6) allude that "events and other entities that are more often directly realised by VPs can also be centers", the only work towards that direction is Kameyama et al. (1993). Also note that, as Kibble (2001) and Poesio et al. (2003) emphasise, the various formulations of CT do not appear to explicitly acknowledge any other factor as being relevant to local coherence.

[20] Note that the formulation of CT in Grosz et al. (1995) defines only one SHIFT transition using only the condition CB(U_n)≠CB(U_{n-1}). Brennan et al. (1987) named SMOOTH-SHIFT "Shifting-1" and ROUGH-SHIFT "Shifting". The more figurative names come from Walker et al. (1994). "CB(U_{n-1}) undef" in Table 2.1 stands for the cases where U_{n-1} does not have a CB (also see section 2.2.4).

	$CB(U_n)=CB(U_{n-1})$ or $CB(U_{n-1})$ undef	$CB(U_n)\neq CB(U_{n-1})$
$CB(U_n)=CP(U_n)$	CONTINUE	SMOOTH-SHIFT
$CB(U_n)\neq CP(U_n)$	RETAIN	ROUGH-SHIFT

Table 2.1: Centering transitions

For each utterance U_n in a discourse segment D consisting of utterances U_1, ... , U_m:

Constraints

C1. There is precisely one $CB(U_n)$.

C2. Every element of $CF(U_n)$ must be realised in U_n.

C3. The $CB(U_n)$ is the highest-ranked element of $CF(U_{n-1})$ realised in U_n.

Rules

R1. If any element of $CF(U_{n-1})$ is realised by a pronoun in U_n, then the $CB(U_n)$ must be realised by a pronoun also.

R2. Transition states are ordered. CONTINUE is preferred to RETAIN, which is preferred to SMOOTH-SHIFT, which is preferred to ROUGH-SHIFT:
CONTINUE>RETAIN>SMOOTH-SHIFT>ROUGH-SHIFT

Rule 1 is often called the pronoun rule. It captures the intuition that the $CB(U_n)$ is often pronominalised.[21] According to Rule 1 no element from the previous utterance can be realised by a pronoun in an utterance unless the CB is realised by a pronoun too. In other words, if there are multiple pronouns in an utterance, realising discourse entities from the previous utterance, then one of these pronouns must realise the CB. In addition, if there is only one pronoun realising entities from U_{n-1}, then this pronoun must be the CB.

Rule 2 claims that some transitions between utterances are more coherent than others by stipulating that these transitions are preferred over others. Measuring coherence is based on an estimate of the hearer's inference load, relative to other choices the speaker had as to how to realise the same propositional content. The most fundamental claim of CT is that if a discourse adheres to the rules and constraints of CT, its coherence will increase and the inference load placed upon the hearer will decrease.

[21] Or deleted in languages like Japanese that allow for zero pronouns. The correspondence between unstressed pronouns in English and null pronouns in languages such as Japanese was established first by Kameyama (1985, 1988). Rule 1 has been extended directly to zero pronouns in Japanese (Walker et al. 1990, 1994), Yiddish (Prince 1994), Turkish (Turan 1995) and Italian (Di Eugenio 1990), among other languages.

The combination of constraints, transition states, and rules makes a set of testable predictions about which interpretations hearers will prefer because they require less processing. Maximally coherent segments are those that require less processing time. For example, a CONTINUE followed by another CONTINUE should require the hearer to keep track of only one main discourse entity, which is currently both the CB and the CP. As a result, discourses that CONTINUE centering the same entity are claimed to be more coherent than those that repeatedly SHIFT from one center to another. Moreover, a single pronoun realising an entity from the previous utterance is the current CB by Rule 1 and can often be interpreted to co-specify the discourse entity realised by CP(U_{n-1}) in one step.

2.2.3 Example discourses

In this section, we present two discourses annotated with CT's data structures and transitions in order to show how the rules and constraints of CT apply to these examples. First, let us consider the following discourse (from Walker et al. 1998a, pp.6-7) where (2.2c) and (2.2c') are alternatives for the third utterance:

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

`CF(Jeff, Dick, car)`

b. He washed the windows as Dick waxed the car.

`CF(Jeff, windows, Dick, car)`

`CB=Jeff, CONTINUE`

c. He soaped the pane.

`CF(Jeff, pane)`

`CB=Jeff, CONTINUE`

c'. He buffed the hood.

`CF(Dick, hood)`

`CB=Dick, SMOOTH-SHIFT`

Walker et al. (1998a) assume the standard CT ranking based on grammatical function in order to compute the CF lists in (2.2). According to this, discourse entities realised in subject position rank more highly than entities realised in object position which are more highly ranked that entities coming from NPs in subordinate clauses or NPs with other grammatical functions.[22]

The first utterance of the discourse does not have a CB by definition. According to the definition of the transitions in Table 2.1, utterance (2.2b) is annotated as a CONTINUE since the CB(2.2a) is

[22] A more precise definition of this ranking appears in Miltsakaki (2002). Section 6.3.3 of chapter 6 discusses the CF ranking in Miltsakaki (2002) in more detail.

undefined and the CB(2.2b) equals the CP(2.2b).[23] When the third utterance in the discourse is (2.2c), it is marked as a CONTINUE transition since Jeff, the CB(2.2c), is the same as the CB(2.2b) as well as the same as the CP(2.2c). In contrast, (2.2c') is a SMOOTH-SHIFT transition, because the CB(2.2c') has changed from the CB(2.2b) although it is the same as the CP(2.2c').

CT predicts that the transition in (2.2c') is less coherent that the one in CB(2.2c). Since (2.2b) is a CONTINUE with the discourse entity Jeff as the CB, the speaker is taken to indicate an intention to talk about the same entity in the subsequent utterance. Indeed, this happens in (2.2c) which continues centering on Jeff. By contrast, despite the indicated intent in (2.2b), the speaker starts talking about Dick in (2.2c'). An indication for the predicted preference of a CONTINUE over a SMOOTH-SHIFT comes from the way that the hearer interprets the pronoun "he" in (2.2c'): "he" is first taken to refer to the CP(2.2b), that is Jeff, but when the hearer processes the rest of (2.2c') she has to revise this interpretation and resolve the pronoun to Dick because the verb "buffed" can only be related to the waxing event. According to Walker et al. (1998a), the combination of the SMOOTH-SHIFT and the use of a pronominal form to realise a new center in (2.2c') are factors that contribute to making (2.2c') a less coherent transition than (2.2c).[24]

In order to postulate the preference of a SMOOTH-SHIFT over a ROUGH-SHIFT for the purposes of pronominal resolution, Brennan et al. (1987) discuss the following (now classic) example, read with the pronouns in (2.3d) destressed:

(2.3) a. Brennan drives an Alpha Romeo.

CF(Brennan, Alpha-Romeo)

b. She drives too fast.

CF(Brennan)

CB=Brennan, CONTINUE

c. Friedman races her on weekends.

CF(Friedman, Brennan, weekends)

CB=Brennan, RETAIN

d. She often beats her.

CF(???, ???)

CB=Friedman, SHIFT

Utterance (2.3d) is characterised by the fact that it achieves a SHIFT. According to Constraint 3, the SHIFT is inevitable because the CP(2.3c) Friedman is realised in (2.3d) as one of the two pronouns,

[23] Different suggestions for the transition in (2.2b) are discussed in section 2.2.4.1.

[24] Hudson-D'Zmura and Tanenhaus (1998) provide experimental evidence that the "garden path" in (2.2c') corresponds to an increase in processing time and a participant's judgement that the discourse concluding with (2.2c') makes less sense than the one finishing with (2.2c).

thus being the CB(2.3d). The question is which of the two pronouns in (2.3d) realises the CB. Note that the formulation of transitions in Grosz et al. (1995) fails to make a choice since their definition of SHIFT does not consider whether the CB(U_n) equals the CP(U_n). Brennan et al. (1987) propose the preference of a SMOOTH-SHIFT over a ROUGH-SHIFT which enables them to successfully bind the pronoun "she" to `Friedman` as shown below:[25]

(2.4) She often beats her.

d. She:`Friedman`, her:`Brennan`

`CF(Friedman, Brennan)`

`CB=Friedman,` SMOOTH-SHIFT

d'. She:`Brennan`, her:`Friedman`

`CF(Brennan, Friedman)`

`CB=Friedman,` ROUGH-SHIFT

2.2.4 Underlying notions of Centering

In this section, we present a more recent analysis of CT into the prerequisite of CONTINUITY and three underlying principles, namely COHERENCE, SALIENCE and CHEAPNESS which was claimed to further simplify CT.[26] Then, we discuss how CT has been used and evaluated by its numerous proponents. We conclude the review of CT with a discussion of its use in the context of NLG.

2.2.4.1 Continuity

Constraint 1 of standard CT can be taken to presuppose that *each utterance in the discourse refers to at least one entity in the utterance that precedes it*. Arguably, this requirement can be seen as a *prerequisite* for the computation of the standard CT transitions in Table 2.1. The definition of the prerequisite of CONTINUITY in terms of CT is as follows:

[25] Walker (1989) performed a manual evaluation of the algorithm for the resolution of pronominal anaphora in Brennan et al. (1987) (known as the BFP algorithm) on a corpus of 281 sentences distributed over texts from 3 genres in comparison with the algorithm in Hobbs (1978). She reports an accuracy of 77.6% for BFP and 81.8% for Hobbs. The main problem with the BFP algorithm is that it can be used directly to resolve **only one** pronoun, that is, the one that is taken to refer to the current CB. However, corpus-based studies have shown that a) many CBs are not pronominalised and b) many non-CB referents are pronominalised (Henschel et al. 2000). Although this is not strictly incompatible with the mainstream definition of Rule 1 (as defined in section 2.2.2), in quite a few cases it makes pronoun resolution according to the BFP algorithm not possible. Hence, subsequent work on CT-based pronoun resolution revised the BFP algorithm substantially (e.g. Strube 1998; Strube and Hahn 1999; Tetreault 2001; Miltsakaki 2002). A critical discussion of the BFP algorithm appears in Kehler (1997).

[26] As the title of the section suggests, we refer collectively to the three underlying principles and their prerequisite as the *underlying notions* of CT.

(2.5) CONTINUITY:

$$Cf(U_{n-1}) \cap Cf(U_n) \neq \emptyset$$

Grosz et al. (1995) do not discuss the effects of violations of Constraint 1 in the coherence of discourse. Kibble and Power (2000, Figure 1) define the additional transition NOCB for the second member of a pair of utterances that do not have any entity in common, suggesting that a NOCB can be considered to be the worst transition causing the highest degradation of entity coherence. Miltsakaki and Kukich (2000b), however, consider the NOCB transition to be a type of ROUGH-SHIFT.

In an attempt to distinguish between different kinds of NOCBs, Di Eugenio (1998, p.128) uses the term CENTER ESTABLISHMENT for an utterance without a CB that corresponds to a global focus shift or contains an entity coreferring with an entity in U_{n-2} when U_{n-1} is an adjunct. Moreover, in Poesio et al. (2002, p.28) the transition that connects two utterances without a CB is called NULL (also in Passoneau 1998), whereas the transition from an utterance with a CB to an utterance that does not have one is called ZERO.

We remind the reader that the inverse case, that is, where U_{n-1} does not have a CB but $CB(U_n)=CP(U_n)$, is classified as a CONTINUE or a RETAIN by Walker et al. (1998a).[27] The additional transition ESTABLISHMENT is often used to refer to such an utterance, which has a CB itself but follows a NOCB transition (e.g. in Kameyama 1998 and Poesio et al. 2002).

2.2.4.2 Coherence and Salience

As Table 2.2 shows, the table of transitions in Brennan et al. (1987) can be rephrased in terms of two general *principles* (Kibble 2001; Beaver 2003). We refer to the first of these principles, i.e. the requirement that $CB(U_n)=CB(U_{n-1})$, as the principle of COHERENCE and to the second one, that is the requirement that $CB(U_n)=CP(U_n)$, as the principle of SALIENCE.[28]

Beaver (2003) and Kibble (2001) notice that ranking COHERENCE over SALIENCE (denoted as COHERENCE>SALIENCE) is a simpler way of stating the preferences over transitions in Rule 2 (that is, CONTINUE>RETAIN>SMOOTH-SHIFT>ROUGH-SHIFT). This is evident from CT's preference of a RETAIN over a SMOOTH-SHIFT. Since a RETAIN only violates COHERENCE and a SMOOTH-SHIFT only violates SALIENCE, the preference of a RETAIN over a SMOOTH-SHIFT is an indirect

[27] See the definition of transitions in Table 2.1 of section 2.2.1.

[28] Kibble (2001) uses the term COHESION instead of COHERENCE for the first of these principles. In traditional text linguistics (e.g. Halliday and Hasan 1976), the term cohesion often refers to the surface cues that communicate several aspects of the discourse structure. Lyons (1981) and De Beaugrande and Dressler (1981) use the term coherence to express the logical consistency of utterances at the content level. Coherence is the term employed here as well since it operates closer to the text structuring task than cohesion. The principles of COHESION and SALIENCE in Kibble (2001) are the same as the constraints COHERE and ALIGN used by Beaver (2003) for his reformulation of the BFP algorithm in terms of Optimality Theory (see section 2.5.2). The relation between COHERENCE and the existence of a NOCB transition in U_{n-1} is investigated in section 3.5 of the following chapter in more detail.

	COHERENCE: $CB(U_n)=CB(U_{n-1})$	COHERENCE∗: $CB(U_n)\neq CB(U_{n-1})$
SALIENCE: $CB(U_n)=CP(U_n)$	CONTINUE	SMOOTH-SHIFT
SALIENCE∗: $CB(U_n)\neq CP(U_n)$	RETAIN	ROUGH-SHIFT

Table 2.2: COHERENCE and SALIENCE in the table of Centering transitions

way of stating that violating COHERENCE is more serious than violating SALIENCE. More generally, reformulating the preferences of Rule 2 directly in terms of the underlying principles instead of the set of transitions is argued to make the CT model simpler and more transparent:

> Given *k* binary constraints it is possible to define *k*! possible rankings for the constraints themselves. If transitions are used instead of constraints, at least 2^k! possible rankings can be defined.
>
> (Beaver 2003, pp.9-10 and p.16)

2.2.4.3 Cheapness

A reformulation of CT, named *Functional Centering* (FC), is defined in Strube and Hahn (1999). Strube and Hahn (1999) introduce the principle of CHEAPNESS in order to improve the way that standard CT accounts for certain cases of pronoun anaphora. CHEAPNESS is defined as follows:

(2.6) CHEAPNESS:

$$CB(U_n) = CP(U_{n-1})$$

In an attempt to specify the set of CT transitions more precisely, Strube and Hahn (1999, Table 20) use the principle of CHEAPNESS to define two additional transitions. According to the "revised" table of transitions in FC shown in Table 2.3, the definitions of CONTINUE and SMOOTH-SHIFT are extended since these transitions are also required to satisfy the principle of CHEAPNESS. Two corresponding transitions which violate the principle of CHEAPNESS are labelled EXPENSIVE CONTINUE and EXPENSIVE SMOOTH-SHIFT.[29]

However, the transitions in Table 2.3 are not central in FC. This is clear by the redefinition of Rule 2 in FC which now favours *cheap transition pairs*, thus giving total priority to the principle of CHEAPNESS over the other two underlying principles of CT:[30]

[29] Note that Strube and Hahn (1999) do not apply CHEAPNESS on the definitions of RETAIN and ROUGH-SHIFT, without an obvious explanation. See section 3.5.1 of the next chapter for more detailed discussion of this issue.

[30] As we mention in section 2.3.1, although the formulation of Rule 2 in FC refers to pairs of transitions, it can be simplified as applying to triples or even pairs of utterances. Not surprisingly, a pair of utterances $< U_{n-1}, U_n >$ is cheap if the $CB(U_n)$ is correctly predicted by the $CP(U_{n-1})$, i.e. $CB(U_n)=CP(U_{n-1})$.

	$CB(U_n)=CB(U_{n-1})$ or $CB(U_{n-1})$ undef	$CB(U_n)\neq CB(U_{n-1})$
$CB(U_n)=CP(U_n)$ and $CB(U_n)=CP(U_{n-1})$	CONTINUE	SMOOTH-SHIFT
$CB(U_n)=CP(U_n)$ and $CB(U_n)\neq CP(U_{n-1})$	EXPENSIVE CONTINUE	EXPENSIVE SMOOTH-SHIFT
$CB(U_n)\neq CP(U_n)$	RETAIN	ROUGH-SHIFT

Table 2.3: Transitions in Functional Centering

Rule 2 in FC

Cheap transition pairs are preferred over expensive ones.

(Strube and Hahn 1999, p.334)

2.2.5 More Centering

CT has motived many cross-linguistic studies in a variety of languages such as a) *Japanese*: Kameyama (1985, 1988, 1998); Iida (1998); Walker et al. (1990, 1994); Matsui (1999) b) *Korean*: Kim et al. (1999) c) *German*: Rambow (1993); Strube and Hahn (1999) d) *Yiddish*: Prince (1994) e) *Hebrew*: Grosz and Ziv (1998) f) *Turkish*: Turan (1995, 1998); Hoffman (1998) g) *Italian*: Di Eugenio (1990, 1996, 1998) h) *Greek*: Dimitriadis (1996); Miltsakaki (2002) i) *Spanish*: Taboada (2002) j) *Finnish*: Kaiser (2000) k) *Hindi*: Prasad (2000); Prasad and Strube (2000), etc.

Moreover, a substantial amount of CT-based work tests the psychological and cognitive plausibility of the model (mainly Hudson-D'Zmura et al. 1986; Hudson-D'Zmura and Tanenhaus 1998; Gordon et al. 1993; Stevenson et al. 1994, 2000; Brennan 1995, 1998), or studies its usefulness for discourse segmentation (Passoneau 1993, 1998), its appropriateness for evaluating student essays (Miltsakaki and Kukich 2000a,b) and its integration with a) Grosz and Sidner's theory of global focus: Hitzeman and Poesio (1998); Walker (1996, 1998); Grosz and Gordon (1999) b) Gundel *et al.*'s givenness hierarchy: Gundel (1998); Walker and Prince (1995); Kaiser (2000) c) theories of information structure: Hoffman (1998); Strube and Hahn (1996, 1999) d) relevance theory: Matsui (1999) e) theories of dynamic, lexical and discourse semantics: Roberts (1998); Cote (1998); Hudson-D'Zmura (1998); Stevenson et al. (1994, 2000) and f) various syntactic phenomena: Grosz and Ziv (1998); Hurewitz (1998); Birner (1998).

As far as computational implementations are concerned, most aspects of the above cited research have been used to specify algorithms for anaphora resolution. In addition to the BFP algorithm (Brennan et al. 1987), a number of algorithms for pronoun resolution, often based on different formulations of CT, have appeared in the recent literature (Strube 1998; Strube and Hahn 1999; Kim et al. 1999;

Tetreault 1999, 2001; Miltsakaki 2002). All these algorithms have been tested on English data, although most of them are applicable to a least one more language, thus providing some evidence for the universality of the theory.

The next two sections review those aspects of CT that our work mostly relates to: the corpus-based evaluation of the theory and its use for the purposes of NLG.

2.3 Corpus-based evaluations of Centering

Starting from the evaluation of the BFP algorithm in Walker (1989), most papers on CT involve some sort of informal corpus-based study that supports, or at least motivates, the reported work. Hence, we will often refer to relevant results from the most detailed evaluations of CT. However, Poesio et al. (2002) were the first to point out that most of the existing corpus-based investigations paid limited attention to the *underspecification of CT*, an issue with interesting methodological and theoretical implications. In this section, we discuss this problem in some detail on the basis of their observations. Then, we present an overview of the methodology and the main results of Poesio et al. (2002) which we consider to be the most complete of all corpus-based evaluations of CT up to now.[31]

2.3.1 Underspecification of Centering

Poesio et al. (2002) start with the observation that CT is best characterised as a parametric theory in that theoretical concepts such as *utterance*, *previous utterance*, *realisation* and *ranking* were intentionally left unspecified in its formulation. In order to present the basic aspects of this underspecification more clearly, let us first repeat the constraints and rules of the theory:

For each utterance U_n in a discourse segment D consisting of utterances U_1, ... , U_m:

Constraints

C1. There is precisely one CB(U_n).

C2. Every element of CF(U_n) must be realised in U_n.

C3. The CB(U_n) is the highest-ranked element of CF(U_{n-1}) realised in U_n.

Rules

R1. If any element of CF(U_{n-1}) is realised by a pronoun in U_n, then the CB(U_n) must be realised by a pronoun also.

[31] Note that a shorter, but more updated, version of Poesio et al. (2002) appears as Poesio et al. (2003). Our discussion is mainly based on Poesio et al. (2002), although we tried to accommodate for the most relevant points from Poesio et al. (2003) as well.

R2. Transition states are ordered: CONTINUE is preferred to RETAIN, which is preferred to SMOOTH-SHIFT, which is preferred to ROUGH-SHIFT.

Starting from the notion of *utterance*, most researchers follow Kameyama (1998) who defined utterance as the tensed clause with the exception of relative clauses and verbal complements which are called "embedded utterance units" and result in updates of the local focus that are then erased, much as in the way the information provided by subordinate discourse segments is erased when they are popped. Suri and McCoy (1994) suggest that some types of tensed adjuncts (in particular, clauses headed by the words "after" and "before") should be treated as embedded utterance units as well. However, Miltsakaki (2002) brings forth arguments from English, Greek and Japanese that the appropriate update unit for topic tracking is the sentence in its traditional sense (i.e. the unit containing the main clause and all the subordinate clauses associated with it) and not the tensed (or finite) clause.

Moving to the definition of Constraint 1, we are quoting what Poesio et al. (2002) call the "strong" formulation of Constraint 1 (Walker et al. 1998a, p.3). A very important point about Constraint 1, first made clear by Poesio et al. (2003), is that its strong formulation makes two claims with respect to the entity coherence of the discourse: first, that all pairs of utterances (except for the first one) satisfy CONTINUITY (see section 2.2.4.1); and that there is not more than one CB in the second member of the pair, what Poesio et al. (2003) call CB UNIQUENESS.[32] A more relaxed formulation of Constraint 1 appears in Walker et al. (1998a, p.3, footnote 2) who mention that C1 is often treated as stating that "there is **not more than one** Cb" (their emphasis). As Poesio et al. (2003) clarify, this formulation does away with CONTINUITY, but retains CB UNIQUENESS. That is, an utterance may or may not have a CB, but when it does satisfy CONTINUITY then the CB is unique.[33]

With respect to the formulation of Constraint 2, the notion *realise* can be interpreted in a strict sense, that is, by taking a center *c* to be realised by a noun phrase *NP* in U_n only if *NP* denotes *c*. In that case *NP directly realises c*. In addition, a center *c* can be counted as *indirectly realised* if it is referred to indirectly by means of a bridging reference (Clark 1977) or a similar kind of functional dependence, e.g. an inferable relation (Prince 1981, 1992).

According to Constraint 3, the *ranking* of CF(U_{n-1}) determines which of the elements that are realised in U_n will be the CB(U_n). Thus, the CF ranking is the main determinant of the transition state that holds between two utterances. Hence, it is not surprising that the definition of the ranking criteria appropriate for different languages (known as the *CF template*, Cote 1998), has been a matter of controversy. In fact, Walker et al. (1994) hypothesise that the CF template is the only language dependent factor within CT. Kameyama (1985) was the first to argue that grammatical role, rather than thematic role which Sidner (1979) used, affected CF ranking. Evidence for many additional

[32] In fact, Poesio et al. (2003) use the results of their evaluation to argue that these two claims be separated.

[33] Note that CB UNIQUENESS contrasts with Sidner's hypothesis that utterances may have two foci and theories which view "topichood" as a matter of degree such as Gundel (1998).

criteria for the CF template have been brought forward in the literature such as a) surface order of realisation: Rambow (1993); Gordon et al. (1993) b) information status: Strube and Hahn (1999) c) semantic role: Stevenson et al. (1994, 2000); Turan (1998); Hoffman (1998); Cote (1998). However, grammatical role is used in most formalisations of CT since it appears to be the most clearly defined concept, by contrast to rather imprecise notions such as thematic role. Grosz et al. (1995) have not excluded other factors, but strongly suggested a CF template for English based on grammatical role which was the definition followed by Brennan et al. (1987) as well.

In a separate line of research, Gordon et al. (1993) and Passoneau (1993) suggested replacing Constraint 3 with operational definitions of the CB(U_n) based on pronominalisation preferences. Gordon et al. (1993) identify the CB(U_n) with the entity that is subject to a *repeated name penalty* which is a slower reading time whenever a full NP is used instead of a pronoun to refer to the CB(U_n). The operational definition of the CB(U_n) in Passoneau (1993) is based on the observed uses of "it" and "that" in a corpus of dialogues.

The cited version of Rule 1 makes the pronominalisation of the CB(U_n) conditional on the existence of another pronoun in U_n. An earlier definition of the rule in Grosz et al. (1983) stated that the CB(U_n) **should** be pronominalised if it is the same as the CB(U_{n-1}). In addition, Gordon et al. (1993) supplement their operational definition of the CB(U_n) with a requirement that the CB(U_n) **should always** be pronominalised.

Last but not least, a further controversy within CT is whether Rule 2 applies to **pairs** or **sequences** of adjacent utterances. Grosz et al. (1995) claimed that Rule 2 applies to the level of **sequences of transitions** stating that sequences of CONTINUEs are preferred to sequences of RETAINs and sequences of RETAINs are preferred to sequences of SHIFTs.

Brennan et al. (1987) apply Rule 2 on an utterance-by-utterance basis for the BFP algorithm. Brennan (1998) suggests that this approach is plausible because psychological research has shown that both human sentence production and interpretation take place incrementally on a phrase by phrase level. In addition, Hudson-D'Zmura and Tanenhaus (1998) show that an immediate provisional interpretation of potentially ambiguous pronouns is made in a way that supports the predictions made by the CT model of Brennan et al. (1987).

Strube and Hahn (1999) as well as Di Eugenio (1998) and Turan (1998) examine how the previous CT transition affects the current one. Strube and Hahn (1999) follow the middle way between the definition of Rule 2 by Brennan et al. (1987) and the one in Grosz et al. (1995) stating that some transition types which receive bad marks in isolation might be more felicitous when occurring in the appropriate context and vice versa. For example, a CONTINUE:CONTINUE sequence is thought to require the lowest processing costs. But a CONTINUE transition that follows a RETAIN implies higher processing costs than a SMOOTH-SHIFT following a RETAIN. This is based on the claim that a RETAIN

should be used where possible before a SHIFT transition to a new CB (Grosz et al. 1995, p.215).

Moreover, Strube and Hahn (1999) notice that according to the table of transitions in Brennan et al. (1987) it is possible for a SMOOTH-SHIFT following a RETAIN to move the CB(U_n) to an entity other than the CP(U_{n-1}) of the RETAIN. In order to express the RETAIN:SMOOTH-SHIFT preference more precisely, Strube and Hahn (1999, p.333) introduce a table of 36 **transition pairs**, labelled as "cheap", "expensive", or "-" (not occurring). As we have shown in section 2.2.4.3, the main idea advocated by Rule 2 in FC is that cheap transition pairs be preferred over their expensive counterparts. However, Kibble (2001) notices that the table of transition pairs in Strube and Hahn (1999) is unnecessarily complicated, since it can be recast simply on the basis of the principle of CHEAPNESS using triples of utterances instead of pairs of transitions.

2.3.2 Using a reliably annotated corpus for evaluation

Poesio et al. (2002) observe that because of the underspecification of CT, the claims made in the model have only been tested fixing upon a particular way of instantiating the parameters, constraints and rules of CT, e.g. by using "finite clause" as the definition of utterance and Constraint 3 (henceforth C3) as the definition of the CB(U_n). This leaves a number of equally plausible theoretical combinations empirically unexplored.

In addition, most of previous corpus-based investigations of CT (such as Walker 1989; Di Eugenio 1996, 1998; Passoneau 1993, 1998; Hurewitz 1998; Kameyama 1998; Byron and Stent 1998; Strube and Hahn 1999; Tetreault 2001) were carried out by a single annotator marking up her corpus according to her subjective judgement. Hence, the idea of using only the information that can be annotated reliably appears to have skipped methodological attention.

Poesio et al. (2002) take Constraint 1 (henceforth C1), Rule 1 (henceforth R1) and Rule 2 (henceforth R2) to be the main claims of CT.[34] They test CT in a more general way than in the previous studies by trying to identify which ways of instantiating the parameters result in the fewest violations of its main claims. In order to carry out this study, they collected GNOME, a corpus of texts from several genres, and annotated it with information that is relevant to the different instantiations of CT.[35]

As we discuss in section 2.5, none of the existing studies on CT specifies a rigid methodology for estimating the entity coherence of the whole text or the collection of texts in a corpus. Poesio et al. (2002) overcome this problem by defining a function that sums up the violations of each claim in the corpus, which serve as a performance measure for each way of specifying the parameters of CT.

[34] Indeed, identifying which are the main claims of CT (especially with respect to C1 which differs from the other two Constraints that are definitions in essence), is pointed out by Poesio et al. (2002) as one of their main contributions.

[35] Section 6.2 of chapter 6 provides more details on the GNOME corpus and its annotation features. Note that using only those features that can be annotated reliably has its own price: It has not been possible for Poesio et al. (2002) to test the claims of CT which are based on features such as the thematic role of an NP for which the annotators were not able to reach acceptable agreement.

Poesio et al. (2002) argue that CT's claims should be viewed as *preferences*, best tested by standard statistical tests, and propose using the signtest for C1 and R1. Their aim is to study the effects of CT's parameterisation much more systematically than before, using appropriate tests of significance for the first time.[36]

Different configurations of CT are first compared according to the extent that they reduce the sum of violations of C1 in the corpus. One of the main findings of Poesio et al. (2002) is that defining the CB(U_n) in ways other than C3 results in a very large number of NOCB (in their terms, NULL or ZERO) transitions. Even when the CB(U_n) is defined according to C3, identifying utterances as finite clauses and only allowing for direct realisation of centers in the computation of CF(U_n) results in more pairs of utterances violating C1 than satisfying it. The number of NOCB transitions is reduced significantly when indirect realisation is specified and sentences are used for the definition of utterance. However, trying to reduce the number of NOCBs results in an increased number of violations of R1 and vice versa. Because of this tradeoff between C1 and R1, it is very difficult to say which is the "best way" of specifying CT.[37]

In order to evaluate the different instantiations of CT with respect to the standard version of R2 due to Brennan et al. (1987), the scoring function of Poesio et al. (2002) sums up the transitions in the corpus for a given specification of the CT parameters. A version of CT is taken to satisfy the canonical ordering of R2 if CONTINUE is found to be more frequent than RETAIN which in turn is found to be more frequent than the various kinds of SHIFT. In other words, if the frequencies of transitions from the corpus correspond to the canonical ordering of R2, then the version of CT that achieves this is taken to account for the entity coherence of the texts. Conversely, finding versions of CT the frequencies of transitions of which satisfy the canonical ordering of R2 is interpreted as providing evidence in favour of R2 as a robust estimator of the entity coherence of the texts in the corpus.

Only a few of the versions of CT invoked by Poesio et al. (2002) return frequencies of transitions that correspond precisely to the canonical ordering of R2. This replicates results from most previous corpus-based studies of CT that often introduce a more lenient criterion for the evaluation of R2 simply stating that CONTINUE should be the most frequent transition.[38] Somehow surprisingly, even under this lenient criterion standard R2 is not verified by most versions of CT tested by Poesio et al.

[36] We believe that this aim is stated much more clearly in the latest version of the paper (Poesio et al. 2003), where the Page Rank test is used to evaluate R2 as well.

[37] The tradeoff between C1 and R1 was one of the main findings of the preliminary evaluation of CT using the GNOME corpus reported in Poesio et al. (2000). Poesio et al. (2003) report an ever bigger tradeoff between C1 and R2 (the latter now being evaluated by the Page Rank test).

[38] However, the transitions used in these studies are not exactly the same as the ones used by Poesio et al. (2002). This has mainly to do with the different treatment of NOCBs, ESTABLISHMENTs and ROUGH-SHIFTs across researchers (see section 2.2.4.1 for a brief discussion).

(2002) including those with the fewest violations of C1.[39] Note that the version of R2 in Grosz et al. (1995) that employs sequences of transitions was not verified in the corpus-based investigation of Poesio et al. (2002). The same is true for the definition of R2 according to FC: all configurations of CT returned more expensive than cheap transition pairs (Poesio et al. 2002, p.66).

In their conclusions, Poesio et al. (2002) emphasise that CT should be supplemented by other coherence inducing factors, a point not discussed extensively in the CT literature, and suggest the ELABORATIONless framework of Knott et al. (2001) as a plausible model of the interaction between entity and rhetorical coherence.[40]

They also point out the difficulty of coming up with the "best way" of specifying CT. They argue, however, that despite the tradeoff between C1 and R1 and the high frequency of transitions other than CONTINUE, treating the main claims of CT as preferences rather than hard constraints makes it possible to find quite a few ways of setting the parameters of CT so that C1 and R1, and perhaps R2, are statistically verified.

2.3.3 Remaining issues in corpus-based evaluation

Although Poesio et al. (2002) present the most methodologically sound corpus-based evaluation of CT, we shall argue in section 2.5.3 that their method of evaluation does not investigate the choices available to an author for structuring a certain set of utterances (Kibble 2001). This is particularly important from an NLG viewpoint, but was not taken into account by Poesio et al. (2002) or by any previous corpus-based study of CT.

Chapter 5 presents a novel corpus-based, search-oriented evaluation methodology which is more suitable for the purposes of text structuring than the one presented by Poesio et al. (2002). In chapter 6, we apply this methodology to a subset of texts from the museum section of the GNOME corpus using the same tools as Poesio et al. (2002).

2.4 Centering and natural language generation

In this section, we review the way that CT motivated recent work on NLG. As a starting point for our discussion we compare the focus rules of TEXT with the preferences of R2. Then, we argue that CT has mainly covered aspects of pronominalisation within NLG, neglecting the original idea of McKeown (1985) that constraints on entity coherence can be used for more than one task in the

[39] This accords with the findings of Passoneau (1998) who reports that dispreferred transitions such as NULL appear more frequently than CONTINUE in her formulation of CT. However, other corpus-based investigations of CT do find that the CONTINUE transition is more frequent than its competitors (e.g. Di Eugenio 1998; Strube and Hahn 1999; Hurewitz 1998). These results are often taken to confirm R2 as be a reliable estimator of the entity coherence of the whole text. Note that R2 appears much more robust when tested by the Page Rank test in Poesio et al. (2003).

[40] This suggestion is discussed in section 6.4.1 of chapter 6 in more detail.

generation process. What is more, most of the recent systems that "pronominalise the CB" do not define it according to C3. This raises the question whether it is indeed possible to make use of C3 for the purposes of NLG.

To address these two issues we review the approach of Kibble and Power (2000) who go directly back to the claims of McKeown (1985). Kibble and Power (2000) integrate pronominalisation and text planning on the basis of CT, using the definition of the CB according to C3. Although we are not directly concerned with the task of pronominalisation, most of our work expands on the approach of Kibble and Power (2000) to text structuring, sharing similar assumptions. As pointed out in section 2.1.6, this section concludes with a discussion of the differences between CT and the framework of text structure in Knott et al. (2001).

2.4.1 Centering and focusing in TEXT

In section 2.1.1, we mentioned that McKeown used Sidner's Theory of Immediate Focus (henceforth, STIF) to control both text structuring and referring expression generation. Although CT and STIF have a lot in common, a direct comparison between them is beyond the scope of this chapter.[41] In this section, we restrict the discussion to the way that McKeown's preferences for immediate focus compare with standard R2.

To begin with, the preferences for changing and maintaining the focus in McKeown (1985) are quite different from the ones of R2. More specifically, changing the current focus to a member of the potential focus list of the previous proposition is preferred to maintaining the focus in order to avoid reintroducing the potential focus at a later point (McKeown 1985, p.62-64). This strategy has the effect of producing "topic clusters" around items just introduced in the discourse. By contrast, R2 favours talking about an established topic instead of shifting the focus to a new one.

On the other hand, McKeown's default focus is similar to the CP(U_n) of standard CT in being directly associated with subjecthood. By contrast, the definition of the default focus in STIF is based on thematic role.[42] However, McKeown's default focus is different from the CP(U_n) and STIF's default focus in that it is used to establish the focus of the **current** utterance and not to predict the focus of the **next** one (McKeown 1985, p.70). In other words, focusing in TEXT is not computed by ranking the entities mentioned in the previous proposition in order of prominence as it happens with CT's C3. By contrast, it is the preferences among the focus rules which mainly specify which argument is recognised as the focused entity. After the focus has been established, the tactical component favours pronominalisation over using a definite description for subsequent reference to an already focused entity (McKeown 1985, p.77).

[41] In general, CT can be seen as a simplification of STIF. The reader is referred to the review of Lecoeuche et al. (1998) and the extended discussion in Poesio and Stevenson (2003) for more details.

[42] According to Walker et al. (1998a, p.3, footnote 1), CP(U_n) roughly corresponds to Sidner's *expected focus*.

Although the focusing rules in TEXT are quite distinct from the rules and constraints of CT, the main intuition behind McKeown's implementation is to use preferences for entity coherence **directly** during text structuring. As we show in the next section, this has been neglected by more recent NLG applications. In these systems, the CB(U_n) is predefined in the database and used for pronominalisation during sentence planning. What these systems have in common with TEXT is that neither is using C3 (or anything similar) in order to define the CB(U_n).

2.4.2 Centering and the generation of referring expressions

A number of NLG practitioners followed the CT-oriented research in anaphora resolution by implementing various pronominalisation algorithms within Reiter's pipeline architecture.[43] However, most of this work does not define the CB(U_n) according to the standard formulation of CT, maybe due to the limitations of reversing R1 for the purposes of generation (recently pointed out by McCoy and Strube 1999 and Henschel et al. 2000)

A typical example of this approach is the algorithm in Passoneau (1993, 1998), where the CB(U_n) is not defined in terms of C3. Similarly, Dale (1992) uses a domain-dependent criterion for identifying the "center" which is semantically defined as the result of the previously described operation (Dale 1992, p.170).

ILEX uses a simplified version of CT for pronominalisation as well (Hitzeman et al. 1997; O'Donnell et al. 1998). We mentioned in section 2.1.4 that each fact node in the content potential is indexed according to the entity that it is primarily about, termed as Arg1. O'Donnell et al. (1998, p.49) allow pronominalisation only when the referent of an NP is the same as the Arg1 of the previous proposition (or of the top nucleus of the previous RS-tree). In their view, the Arg1 is equivalent to the CB(U_n). In our opinion, the Arg1 is better seen as specifying a preferrence for the CP(U_n), i.e. the most prominent argument of the proposition/RS-tree. If the referent of an NP is the same as the CP(U_{n-1}), it is also the CB(U_n) in accordance with the standard formulation of CT.

Hence, we interpret the algorithm for pronominalisation in ILEX as *pronominalising the CB(U_n) if it is the same as the CP(U_{n-1})*. This algorithm reverses the main preference for anaphora resolution in Strube and Hahn (1999) which is directly based on the principle of CHEAPNESS, viz. the requirement that CB(U_n)=CP(U_{n-1}).

The reluctance of the recent approaches to define the CB(U_n) explicitly using C3 raises the question whether a straightforward implementation of CT for the purposes of NLG is desirable or even

[43] As we mentioned in section 1.1 of the introductory chapter, although there is a lot of variability in the structure of NLG systems, there is considerable consensus that the process of generation breaks down logically to at least six tasks, namely content determination, text structuring, aggregation, referring expression generation, lexicalisation and surface realisation (Reiter 1994; Reiter and Dale 1997, 2000). Pronominalisation is usually considered to be a subtask of referring expression generation. The modules implementing (some of) these tasks are often arranged in a pipeline fashion where the output of one module serves as the input for the next.

possible. This issue is discussed in the next section.

2.4.3 Integrating Centering with text generation

In his discussion of the appropriateness of CT for the purposes of NLG, Kibble (1999) argues that COHERENCE and SALIENCE belong to different tasks within the pipeline architecture. According to Kibble (1999), COHERENCE can be responsible for ordering propositions to maintain referential continuity, a task related to text structuring, whereas SALIENCE can be used to choose a construction that makes the CB(U_n) prominent within a clause or sentence, a matter relevant to sentence planning. In addition to this, Kibble (1999) points out that there might not be a single point in the generation process for making a choice between one or the other type of transition as defined by standard CT.

The conclusion of Kibble (1999) is that in NLG "the topic" should not be defined according to C3. Instead, it should appear prerecorded in the database, as in the systems reviewed in the previous section. Kibble (1999) suggests that, in order to implement CT for text structuring, the text planner can be used to independently designate the CB(U_n) if the topic of U_n is an argument of U_{n-1}. The text should be organised so that the same CB is maintained over a sequence of clauses. Then, the realisation of U_{n-1} can be planned in order to promote an entity to the highest-ranked subject position in U_{n-1} if this entity is either the CB(U_n) or, less preferably, the CB(U_{n-1}).

However, Kibble and Power (2000) claim that designating the topic independently as part of the semantic input as suggested by Kibble (1999) does not solve the problem of identifying the topic. Following Prince (1999), they propose that the topic should be identified with the CB(U_n) as defined by CT. Kibble and Power (2000) treat the task of identifying the CB(U_n) as a constraint satisfaction problem, assuming that certain options for syntactic realisation can be predicted on the basis of the argument structure of predicates. In their system, the potential CBs of a proposition U_n are given by the intersection between CF(U_n) and CF(U_{n-1}), which consists of all the arguments that the two propositions have in common. The potential CPs of U_n are all those referents in the current proposition that can be realised in the subject position, a decision based on case roles within the proposition.

Note that Kibble and Power (2000) return to the basic intuition of McKeown (1985) that a theory of entity coherence could fit the generation process in more than one task, namely text structuring and sentence planning. Their operationalisation of CT means that transitions can be calculated as part of text structuring, contrary to the arguments in Kibble (1999).

Our work on text structuring is a systematic attempt to determine whether the constraints and rules of CT can be turned round to guide the text structuring process, extending the approach of Kibble and Power (2000). In this sense, we assume that the CB(U_n) is computed on the fly during text structuring according to the standard apparatus of CT. After the extended review of CT, we are now ready to discuss how well it fits to the framework of text structure in Knott et al. (2001).

2.4.4 Centering beyond ELABORATION

In this section, we compare the preferences of CT with the ELABORATIONless framework that underlies the text structuring component of ILEX. To do this we repeat example (2.1) of section 2.1.4 as example (2.7), annotated with the data structures of standard CT, namely CF lists, CBs and the transitions in Table 2.1:[44]

(2.7) C_1, E_1: `J-999`

a. This piece is a necklace.

`CF(J-999, necklace)`

b. It was designed by a jeweller called Jessie King.

`CF(J-999, King),`

`CB=J-999,` CONTINUE

c. It was designed in 1905.

`CF(J-999, 1905),`

`CB=J-999,` CONTINUE

d. It is made of silver.

`CF(J-999, silver),`

`CB=J-999,` CONTINUE

C_2, E_2: `King`

e. Jessie King was a famous designer.

`CF(King, designer),`

NOCB

fS. She was Scottish,

[44] Since all utterances in example (2.7) correspond to fact nodes in ILEX's content potential, the CF lists in example (2.7) are computed accordingly. That is, the entities in the CF list correspond to the arguments of the facts in the content potential. In this sense, the entity `necklace` is part of the CF(2.7a) even though it is evoked by a predicative NP. In addition, the correspondence between NPs and entities is not strictly one-to-one: for example, the entity `King` is evoked by a complex NP in (2.7b). Further to this, in the computation of the CF list we do not account for bridging relations between entities related in the domain ontology of ILEX. To compute the CF list of local RS-trees such as the one represented by (h) and its satellite (hS) we extend the standard CF template of Walker et al. (1998a), mentioned in section 2.2.3, to cover the nucleus-satellite distinction: According to this, the Arg1 of the top nucleus corresponds to the CP of the RS-tree (see section 6.3.3 of chapter 6 for more details). In addition, we take the cleft sentence (2.7g) to manifest the promotion of an Arg2 to the CP position in accordance with the assumptions of Kibble and Power (2000). Finally, instead of identifying the Arg1 of the propositions or RS-trees with the CB(U_n) we take it to correspond to the CP(U_n), whereas the CB(U_n) is computed on the fly according to C3.

f. but she worked in London.

`CF(King, London),`

`CB=King,` CONTINUE

g. It was in London that this piece was made.

`CF(London, J-999),`

`CB=London,` SMOOTH-SHIFT

C_3, E_3: `Arts-and-Crafts-style`

hS. Like the previous piece,

h. this piece is in the Arts-and-Crafts style.

`CF(J-999, Arts-and-Crafts-style, J-888),`

`CB=J-999,` SMOOTH-SHIFT

iS. Although the previous piece had a simple shape,

i. Arts-and-Crafts style jewels tend to be elaborate;

iS. for instance, this piece has detailed florals.

`CF(Arts-and-Crafts-style, J-888, J-999),`

`CB=J-999,` RETAIN

The first question we have to address in our attempt to recast ILEX's framework of text structure in terms of standard CT is whether CT applies to the whole of the discourse or should simply be restricted within the entity chain. Knott et al. (2001) maintain that entity chains are equivalent to the focus spaces of Grosz and Sidner (1986). This might suggest that CT is only allowed to apply within the entity chains with the first unit of each chain taken to correspond to a segment-initial utterance.

The main problem with this assumption is that in Grosz and Sidner's theory of discourse structure, the global attentional state is dependent on the **intentions** of the conversational participants. Due to the opportunistic nature of intentionality within the descriptive genre, the equation of entity chains with Grosz and Sidner's focus spaces is difficult to maintain. In this sense, (2.7) is better seen as a single segment where CT controls the local attentional structure between and within entity chains.

C_1 is a sequence of text spans adhering to the rigid definition of a legal entity chain in Knott et al. (2001). All the fact nodes in C_1 have `J-999` as their Arg1. This defines utterances (2.7b-d) as a sequence of CONTINUE transitions in our CT analysis. Note that Knott et al. (2001) do not explain how the set of propositions are ordered with respect to each other within C_1. This turns out to be quite important when we consider the move to the second entity chain C_2, which according to Knott et al.

(2001) is achieved via a resumption relation to the discourse-old entity `King`. At this point we identify **the main incompatibility** between the framework of text structure in ILEX and the preferences of CT: achieving a resumption from (2.7e) to (2.7b) results in a NOCB transition in (2.7e). Note that the NOCB in (2.7e) can be easily undone if (2.7b) is placed as the last utterance of C_1, so that CONTINUITY is preserved within and across the first two entity chains.

Interestingly, although Knott et al. (2001) claim that all propositions within a legal entity chain should have the same Arg1 (unless they are part of a local RS-tree), this is clearly not the case for (2.7g), since the Arg1 of (2.7g) is not `King`, the global focus of C_2. Instead, according to our CT analysis, (2.7g) shifts the CB(U_n) to `London` **within** the second entity chain. This shows that an entity chain in the framework of text structure underlying ILEX can employ a variety of transitions in addition to CONTINUE.

This becomes more evident when C_3 is considered. After being shifted from `King` to `London` in the second entity chain, the CB is placed back to `J-999` at the end of the discourse. Note that according to this analysis the entity `Arts-and-Crafts-style` never becomes the CB. Rather, the discourse finishes with a RETAIN which creates the expectation that the next utterance after the last local tree of (2.7) might center on the `Arts-and-Crafts-style`.

We conclude that the entity chains of ILEX can be analysed into various CT transitions. However, the main incompatibility between the CT analysis and the framework of ILEX is the resumption relation from (2.7e) to (2.7b). Knott et al. (2001) acknowledge the need for more empirical work to investigate the structural properties of a resumption. Hence, we provisionally take resumption to be a constraint which is specific to their domain of application.

Crucially, the existence of resumptions is not the only theoretical argument of Knott et al. (2001) in favour of an ELABORATIONless framework of RST relations. Therefore, the question remains whether, modulo resumptions, the framework of text structure in ILEX can be supplemented with an account of entity coherence based on notions from CT. However, there is a significant modification in the way that our work addresses this question.

Following the suggestions of Kibble and Power (2000) and Knott et al. (2001), this thesis is a systematic attempt to use notions from CT to specify a model of entity coherence for the purposes of text structuring in NLG. In an important deviation from them, however, entity coherence is isolated as the most relevant factor for text structuring, while rhetorical relations are considered only to the limited extent that the datasets available for the study allow us. Despite this modification, we believe that considering entity coherence as the **only factor** for descriptive text structuring is an interesting question on its own and comes closer to the various formulations of CT which do not appear to acknowledge other constraints as being relevant to characterising textual coherence.

In order to specify how CT can be used for the purposes of text structuring in more detail, some more discussion of the theory is in order. This is motivated by our observation that despite the partial compatibility between the two models of text structure, there seems to be an intuition that entity chains capture, but the transitions of CT fail to address. Standard CT applies on a pair of utterances, whereas entity chains are an attempt to estimate the entity coherence of longer spans of text within a segment. Clearly, in NLG there is a need for a more global estimate of entity coherence than the one provided by C1 and R2. Note that this seems to arise when CT is used for interpretation as well as shown by the underspecification of the window of application of R2 (pairs vs sequences of utterances, as already discussed in section 2.3.1). In the next section, we discuss the problems that need to be confronted for CT to be able to estimate the entity coherence of longer spans of text directly, without resorting to an intermediate representation such as the entity chain.

2.5 Applying Centering to longer spans of text

In this section, we discuss whether the preferences underlying R2 and C1 can be used to estimate the entity coherence of text spans longer than a pair of utterances. We begin with the inappropriateness of R2 as a predictor of the RETAIN:SHIFT pattern which has been claimed to be the preferred way for shifting the CB(U_n). This leads to the conclusion that an estimate of entity coherence that extends to more than two utterances is necessary for CT to account for the alleged RETAIN:SHIFT preference.

Then, we look at the more general question of whether CT can be used to estimate the coherence of a structure that spans across several utterances. We suggest that, although summing up the transitions in a text is the first step to this direction, comparing these sums with the preferences underlying R2 and C1 directly might not be the most appropriate way to estimate the entity coherence of the whole text given a certain content.

To support this argument, we present examples that include transitions dispreferred by R2 and C1. We show that in order to account for the existence of dispreferred transitions in a text of attested coherence, one has to consider the choices available to an author when structuring a certain set of utterances. One way to do this is by employing a search-oriented strategy which views the preferences underlying R2 and C1 as a relative rather than an absolute measure of entity coherence.

What is more important for our purposes, the discussion in this section shows that defining a scoring function of entity coherence for the complete text structure is a more appropriate alternative than trying to use R2 incrementally for text structuring. We conclude the chapter by formalising the first research question that our work addresses.

2.5.1 RETAIN as a prediction for a SHIFT

Consider the following example, where (2.8c-d) and (2.8c'-d') are different context-independent realisations of the propositions that follow utterance (2.8b):[45]

(2.8) a. This exhibit is an amphora.

`CF(exhibit1, amphora)`

b. This exhibit comes from the archaic period.

`CF(exhibit1, archaic-period)`

`CB=exhibit1,` CONTINUE

c. This exhibit was decorated by an artist known as the "painter of Kleofrades".

`CF(exhibit1, painter-of-Kleofrades)`

`CB=exhibit1,` CONTINUE

d. The "painter of Kleofrades" used to decorate big vases.

`CF(painter-of-Kleofrades, entity-4049)`

`CB=painter-of-Kleofrades,` SMOOTH-SHIFT

c'. An artist known as the "painter of Kleofrades" decorated this exhibit.

`CF(painter-of-Kleofrades, exhibit1)`

`CB=exhibit1,` RETAIN

d'. The "painter of Kleofrades" used to decorate big vases.

`CF(painter-of-Kleofrades, entity-4049)`

`CB=painter-of-Kleofrades,` SMOOTH-SHIFT

The CB of both (2.8c) and (2.8c') is the entity `exhibit1`. Note that in both cases this CB is the same as the CB of (2.8b). The transition of (2.8c) is a CONTINUE because `exhibit1` is both the CB(2.8c) and the CP(2.8c). By contrast, in (2.8c') it is the `painter-of-Kleofrades` that is promoted to the CP position. Since the CB(2.8c') is the same as the CB(2.8b) but distinct from the CP(2.8c'), the transition of (2.8c') is a RETAIN.

According to Grosz et al. (1995, p.215), a RETAIN ideally should be used to introduce a SHIFT in the following utterance. Brennan (1995, 1998), Turan (1995, 1998), and Di Eugenio (1996, 1998)

[45] Most of the utterances in the examples that follow realise propositions derived from the database of the MPIRO system, which is the multilingual extension of ILEX (Isard et al. 2003). The procedure used to realise database propositions out of context is explained in section 9.2.3 of chapter 9. The CF lists of the utterances are computed in accordance with the assumptions stated in footnote 44 in section 2.4.4. An example of the computation of the CF lists from propositions in MPIRO is given in section 7.3 of chapter 7. A major difference between ILEX and MPIRO is that, at the time of writing these lines, MPIRO's database does not represent rhetorical relations. Like ILEX, the propositions in MPIRO correspond to binary predicates, the Arg2 of which can often be an entity such as `entity-4049` in (2.8d). This entity is not represented in the domain ontology of the system and is realised by canned text.

report studies on pronominalisation in different languages which are often taken to support the hypothesis that a RETAIN in U_{n-1} is a signal of an intention to (SMOOTH-)SHIFT the CB(U_n) to another entity by realising the CB(U_{n-1}) in a lower-ranked position in CF(U_{n-1}).[46] In addition, Brennan et al. (1987) suggest that a computational system for generation should try to use a RETAIN as a signal for an impending SMOOTH-SHIFT, so that after a RETAIN, a SMOOTH-SHIFT will be preferred rather than a CONTINUE.

Crucially, all these suggestions aim at promoting a (SMOOTH-)SHIFT instead of a CONTINUE **after** a RETAIN, rather than choosing between a RETAIN over a CONTINUE **before** a (SMOOTH-)SHIFT. As indicated by Kibble and Power (2000), although an incremental algorithm for text structuring based on R2 is conceivable, if it is applied on an utterance-by-utterance basis to structure the utterances in (2.8), the CONTINUE in (2.8c) will be preferred over the RETAIN in (2.8c') as the most coherent transition following (2.8b). However, it is the RETAIN:SMOOTH-SHIFT sequence that is taken to represent the globally best solution for (2.8) according to standard CT.

We identify this problem as the RETAIN:SMOOTH-SHIFT *inadequacy of R2* (RSI). In order to account for the RETAIN:SMOOTH-SHIFT pattern, R2 must be supplemented with a global estimate of the text structure that extends to at least a triple of utterances. Such an estimate is provided by the principle of CHEAPNESS which advocates the RETAIN:SMOOTH-SHIFT sequence as one of the cheap transition pairs, by contrast to the CONTINUE:SMOOTH-SHIFT sequence in (2.8c-d) which belongs to the expensive transition pairs.[47]

In section 3.2.1 of chapter 3, we use notions from Optimality Theory (Prince and Smolensky 1997) to define a simpler version of this estimate. By analysing (2.8) in terms of the competition between the principles of CHEAPNESS and SALIENCE, we provide a simple solution to RSI, without having to resort to the more complicated definition of 36 transition pairs in Strube and Hahn (1999).

[46] None of the corpus-based studies of CT, including Poesio et al. (2002), reports a substantial amount of RETAIN:(SMOOTH-)SHIFT pairs. In other words, most of the times (SMOOTH-)SHIFTing the CB(U_n) is not preceded by a RETAIN in U_{n-1} in the way assumed by Grosz et al. (1995). The studies on pronominalisation in Brennan (1995, 1998), Di Eugenio (1996, 1998) and Turan (1995, 1998) provide **indirect** evidence in favour of the RETAIN:(SMOOTH-)SHIFT pattern. Brennan (1995, 1998) reports that an entity first realised in object position cannot be pronominalised in the next utterance. In addition, Di Eugenio (1996, 1998) and Turan (1995, 1998) report that a CONTINUE that follows a RETAIN is just as likely to realise the subject with a strong pronoun as with a null pronoun whereas a CONTINUE that follows another type of transition is much more likely to use a null pronoun in the subject position. This was taken to provide evidence in favour of the hypothesis that a RETAIN signals an upcoming (SMOOTH-)SHIFT rather than a CONTINUE. If the (SMOOTH-)SHIFT does not occur, then the speaker must use an NP other than a null pronoun to prevent the hearer from misinterpreting the utterance. However, as Karamanis (2001) explains, Di Eugenio's remarks on pronominalisation should not be restricted to the RETAIN:CONTINUE pattern but can be extended to all pairs of utterances whose second member violates the principle of CHEAPNESS.

[47] Note, however, that this only holds when grammatical function is used to compute the CF list. In this case, utterance (2.8d) can be indeed classified as an EXPENSIVE SMOOTH-SHIFT. If information structure is used for the ranking of the CF list, as Strube and Hahn (1999) suggest, both (2.8c) and (2.8c') are CONTINUEs and the utterance that follows them will violate CHEAPNESS.

2.5.2 Estimating the coherence of the whole text

RSI arises from conflicting predictions within CT caused by the inability of the canonical ordering in standard R2 to account for the suggestion that a RETAIN is the preferred way of introducing a SHIFT. As we mentioned in the previous section, corpus-based evidence in favour of the RETAIN:SHIFT preference is at best inconclusive. Our discussion of RSI does not aim at supporting or rejecting the view of Grosz et al. (1995), but to introduce a more general problem of CT, that is, the inability to use R2 directly in order to estimate the coherence of a text that consists of more than one pair of utterances.

As Beaver (2003) notices, the utterance-by-utterance classification of transitions in Brennan et al. (1987) and indeed the bulk of later literature in CT do not provide a clear way to estimate the coherence of a complete text.[48] In other words, what CT can do directly is to compare two different transitions from an utterance on the basis of R2. Because the preferences of R2 in Brennan et al. (1987) act at a sentence-by-sentence level, CT does not directly provide a *scoring function* for estimating the coherence of a structure that spans across several pairs of utterances.

Beaver (2003) reformulates the Centering algorithm for anaphora resolution in Brennan et al. (1987) into a set of violable constraints in the spirit of Optimality Theory (OT). His model, namely *Centering in Optimality Theory* (COT), is not only relevant to anaphora resolution, but is also thought to apply to the evaluation of complete texts:[49]

> It is possible to apply COT to compare the felicity of arbitrary large texts. [To do this] it is necessary to decide how to count violations of constraints in different sentences. To demonstrate the possibility of optimising entire texts, I propose that we count violations in a multi-sentence discourse in the most obvious way: we form one tableau using the standard COT constraint ranking, we enter violations of each constraint in the column corresponding to the violated constraint regardless of the sentence in which the violation occurred and then select the optimal candidate using the standard OT method.
>
> (Beaver 2003, section 5.4)

In general, the main focus of Beaver (2003) is placed upon the equivalence of COT with the model of Brennan et al. (1987), so most of the specified constraints and the discussed examples do not have to do with the estimation of the coherence of the whole text per se, but are inspired by ways of facilitating pronoun resolution in comparison to the algorithm in Brennan et al. (1987). As a result, the evaluation of complete text structures is discussed very briefly using only two examples from Grosz and Sidner (1998).

[48] Neither do the immediate focus rules of McKeown (1985). As we mentioned in section 2.4.1, the difference between the way that McKeown (1985) interprets STIF and standard R2 is that McKeown prefers to shift the focus in the second member of a given pair of utterances whereas R2 maintains it. However, both models define local preferences operating at the level of a pair of utterances. We cannot see a way of estimating the coherence of a complete text without translating the rules of McKeown (1985) into a scoring function of text structure in the sense discussed in this section. Defining such a function lies beyond the scope of the thesis which is restricted to investigating CT's potential for this purpose.

[49] Various examples of applying Beaver's methodology are presented in the next chapter.

However, the remark of Beaver (2003) about estimating the coherence of a **whole** text by counting sums of violations is quite useful for the following reason: As we mentioned in section 2.3.2, none of the existing corpus-based investigations of CT formally specifies a rigid methodology for estimating the coherence of a complete text structure. However, what most of them do practically is to count **sums of transitions** in a text. Thus, they informally define a scoring function of entity coherence which sums up the transitions in different utterances. Then, the frequencies of transitions are compared to the preferences that underlie the basic claims of CT. C1 and R2 are taken to be robust estimators of the entity coherence of the texts in the corpus, if a certain specification of CT returns frequencies of transitions that correspond to these preferences.

At this point, we need to emphasise that counting sums of COT violations or standard CT transitions is by no means specific to these versions of CT. For the sake of simplicity, in the examples that follow we employ the scoring function for standard C1 and R2 from Poesio et al. (2002). In chapter 3, we show how this approach can be extended to the definition of other scoring functions of entity coherence based on the different ways of specifying CT reviewed in section 2.2.4.

In the next section, we argue that summing up the transitions in a text, although necessary, might not be the most reliable way for estimating the coherence of a text structure given a certain content. To support this claim we present two examples which show that considering the alternative sequences of the utterances that a text consists of can be more enlightening than comparing the frequencies of transitions with the underlying preferences of R2 and C1.

2.5.3 Looking at alternative sequences of utterances

As we saw in the preceding sections, R2 and C1 are the main predictions of standard CT with respect to the entity coherence of a pair of utterances. However, CT is unclear on how to use them in order to estimate the entity coherence of longer text spans. What most of the existing corpus-based investigations of CT did in response to this was to define scoring functions of entity coherence based on the sums of transitions in a text. The frequencies of the transitions are often compared with the following absolute preferences: a) CONTINUE is the most frequent transition (abs-CONT) and b) minimise NOCB (abs-NOCB). When the frequencies of transitions are found to adhere to these preferences, CT is taken to be a reliable estimator of the entity coherence of a text.

Kibble (2001) was the first to point out that the methods followed in previous corpus-based evaluations of CT are incomplete:

> [...] corpus analysis itself is not sufficient to evaluate the claims of CT without taking into account the underlying semantic content of a text. That is, statistics about the relative frequency of occurrences of different transition types [...] do not take account of the choices available to an author.
>
> (Kibble 2001, p.582)

The suggestion of Kibble (2001) is very important since it relates to a quote from Grosz et al. (1995):

> Rule 2 provides a constraint on speakers, and on natural language generation systems. They should plan ahead to minimise the number of shifts. [...] To empirically test the claim made by Rule 2 requires examination of differences in inference load of alternative multi-utterance sequences that differentially realise the same content.
>
> (Grosz et al. 1995, p.215)

To our knowledge, the suggestion of Grosz et al. (1995) to examine **alternative sequences of utterances** that differentially realise the same content has not been followed in any of the existing corpus-based evaluations of CT. Note that in Grosz et al. (1995) the preferences that underlie C1 and R2 are viewed as a way of discriminating an attested sequence of utterances from its alternatives.[50]

The suggestions of Kibble (2001) and Grosz et al. (1995) are very important not only from an interpretation but also from a generation point of view. In NLG, and especially in the generation of structures for descriptive texts, most (if not all) of the semantics that have to be communicated are available to the system before text structuring. The text structuring component needs to output a coherent sequence of propositions from a large set of potential alternatives. Thus, the problem of **choice** between solutions, central to any decision in NLG, needs to be considered in the definition of a model of entity coherence based on CT for the purposes of text structuring.

In the remainder of this section we present two examples that elaborate on the quotations of Kibble (2001) and Grosz et al. (1995). We argue that comparing the frequency of transitions with abs-CONT and abs-NOCB, without considering their alternatives, is an incomplete way for estimating the coherence of a text given a certain content. We conclude that judging the coherence of a complete structure sufficiently requires searching through different sequences of utterances that realise the same content.

Clearly, this thesis presents only one way of investigating the problem of choice, the one that arises from a specific NLG viewpoint. For instance, our argument in the next section is based on the assumption that content determination is done strictly before text structuring, that is, we ignore the possibility of changing the propositional content in order to make the text structure more coherent. Additional complications that have to do with decisions such as segmentation, aggregation, etc. are also not taken into account. Even with these modifications, to our knowledge, this thesis represents the first, albeit limited, attempt to account for the problem of choice in the corpus-based evaluation of CT.

[50] As Grosz et al. (1995) talk about "differences on inference load", the best way to test their claim is by perceptual experiments and indeed this path was followed by many proponents of CT (see section 2.2.5 for relevant citations). However, as Poesio et al. (2002) very convincingly argue, these experiments are very difficult to take place on a large scale, while it is practically impossible to address CT's underspecification with them. We maintain that the same is true for the problem of estimating the coherence of a complete text structure given a propositional content. Hence, corpus-based research represents the most realistic alternative, especially for an extensive study. See section 4.6 of chapter 4 for an elaboration on this point.

2.5.4 Estimating the coherence of the whole text structure requires search

In order to demonstrate how important it is to account for alternative ways of ordering the same set of utterances, consider the following example, adapted from a human text describing a museum artefact in the MPIRO domain:[51]

(2.9) a. This exhibit is an amphora.

`CF(exhibit1, amphora)`

b. Amphoras have an ovoid body and two looped handles, reaching from the shoulders up.

`CF(amphora, entity-3908),`

`CB=amphora, CONTINUE`

c. Amphoras were produced in two major variations: type A and the type with a neck.

`CF(amphora, typeA, type-neck)`

`CB=amphora, CONTINUE`

d. This exhibit is a type A amphora.

`CF(exhibit1, typeA)`

`CB=typeA, ROUGH-SHIFT`

e. This exhibit comes from the archaic period.

`CF(exhibit1, archaic-period)`

`CB=exhibit1, SMOOTH-SHIFT`

f. This exhibit was painted using the red figure technique.

`CF(exhibit1, red-figure-technique)`

`CB=exhibit1, CONTINUE`

Assuming a scoring function that calculates the sums of standard CT transitions, the structure in (2.9) has three CONTINUEs, one ROUGH-SHIFT in position (2.9d), and one SMOOTH-SHIFT in position (2.9e). Note that the text in (2.9) satisfies abs-CONT because it has more CONTINUEs than (SMOOTH or ROUGH) SHIFTs. Abs-NOCB is also satisfied since none of the pairs of utterances in (2.9) violates C1, i.e. there are no NOCBs in (2.9).

Do abs-CONT and abs-NOCB account for the coherence of (2.9) after all? One is tempted to be affirmative since (2.9) has more preferred than dispreferred transitions, thus roughly satisfying the intuition behind R2, and no violations of CONTINUITY, adhering to C1 for every pair of utterances.

[51] Although utterances (2.9c) and (2.9d) appear in the human text, they do not correspond to propositions in the MPIRO database. For this reason we computed their CF lists using the ranking assumed by the standard version of CT.

Therefore, R2 and C1 as manifested by abs-CONT and abs-NOCB could be taken as robust estimators of the coherence of (2.9).

However, there are a couple of questions that remain unaddressed by abs-CONT and abs-NOCB. For example: Why is the structure at (2.9) not a series of CONTINUEs as the canonical ordering of R2 would optimally predict? Isn't it at odds with R2 that two of the transitions in (2.9) are SHIFTs? How is it possible for a coherent text like (2.9) to have a ROUGH-SHIFT, the most dispreferred transition according to R2? Is there a way to violate C1 given the utterances in (2.9)? All these questions are summed up to the following one: *What are the alternatives to* (2.9)? Or to put it more generally:

> Given a discourse D consisting of an ordered set of utterances UD $= \{U_1, \ldots, U_n\}$, is it possible for the alternative sequences of the members of UD to differ from D according to the way a scoring function based on some formulation of CT estimates their entity coherence?

In order to address this question, we search through the alternative sequences of the utterances in (2.9) and record their transitions.[52] Then, we can compare the sums of transitions in the alternative orderings with the sums from (2.9). First, we notice that 116/120 (96.67%) of the possible orderings have at least one NOCB transition. This shows that although it is possible to put the utterances in (2.9) in such an order as to violate CONTINUITY, (2.9) is one of the 4 sequences of utterances that does not exhibit this property. Arguably, abs-NOCB cannot account for this fact; only a relative account of the preference that underlies C1, such as the one provided here, manages to estimate the amount of differentiation of (2.9) from its alternatives with respect to the property NOCB.

Crucially for abs-CONT, each of these 4 sequences has three CONTINUEs, one SMOOTH-SHIFT and one ROUGH-SHIFT as their transitions. Hence, given the set of utterances that (2.9) consists of, both SHIFTs in (2.9) are unavoidable under any ordering that satisfies C1 for every pair of utterances. As a result, using abs-CONT to estimate the coherence of (2.9) is not so informative since the same sums of transitions emerge in all structures that minimise on NOCB.

In summary, using frequencies of transitions to evaluate the coherence of (2.9) in direct comparison with abs-CONT (or the canonical ordering of R2) does not account for the existence of two dispreferred transitions. Looking at the alternative sequences of utterances provides a clearer explanation on what are the possible choices for ordering the utterances in (2.9). This way of estimating the coherence of (2.9) reveals that the ROUGH-SHIFT and the SMOOTH-SHIFT in (2.9) are unavoidable for any sequence of utterances that avoids a NOCB transition.

[52] As Dimitromanolaki and Androutsopoulos (2003) report, there is a strong preference for the first utterance in each ordering always to be "*This exhibit is an amphora*". We view this convention as a piece of domain communication knowledge (Kittredge et al. 1991) that our investigation needs to account for. Therefore, what one needs to do is to permute the utterances in positions (2.9b-f), keeping the utterance in (2.9a) as the first utterance in each possible sequence of utterances. We did this for examples (2.9) and (2.10) and recorded the transitions for 5!=120 possible orderings.

Since (2.9) is distinguished from its alternatives by the lack of NOCB transitions, one might be tempted to believe that estimating the coherence of a structure in terms of abs-NOCB might be just enough. However, our main claim is that a structure of attested coherence needs to be compared with its alternatives as far as the preference for avoiding NOCBs is concerned as well. This is made clear by the following example which comes from a corpus of coherent sequences of propositions derived from the MPIRO database:[53]

(2.10) a. This exhibit is an amphora.

`CF(exhibit1, amphora)`

b. This exhibit was painted using the red figure technique.

`CF(exhibit1, red-figure-technique),`

`CB=exhibit1, CONTINUE`

c. In the red figure technique, the background was painted black and the figures that were predesigned had the natural color of the clay.

`CF(red-figure-technique, entity-2474),`

`CB=red-figure-technique, SMOOTH-SHIFT`

d. The red figure technique is the opposite of the black figure technique.

`CF(red-figure-technique, black-figure-technique),`

`CB=red-figure-technique, CONTINUE`

e. This exhibit was decorated by an artist known as “the painter of Kleofrades”.

`CF(exhibit1, painter-of-Kleofrades),`

`CB=exhibit1, NOCB`

f. “The painter of Kleofrades” used to decorate big vases.

`CF(painter-of-Kleofrades, entity-4049),`

`CB=painter-of-Kleofrades, CONTINUE`

The structure in (2.10) has one NOCB transition in position (2.10e). Using the preference which underlies C1 absolutely condemns the structure as incoherent since the NOCB transitions are not minimised between all pairs of utterances.

However, simply summing up the pairs of utterances that violate C1 is not enough for estimating the coherence of a text like (2.10). What one needs to do in addition is to search through the search space of possible orderings to investigate whether there exists a structure that minimises the observed violations. Profiling the search space in such a way for (2.10) reveals that there are no orderings with

[53] See Dimitromanolaki and Androutsopoulos (2003) and section 7.2 of chapter 7 for more details on this corpus.

zero NOCBs! The structure in (2.10) is a member of a set of 18 orderings, all of which include a pair of utterances violating C1. Crucially, this is the best that one can get with the set of utterances in (2.10) as far as C1 is concerned, since 85% of the alternative sequences return more NOCBs than (2.10). Again, regarding the preference behind C1 absolutely fails to account for this fact which can only emerge if alternative sequences of the utterances in (2.10) are considered.

With respect to the preference behind R2, 10 of the sequences that optimise on the number of NOCBs return fewer CONTINUEs than (2.10). Therefore, there are only 7 alternative sequences of utterances (5.83%) that have the same transitions as (2.10). This time, the number of CONTINUEs is found to positively discriminate (2.10) against its alternatives.

Similarly to what was argued in section 2.5 above, the discussion in this section points out the difficulty of using R2 incrementally for generating an order for the utterances in (2.9), as such an algorithm might be confronted with a locally worst, yet globally best, choice at any time during the text structuring process.

It is then the search-based approach to text structuring reviewed in section 2.1.5 that makes use of the required global measure and represents a more appropriate alternative than the deterministic use of R2. Hence, it seems that defining a scoring function of entity coherence such as the one provided by Poesio et al. (2002) appears to be the most appropriate solution for the problem of how CT should be formulated for the purposes of text structuring. And it should not be a coincidence that, to our knowledge, the only two implemented versions of CT for text structuring, namely Kibble and Power (2000) and the genetic algorithm in Cheng (2002), do make use of the search-based approach (although, as the next chapter points out, the ways that CT is formulated there represent only two of the many possible options).

Going back to the point raised in the previous section, when the sequence of utterances in (2.9) and (2.10) are compared with their alternatives, computing the sum of CONTINUEs and the sum of NOCBs provides a more informative estimate of their entity coherence than comparing these sums with absolute preferences.

Clearly, a more extended search-oriented operation such as the one presented in this section is required in order to specify how general the phenomena in (2.9) and (2.10) are.[54] In chapter 5, we pursue the issue further by introducing a search oriented, corpus-based methodology especially devised for our purposes. Although we indicate that determining how coherent a structure is (compared to the alternatives for the same content) requires search, our empirical research cannot investigate this issue exhaustively. Instead, we restrict ourselves to raising the question with respect to the standard

[54] It should be clear to the reader that we refer to two distinct uses of search: (a) the one that is presented in this section that is used **prior** to the actual generation, but is essential for a scoring function to be seen as a robust estimator of the entity coherence of a complete text structure (b) the one used **during** the actual generation in systems such as Mellish et al. (1998a) to output the best scoring solution. Obviously, (a) is the use that the thesis is concerned with.

corpus-based evaluation of CT, but attempt to explore it under a specific NLG perspective.

2.6 A research question for search-based text structuring

In summary, the previous section shows that CT needs to be extended in order to:

a) Resolve RSI in favour of the RETAIN:SHIFT pattern (section 2.5.1).

b) Estimate the entity coherence of the whole text (section 2.5.2).

c) Examine alternative sequences of utterances (section 2.5.3 and section 2.5.4).

The discussion points out that it is difficult to use R2 directly on an utterance-by-utterance basis to generate a complete text structure. An alternative to the incremental use of R2 is provided by the global scoring functions typically used in the corpus-based evaluation of the theory. Hence, it seems to us that defining a scoring function of entity coherence appears to be the most plausible way of representing CT for the the purposes of text structuring.

Although we have already introduced the scoring function of entity coherence from Poesio et al. (2002), we feel that this might not be the only possibility, given the many different ways of formulating CT. Thus, the first research question that our work addresses is the following:[55]

Q1: How can CT be used to define an evaluation metric of entity coherence for search-based descriptive text structuring?

In an attempt to discuss in detail a number of potential solutions to question (Q1), we begin the following chapter with a review of some metrics of entity coherence already in use for the purposes of text structuring including the ones in Mellish et al. (1998a), Cheng (2002) and Kibble and Power (2000). Then, we define additional metrics based on the different specifications of CT.

[55] See section 2.1.5 and the beginning of the next chapter for a clarification of the difference between a scoring function and an evaluation metric.

Chapter 3

Defining metrics of entity coherence

The previous chapter argues that the most appropriate way to use CT to aid the text structuring process is by defining a metric of entity coherence. This chapter discusses possible ways of defining such a metric, starting with an investigation of existing metrics of text structure that use notions from CT. Then, we define additional metrics of entity coherence based on the different formulations of CT. We conclude the chapter with the next question that our empirical work needs to investigate:

Q2: Which metrics of entity coherence constitute the most promising candidates for text structuring?

3.1 Existing metrics of entity coherence

We define a *scoring function* of entity coherence as a simple function that returns a score (or a set of scores) S for the entity coherence of a text structure T. An *evaluation method* of entity coherence uses S to compare T with one or more alternatives. When the scoring function is supplemented with an evaluation method, then it constitutes an *evaluation metric* of entity coherence.

In the previous chapter, we discussed one of the informal CT-based scoring functions of Poesio et al. (2002). This scoring function sums up the number of transitions in a text and can be formalised as follows for our purposes:[1]

- Scoring function in Poesio et al. (2002):

  Sum(NOCB), Sum(CONTINUE), Sum(RETAIN), Sum(SMOOTH-SHIFT), Sum(ROUGH-SHIFT)

Poesio et al. (2002) evaluate the different ways of specifying the parameters of CT, among other criteria, according to the extent that they minimise the sum of NOCBs and maximise the sum of CONTINUEs

[1] As mentioned in the previous chapter, we do not follow Poesio et al. (2002) in distinguishing NOCBs into ZERO and NULL transitions. A plus (+) is used to denote that the scores which are calculated by the scoring function are added up (see section 3.1.2 for an example), whereas a comma denotes that the set of scores is passed to the evaluation method without being added up.

in the GNOME corpus. In section 2.5.4 of the previous chapter, we presented two examples of applying this function, and argued that this method of evaluation does not account for the problem of choice which is particularly important from an NLG viewpoint. Chapter 5 presents such a search-oriented, corpus-based methodology especially devised from an NLG perspective.

Before entering the discussion in chapter 5, however, we feel that there exists another question that remains unaddressed: Are the suggestions in Poesio et al. (2002) the **only possible** CT-based scoring functions of entity coherence?

In this section, we start investigating that question by reviewing other existing scoring functions as well as complete evaluation metrics of entity coherence that have been used to guide the text structuring process in NLG. As we mentioned in section 2.1.5 of the previous chapter, these metrics are associated with a view of text structuring as a formal search problem where the metric is used to select a candidate solution between its alternatives.

3.1.1 Metrics in stochastic ILEX

As we mentioned in section 2.1.5 of the previous chapter, Mellish et al. (1998a) define an intuitive scoring function which employs entity-based features of coherence as well as other parameters of text quality. Some of the entity-based features of this function draw upon CT, although Mellish et al. (1998a) do not make any direct reference to it. They acknowledge, however, that integrating a formal model of entity coherence with their approach would be worthwhile.

Cheng (2002) builds upon the remarks of Mellish et al. (1998a) by presenting a genetic algorithm which handles the interaction between text structuring and aggregation in the ILEX domain. Cheng uses her own intuitions to specify preferences for rhetorical relations, entity coherence, aggregation, and their interactions (Cheng 2002, p.127). Her function extends the scoring scheme of Mellish et al. (1998a) with features weighted according to these preferences (Cheng 2002, pp.186-188). A series of evaluation experiments are employed to show that the intuitions underlying her scoring function are correct (Cheng 2002, Chapter 8).

Cheng formulates entity coherence using standard CT, weighting the transitions in her scoring function according to the preferences of R2.[2] Because the main focus of Cheng (2002) is on the interaction of aggregation with text structuring, replicating the scoring function for entity coherence outside her stochastic system using her exact weights would not make much sense. Instead, the unconditional preferences of standard R2 are captured in terms of Optimality Theory (OT) without specific numerical weights (see section 3.4.1 for more details).

[2] A novel transition called ASSOCIATE SHIFT which captures a bridging relation between two entities in adjacent utterances is introduced in the formulation of CT transitions by Cheng (2002, p.142). Since we do not investigate the effect of indirect realisation we take the preferences between transitions in Cheng (2002) to be equivalent to the ones in standard R2.

3.1.2 Summing up the underlying notions

Kibble and Power (2000) use the reformulation of CT into the prerequisite of CONTINUITY and the three underlying principles of CT (namely COHERENCE, CHEAPNESS and SALIENCE) in the definition of their scoring function of entity coherence.[3] This function sums up the number of times each candidate structure violates each CT notion and then adds the four resulting sums together. Sum(NOCB), Sum(COH∗), Sum(CHEAP∗) and Sum(SAL∗) stand for the sums of the violations of CONTINUITY, COHERENCE, CHEAPNESS and SALIENCE respectively:

i. Scoring function of entity coherence in Kibble and Power (S.KP):

   Sum(NOCB)+Sum(COH∗)+Sum(CHEAP∗)+Sum(SAL∗)

In order to explain how S.KP works more clearly, let us apply it to examples (3.1) and (3.2), assuming that these examples correspond to two (of the) candidate solutions for text structuring. The utterances in the examples are annotated with the violations of the underlying notions of CT in addition to the standard CT transitions.[4] For instance, the ROUGH-SHIFT in (3.1d) violates all three underlying principles of CT, only satisfying the prerequisite of CONTINUITY:

(3.1) a. This exhibit is an amphora.

`CF(exhibit1, amphora)`

b. Amphoras have an ovoid body and two looped handles, reaching from the shoulders up.

`CF(amphora, entity-3908),`

`CB=amphora, CONTINUE`

CHEAP∗

c. Amphoras were produced in two major variations: type A and the type with a neck.

`CF(amphora, typeA, type-neck)`

`CB=amphora, CONTINUE`

d. This exhibit is a type A amphora.

`CF(exhibit1, typeA)`

`CB=typeA, ROUGH-SHIFT`

[3] Following the terminology in section 2.2.4 of chapter 2, we refer collectively to the three underlying principles and their prerequisite as the underlying notions of CT.

[4] Example (3.1) is the same as example (2.9) in section 2.5.4 of chapter 2. Clearly all metrics presented in this chapter can apply to more than two candidate structures. In this chapter, the discussion is restricted to only two examples for the sake of simplicity.

COH∗, CHEAP∗, SAL∗

e. This exhibit comes from the archaic period.

CF(exhibit1, archaic-period)

CB=exhibit1, SMOOTH-SHIFT

COH∗

f. This exhibit was painted using the red figure technique.

CF(exhibit1, red-figure-technique)

CB=exhibit1, CONTINUE

In (3.2), utterance (b) appears between (d) and (e). This creates two violations of CONTINUITY in positions (3.2b) and (3.2e):

(3.2) a. This exhibit is an amphora.

CF(exhibit1, amphora)

c. Amphoras were produced in two major variations: type A and the type with a neck.

CF(amphora, typeA, type-neck)

CB=amphora, CONTINUE

CHEAP∗

d. This exhibit is a type A amphora.

CF(exhibit1, typeA)

CB=typeA, ROUGH-SHIFT

COH∗, CHEAP∗, SAL∗

b. Amphoras have an ovoid body and two looped handles, reaching from the shoulders up.

CF(amphora, entity-3908)

NOCB

e. This exhibit comes from the archaic period.

CF(exhibit1, archaic-period)

NOCB

f. This exhibit was painted using the red figure technique.

CF(exhibit1, red-figure-technique)

CB=exhibit1, CONTINUE

	Violations of Centering notions				S.KP
Text	NOCB	COH∗	CHEAP∗	SAL∗	Total
(3.1)	-	d, e	b, d	d	5
(3.2)	b, e	d	c, d	d	6

Table 3.1: Violations of Centering notions and scores for examples (3.1) and (3.2) according to the scoring function S.KP

Table 3.1 summarises the violations of the underlying notions of CT for the two examples.[5] The last column of the Table reports the total number of violations for each example which corresponds to the score returned by S.KP. As Table 3.1 shows, the total number of violations in (3.1) is 5, whereas the total number of violations in (3.2) is 6. Thus, the structure with the smaller sum of violations of the CT notions is (3.1). If S.KP is supplemented with an evaluation method which selects the structure with the smallest number of CT violations as a better solution for text structuring than its competitor(s), then (3.1) wins the competition with (3.2).

The combination of the S.KP scoring function with the preference for the structure with the smallest number of violations of the CT notions gives rise to the following metric of entity coherence, which we call **M.KP**:

- Metric of entity coherence employing S.KP (M.KP):
  - **scoring function**:
    Sum(NOCB)+Sum(COH∗)+Sum(CHEAP∗)+Sum(SAL∗)
  - **evaluation method**:
    Prefer the solution with the lowest score

3.1.3 Isolating the effects of entity coherence

The scoring function of entity coherence in Kibble and Power (2000) is part of a larger evaluation module that applies a battery of tests to a restricted set of candidate solutions and selects the one with the lowest total cost. Kibble and Power (2000) claim that a candidate solution that does worse than another competitor according to S.KP can still be selected over its alternative if it is assigned with a better score for certain stylistic preferences. These preferences are related to favourable ways of realising the underlying rhetorical structure of a candidate text structure.

[5]The columns in the middle of Table 3.1 are headed by the violations of the CT notions. Following the approach of Beaver (2003), quoted in section 2.5.2 of chapter 2, the cells beneath each violation report the utterances in which the violation occurs for each example.

As in Mellish et al. (1998a), the interaction between entity and rhetorical coherence in Kibble and Power (2000) is specified intuitively, a point clearly acknowledged by the authors. Hence, these metrics of entity coherence for text structuring provide reasonable indications for, rather than a complete solution to, the complex issue of how different models of coherence interact with each other.

As we have clearly mentioned section 2.4.4 of the previous chapter, specifying a model for this interaction is beyond the scope of this thesis. Instead, we attempt to estimate the importance of entity coherence on characterising a descriptive text structure. Hence, treating S.KP in isolation for the definition of M.KP seems appropriate.

This does not mean, however, that the effects of other coherence-inducing mechanisms are completely ignored in our work. In chapter 6, we investigate the interaction of rhetorical and entity coherence in a subset of the GNOME corpus based on the assumptions of Knott et al. (2001). In chapter 8, we discuss how an additional constraint on entity coherence can supplement our general experimental approach.

All in all, the main argument in this chapter is that the different formulations of CT give rise to many metrics of entity coherence, M.KP being one of the numerous possibilities. Identifying the most appropriate metrics for text structuring is an empirical issue as our experiments in subsequent chapters show.

3.2 An OT ranking of the underlying principles

Although Kibble and Power (2000) mention that each CT notion that their metric employs may be assigned a different cost, in practice they decide that all of them be weighted equally. However, this decision does not stem from any empirical finding or theoretical claim of CT.

It is notable that calculating the sum of violations for each underlying notion of CT in S.KP is similar to the way Beaver (2003) computes the violations of COT constraints. The main difference is that COT violations are *ranked* in a standard OT fashion instead of being summed up. Another difference is that COT only employs two of the underlying principles of CT (that is, COHERENCE and SALIENCE as reported in section 2.2.4.2 of chapter 2). Thus, as shown by Beaver (2003), COT is equivalent to standard CT.

A metric of entity coherence based on standard CT is defined in section 3.4.1. This section argues in favour of an **OT ranking** of the CT notions. To support this argument we first remind the reader of the RETAIN:SMOOTH-SHIFT *inadequacy of R2* (RSI) as discussed in section 2.5.1 of chapter 2. Although RSI cannot be resolved when the underlying principles of CT are taken to be of equal importance as in S.KP, a simple solution to RSI can be defined by considering the three underlying principles of CT as ranked violable constraints in the sense of OT. This, in turn, gives rise to an eval-

uation method that ranks the sums of the violations of the CT principles according to the preference order defined for the resolution to RSI.

In the remainder of the section, we discuss some alternatives for the ranking of the CT notions. Then, we define metrics that further reduce the complexity of the CT framework by considering only a few of the underlying notions as contributors to the overall coherence of the text.

The chapter continues with the definition of a metric closer to the standard formalisation of R2 and a detailed discussion of less explored aspects of the relation between CT principles and transitions. We conclude by defining the next aim of our research as the comparison of some of these metrics on an empirical basis.

3.2.1 Resolving RSI

In order to motivate the reformulation of M.KP in terms of OT, let us repeat example (2.8) from section 2.5.1 of chapter 2 as example (3.3), annotated with the violations of the underlying principles of CT in addition to the standard CT transitions:

(3.3) a. This exhibit is an amphora.

`CF(exhibit1, amphora)`

b. This exhibit comes from the archaic period.

`CF(exhibit1, archaic-period)`

`CB=exhibit1, CONTINUE`

c. This exhibit was decorated by an artist known as the “painter of Kleofrades”.

`CF(exhibit1, painter-of-Kleofrades)`

`CB=exhibit1, CONTINUE`

d. The “painter of Kleofrades” used to decorate big vases.

`CF(painter-of-Kleofrades, entity-4049)`

`CB=painter-of-Kleofrades, SMOOTH-SHIFT`

COH∗, CHEAP∗

c’. An artist known as the “painter of Kleofrades” decorated this exhibit.

`CF(painter-of-Kleofrades, exhibit1)`

`CB=exhibit1, RETAIN`

SAL∗

d’. The “painter of Kleofrades” used to decorate big vases.

`CF(painter-of-Kleofrades, entity-4049)`

CB=painter-of-Kleofrades, SMOOTH-SHIFT

COH∗

As we discussed in section 2.5.1, RSI emerges from the inability of R2 to resolve the competition between (3.3c-d) and (3.3c'-d') in favour of the RETAIN:SMOOTH-SHIFT sequence in (3.3c'-d'). Note that since both (3.3a-b-c-d) and (3.3a-b-c'-d') violate the underlying principles two times in total, M.KP cannot discriminate between them at all.

Kibble and Power (2000) suggest that the principle of CHEAPNESS can provide a solution to RSI. Indeed, the RETAIN:SMOOTH-SHIFT sequence in (3.3c'-d') is one of the cheap transition pairs according to Strube and Hahn (1999), by contrast to the CONTINUE: (EXPENSIVE) SMOOTH-SHIFT sequence in (3.3c-d) which belongs to the expensive transition pairs. Since cheap transition pairs are preferred over expensive ones, the preference for (3.3c'-d') over (3.3c-d) is predicted. However, as Kibble (2001) mentions, the definition of 36 transition pairs is unnecessarily complicated, since its only purpose is to define the priority of the principle of CHEAPNESS over the other principles of CT.

In this section, we propose a solution to RSI by defining an OT ranking **directly** on the three underlying principles of CT. That is, we view the principles of CT as violable constraints in an OT way and rank them according to the following preference which we will conventionally name POT1 (from CT *Principles in OT - Ranking 1*):[6]

ii. CT Principles in OT - Ranking 1 (POT1):

COHERENCE>CHEAPNESS>SALIENCE

This preference states that violating COHERENCE is more serious than violating CHEAPNESS which, in turn, is more serious than violating SALIENCE. Ranking COHERENCE over SALIENCE is motivated by the reanalysis of the table of standard CT transitions in terms of these two principles, discussed in section 2.2.4.2 of chapter 2, and the precedence of the RETAIN transition over a SMOOTH-SHIFT as defined by standard R2. In what follows, we explain how ranking CHEAPNESS over SALIENCE resolves RSI in favour of the RETAIN:SMOOTH-SHIFT pattern.

The violations of the underlying principles in the competing sequences of utterances in (3.3) are summarised in Table 3.2. According to OT (Prince and Smolensky 1997), the optimal structure is the one that returns fewer violations of the most highly ranked constraint on which the competing structures differ. Because both configurations in Table 3.2 violate COHERENCE the same number of times, this constraint cannot decide the competition. Crucially, the RETAIN in (3.3c') violates SALIENCE whereas the SMOOTH-SHIFT in (3.3d) violates CHEAPNESS. Since CHEAPNESS is ranked higher than SALIENCE, POT1 decides the competition in favour of (3.3a-b-c'-d') that contains the

[6]Note that no crosslinguistic predictions are associated with this ranking, contrary to standard assumptions in OT.

Text	Violations of Centering principles		
	COH∗	CHEAP∗	SAL∗
(3.3a-b-c-d)	d	d	
(3.3a-b-c'-d')	d'		c'

Table 3.2: Violations of Centering principles in the utterances of example (3.3)

RETAIN:SMOOTH-SHIFT sequence. Arguably, ranking the CT principles directly provides a much more straightforward solution to RSI than resorting to the 36 pairs of transitions in FC.

A new metric of entity coherence can be defined on the basis of the solution to RSI discussed so far. In order to emphasise its relation with the ranking defined by POT1, the metric bears the name **M.POT1**. The scoring function of this metric computes the sums of the violations of the four CT notions. However, instead of adding up the sums as in M.KP, the scoring function of M.POT1 communicates with an OT-like evaluation method.

The evaluation method of M.POT1 examines the sums of violations of the three underlying principles in a way that abides with the resolution to RSI. In addition, the sum of violations of the prerequisite of CONTINUITY is ranked as the most important penalty. This is motivated by the fact that if CONTINUITY is violated between two utterances one cannot apply any of the underlying principles to the second member of the pair.[7]

- RSI-motivated metric of entity coherence (M.POT1):
  - **scoring function**:
    Sum(NOCB), Sum(COH∗), Sum(CHEAP∗), Sum(SAL∗)
  - **evaluation method**:
    Sum(NOCB)>Sum(COH∗)>Sum(CHEAP∗)>Sum(SAL∗)

Applying M.POT1 to examples (3.1) and (3.2) works as follows: The most serious violations of entity coherence in Table 3.1 are the NOCB transitions, each corresponding to a violation of CONTINUITY. There are no NOCBs in (3.1), while CONTINUITY is violated twice in (3.2). Hence, (3.1) is optimal when compared to (3.2) using M.POT1 because it returns less violations of the most highly ranked constraint.

Note that according to mainstream OT no other lower ranked constraint violated by either of the candidates is taken into account by the evaluation method of M.POT1 in determining the winner.

[7]Note that the prioritising CONTINUITY is expressed purely in the way that the sums of violations are inspected by the evaluation method. If CONTINUITY is violated by the second member of a pair of utterances, the scoring function of M.POT1 returns 0 (that is, no violation) for each underlying principle of CT, thus avoid penalising the utterance more than once for the same defect.

Only the most highly ranked constraint on which the two candidates differ (in this case CONTINUITY) matters. The sum of violations of COHERENCE will only be considered if the candidate structures are found to have the same number of NOCBs. If the sum of violations of COHERENCE is the same as well, then the violations of CHEAPNESS will be considered to resolve the competition between the structures and so on. If the candidate structures have the same number of violations for each CT notion, then they are considered equivalent by M.POT1.

Crucially, the competition would be decided in favour of example (3.2), were COHERENCE ranked as the most serious constraint over CONTINUITY. This shows that the ranking of the CT notions is very important in the way that an OT-inspired evaluation metric of entity coherence works. This issue is discussed in the next section in more detail.

3.2.2 Alternative POT rankings

As we mentioned in the previous chapter, the argumentation in favour of the RETAIN:SHIFT pattern is not infallible. Thus, the evaluation method of M.POT1 employs only one of the possible ways of ranking the underlying notions of CT in OT terms. This ranking needs to be subjected to empirical justification in the same way as the decision to sum up the violations in S.KP.

Indeed, Kibble (2001) argues against the priority of COHERENCE over SALIENCE. He tentatively suggests the following preference in order to resolve the conflicts between the underlying principles.

iii. Kibble's ranking of CT Principles:

$\{$CHEAPNESS, SALIENCE$\}$ >COHERENCE

According to this ranking, the violations of CHEAPNESS and SALIENCE have the same effects on the entity coherence of the discourse. Moreover, violating either of these constraints is more serious than violating COHERENCE.

Alongside Kibble (2001), we believe that the exact ranking of the notions of CT remains an open question. As we have already mentioned in the previous section, a very interesting parameter in addition to the ones considered by Kibble (2001) is the ranking of the sum of NOCB transitions. Table 3.3 shows how two extreme possibilities for ranking of the sum of NOCBs give rise to 11 metrics in addition to M.POT1.

The metrics in Table 3.3 can be divided into two sets: Like M.POT1, the sum of NOCBs is the first score to be examined in metrics M.POT2 to M.POT6. However, these metrics differ in the way that the evaluation method prioritises the sums of violations of the CT principles. In all remaining metrics the sum of NOCB transitions is the least important violation. The last choice reflects the view of standard CT which does not discuss the effects of the NOCB transition on the coherence of

Metric	Evaluation method
M.POT2	Sum(NOCB)>Sum(COH∗)>Sum(SAL∗)>Sum(CHEAP∗)
M.POT3	Sum(NOCB)>Sum(SAL∗)>Sum(CHEAP∗)>Sum(COH∗)
M.POT4	Sum(NOCB)>Sum(SAL∗)>Sum(COH∗)>Sum(CHEAP∗)
M.POT5	Sum(NOCB)>Sum(CHEAP∗)>Sum(COH∗)>Sum(SAL∗)
M.POT6	Sum(NOCB)>Sum(CHEAP∗)>Sum(SAL∗)>Sum(COH∗)
M.POT7	Sum(COH∗)>Sum(CHEAP∗)>Sum(SAL∗)>Sum(NOCB)
M.POT8	Sum(COH∗)>Sum(SAL∗)>Sum(CHEAP∗)>Sum(NOCB)
M.POT9	Sum(SAL∗)>Sum(CHEAP∗)>Sum(COH∗)>Sum(NOCB)
M.POT10	Sum(SAL∗)>Sum(COH∗)>Sum(CHEAP∗)>Sum(NOCB)
M.POT11	Sum(CHEAP∗)>Sum(COH∗)>Sum(SAL∗)>Sum(NOCB)
M.POT12	Sum(CHEAP∗)>Sum(SAL∗)>Sum(COH∗)>Sum(NOCB)

Table 3.3: Some alternative rankings of the violations of Centering notions

the discourse. Again, each metric examines the sums of violations of the underlying principles in a different order of priority.

The metrics in Table 3.3 were not used in the experiments reported in the chapters that follow. In order to somehow contain the effort, we only considered M.POT1 that was given theoretical priority because of its relation to the resolution of RSI, which, in our view, is a very interesting debate within CT. Nevertheless, we regard experimenting with different rankings of the underlying principles more extensively as an appealing direction for empirical future work.[8]

3.3 Simpler metrics of entity coherence

This section presents metrics that are simpler than M.KP and M.POT1 in the sense that they employ fewer (combinations of) CT principles. This is based on the assumption that only specific violations of CT notions cause incoherence in a structure.

3.3.1 Computing only the violations of CHEAPNESS

According to Strube and Hahn (1999) the principle of CHEAPNESS is to be given complete priority over the other notions of CT for the purposes of anaphora resolution. We are interested to see how suitable a metric based on the formulation of R2 in FC (see section 2.2.4.3 of chapter 2) could be

[8]This could be extended to the remaining 12 possible ways of ranking the underlying principles not illustrated in Table 3.3.

for the text structuring task. For this reason we devise the following metric that computes the sum of violations of CHEAPNESS only and compares the candidate structures according to these scores (**M.CHEAP**):

- FC metric of entity coherence (M.CHEAP):
  - **scoring function**:
    Sum(CHEAP$*$)
  - **evaluation method**:
    Prefer the solution with the lowest score

Note that if CHEAPNESS is used as the only criterion for evaluating the entity coherence of (3.1) and (3.2), then both structures are considered to be equivalent since each violates CHEAPNESS twice (see Table 3.1).

3.3.2 Computing only the violations of CONTINUITY

Under a slightly more radical view of entity coherence one can do away with the underlying principles of CT and only rely on the prerequisite of CONTINUITY. **M.NOCB** is the metric that uses the sum of violations of CONTINUITY only, ignoring all other aspects of entity coherence as defined by the three CT principles:

- Metric of entity coherence based only on CONTINUITY (M.NOCB):
  - **scoring function**:
    Sum(NOCB)
  - **evaluation method**:
    Prefer the solution with the lowest score

Looking at Table 3.1 again, the evaluation method of M.NOCB favours (3.1) over (3.2) as the structure that violates CONTINUITY less times. Hence, the result for this example is the same as using M.POT1. The relationship between M.POT1 and M.NOCB is defined more generally as follows: Assuming that M.NOCB returns t_1 NOCB transitions for structure T_1 and t_2 NOCB transitions for structure T_2, when $t_1 < t_2$, then T_1 is preferred over T_2 by the evaluation methods of both M.NOCB and M.POT1. Conversely, if $t_1 > t_2$, then both metrics favour T_2. Thus, it is only when $t_1 = t_2$ that using M.POT1 might return different results from using M.NOCB.

When $t_1 = t_2$, the evaluation method of M.NOCB considers T_1 to be equivalent to T_2 with respect to their entity coherence. However, as we mentioned at the end of in section 3.2.1, the evaluation

method of M.POT1 does not rest its case so easily. Instead, the sums of violations of the three underlying principles of CT are consulted in the order defined by their ranking in POT1. The next most highly ranked CT constraint on which the two candidates differ decides the competition. T_1 and T_2 are considered equivalent by the evaluation method of M.POT1 only if its scoring function returns the same scores for each violation of the underlying notions of CT.

As we discussed in section 2.3.2 of chapter 2, the sum of NOCBs is used by Poesio et al. (2002) as the measure of violations of C1 in the evaluation of different versions of CT. This metric is also used in the stochastic text structuring system of Karamanis and Manurung (2002). Because M.NOCB only uses the prerequisite of CONTINUITY, it is recognised as the simplest among the metrics discussed in this chapter.[9] As we mention in chapter 5, because of its simplicity, M.NOCB serves as the **baseline** for our corpus-based experiments which aim to specify which of the different metrics are the most promising candidates for the purposes of text structuring.

3.3.3 ROUGH-SHIFT as a source of incoherence

A view of entity coherence closer to standard CT comes from Miltsakaki and Kukich (2000a,b). Miltsakaki and Kukich (2000a,b) supplemented a system for grading student essays with a measure of entity incoherence based on the percentage of ROUGH-SHIFT and NOCB transitions.[10] They show that this modification improves the accuracy of the grades generated by the system when compared with grades from human experts. Our next metric, namely **M.MIL**, is inspired by the combination of the sum of ROUGH-SHIFT and NOCB transitions as an estimate of incoherence in the scoring function of Miltsakaki and Kukich (2000a,b):[11]

- M.MIL metric:
  - **scoring function**:

    Sum(NOCB)+Sum(ROUGH-SHIFT)
  - **evaluation method**:

    Prefer the solution with the lowest score

[9] M.CHEAP also employs only one CT notion, namely CHEAPNESS, but for CHEAPNESS to be satisfied or violated CONTINUITY needs to apply first. For this reason, M.NOCB is considered to be simpler than M.CHEAP.

[10] Because of the impreciseness of standard CT about the NOCB transition discussed in section 2.2.4.1 of chapter 2, Miltsakaki and Kukich (2000a,b) only talk about ROUGH-SHIFT transitions. That is, they do not make the distinction between a violation of CONTINUITY, represented by a NOCB transition and a ROUGH-SHIFT where CONTINUITY is preserved but both COHERENCE and SALIENCE are violated (i.e $Cb(U_n) \neq Cb(U_{n-1}) \neq Cp(U_n)$ as shown by the definition of ROUGH-SHIFT in Table 2.1 in section 2.2.1 of chapter 2). That Miltsakaki and Kukich (2000a,b) include the violations of CONTINUITY in their scoring function is clear from the discussion of the example of an incoherent essay in Miltsakaki and Kukich (2000a), the second transition of which does not have a CB and is marked as a ROUGH-SHIFT.

[11] As chapter 5 explains, each metric which is subject to our experimental methodology evaluates structures of the same length. For this reason, we can define M.MIL directly in terms of the sums of ROUGH-SHIFT and NOCB transitions rather than their percentages as done by Miltsakaki and Kukich (2000a,b).

Text	NOCB	ROUGH-SHIFT	M.MIL
(3.1)	-	d	1
(3.2)	b, e	d	3

Table 3.4: Scores for examples (3.1) and (3.2) according to metric M.MIL

The last column of Table 3.4 reports the scores of (3.1) and (3.2) according to the scoring function of M.MIL. Example (3.1) has a ROUGH-SHIFT transition in utterance (3.1d) and no violations of CONTINUITY, so its overall score is 1. Example (3.2) has a ROUGH-SHIFT transition in utterance (3.2d) and two violations of CONTINUITY in (3.2b) and (3.2e), so its score is 3. Example (3.1) is preferred over (3.2) according to the evaluation method of M.MIL as the structure that returns a lower score for entity incoherence.

3.3.4 What is a SHIFT?

The reader might recall from the discussion of standard CT in the previous chapter that Grosz et al. (1995) do not follow the distinction between SMOOTH- and ROUGH-SHIFT as introduced by Brennan et al. (1987). For them, the second member of a pair of utterances that violates COHERENCE is simply classified as a SHIFT. In other to accommodate for this definition of SHIFT, we modify the scoring function of M.MIL so that instead of the sum of ROUGH-SHIFTs it computes the sum of all utterances that violate COHERENCE in addition to the violations of CONTINUITY.[12] The resulting metric is called **M.SH**:

- Revision of M.MIL using SHIFTs (M.SH):
  - **scoring function**:
    Sum(NOCB)+Sum(COH$*$)
  - **evaluation method**:
    Prefer the solution with the lowest score

M.SHOT1 is a version of M.SH in which violating CONTINUITY is considered to be more severe than violating COHERENCE. To express this we devise the evaluation method of M.SHOT1 so that the sums of violations are considered in the order defined by the evaluation method in M.POT1:

- M.POT1 revision of M.SH (M.SHOT1):

[12] Clearly, a violation of COHERENCE (COH$*$) corresponds to the SHIFT transition between two utterances. See Table 2.2 in section 2.2.4.2 of chapter 2.

	M.SHOT1		M.SH
Text	NOCB	COH∗	Total
(3.1)	-	d, e	2
(3.2)	b, e	d	3

Table 3.5: Scores for examples (3.1) and (3.2) according to metrics M.SH and M.SHOT1

– **scoring function**:

  Sum(NOCB), Sum(COH∗)

– **evaluation method**:

  Sum(NOCB)>Sum(COH∗)

Table 3.5 reports the scores for examples (3.1) and (3.2) according to M.SH and M.SHOT1. The total number of NOCBs and SHIFTs for (3.1) is 2, whereas the total for (3.2) is 3. Hence, M.SH decides in favour of (3.1). The same is true for M.SHOT1 which only considers the violations of CONTINUITY in the same way as M.POT1 and M.NOCB. The difference between M.POT1, M.NOCB and M.SHOT1 has to do with the number of additional scores that the evaluation method takes into account when the candidate structures have the same scores for the number of NOCBs.

3.4 Transition-based metrics of entity coherence

The metrics presented so far are based on the logic that if the scoring function counts violations of CT notions or incoherent transitions, examples (3.1) and (3.2) should be compared with respect to the extent they minimise these violations. However, the standard formulation of R2 places emphasis on **maximising** preferred transitions such as CONTINUE, instead of **minimising** violations.

Although recasting standard transitions in terms of (some of) the underlying principles simplifies the CT framework (see section 2.2.4.2 of chapter 2), using transitions in the scoring function of an evaluation metric of entity coherence is not completely out of place. Note that metrics such as M.KP and M.POT1 are agnostic with respect to the occurrence of violations of different underlying principles in the same utterance. By contrast, certain formulations of CT define a variety of transitions as a vocabulary for the way that such violations are combined in the same utterance.

In this section, we discuss metrics of entity coherence that employ transitions as a way to express certain combinations of the underlying principles of CT. Like the scoring function of Poesio et al. (2002) which is used for the evaluation of different versions of CT on the basis of R2, the scoring function of these metrics computes the **sum of transitions** in a candidate structure. Then, their evaluation method promotes the structure with the highest number of preferred transitions.

We start by ranking the sums of transitions according to the preferences of R2. Then, we investigate the relation between the underlying principles and the various transitions of CT in more detail.

3.4.1 The BFP metric

The scoring function of the BFP metric (**M.BFP**) computes the sums of standard CT transitions in a structure. Then, its evaluation method compares the candidate structures by examining the sum of transitions in the order specified by the standard formulation of R2:

- Standard transition-based metric of entity coherence (M.BFP):
  - **scoring function**:
    Sum(CONTINUE), Sum(RETAIN), Sum(SMOOTH-SHIFT), Sum(ROUGH-SHIFT)
  - **evaluation method**:
    Sum(CONTINUE)>Sum(RETAIN)>Sum(SMOOTH-SHIFT)>Sum(ROUGH-SHIFT)

Table 3.6 shows the standard CT transitions for (3.1) and (3.2). This time, the structures are compared with respect to the number of coherent transitions in the order indicated by the preferences of R2. The first score to be examined is the sum of CONTINUE transitions. Because (3.1) has more CONTINUEs than (3.2), it is declared to be the winner of the competition in one go. Only if the two structures were found to have the same number of CONTINUEs would the sum of RETAINs be examined. Moreover, the fact that (3.1) has one SMOOTH-SHIFT more than (3.2) is irrelevant, since the sum of SMOOTH-SHIFTs is examined only when the structures have the same scores for the two more highly ranked transitions.

The evaluation method of M.BFP resembles the OT-inspired evaluation method of a principle-based metric such as M.POT1, with the exception that the scoring function of M.POT1 counts sums of violations of CT notions and its evaluation method prefers the candidate structure with the **smallest** number of the most severe violation. By contrast, the scoring function of M.BFP counts sums of transitions and its evaluation method gives preferences to the structure with the **highest** number of the most preferred transition.

However, a scoring function that computes the sum of a **preferred** transition Sum(TRAN), can be trivially changed into a scoring function closer to standard OT that computes Sum(TRAN$*$), i.e. the number of **dispreferred** transitions other than TRAN. The relation between Sum(TRAN) and Sum(TRAN$*$) is given by the following equation:

iv. Sum(TRAN$*$)=(n$-$1)$-$Sum(TRAN)

Text	Centering Transitions			
	CONTINUE	RETAIN	SMOOTH-SHIFT	ROUGH-SHIFT
(3.1)	b, c, f	-	e	d
(3.2)	c, f	-	-	d

Table 3.6: Standard Centering transitions in examples (3.1) and (3.2)

where n stands for the number of utterances a structure consists of. Then, an evaluation method can be defined so that it prefers the structure with the smallest number of Sum(TRAN$*$).

For an evaluation metric such as M.BFP that employs more than one transition the scoring function could be redefined to return the list of scores of dispreferred transitions. The evaluation method then examines these scores accordingly, showing preference for the structure with the smallest number of the most highly ranked dispreferred transition:

v. Definition of M.BFP using dispreferred transitions:

- **scoring function**:

  Sum(CONTINUE$*$), Sum(RETAIN$*$), Sum(SMOOTH-SHIFT$*$), Sum(ROUGH-SHIFT$*$)

- **evaluation method**:

  Sum(CONTINUE$*$)>Sum(RETAIN$*$)>Sum(SMOOTH-SHIFT$*$)>Sum(ROUGH-SHIFT$*$)

Although the two ways of defining M.BFP are equivalent, the one in the beginning of the section is more straightforward and was the one used in our implementation of M.BFP.

Note that, as in Brennan et al. (1987), NOCBs are not taken into account for the definition of transitions in M.BFP. It should also be made clear to the reader that M.BFP is different from the principle-based metrics in the following sense: Assume that the number of NOCB transitions for structure T_1 is t_1 and the number of CONTINUEs is c_1. In addition, structure T_2 has t_2 NOCBs and c_2 CONTINUEs. If both $t_1 > t_2$ and $c_1 > c_2$ hold, T_1 will loose the competition with T_2 according to M.NOCB (and its extensions such as M.POT1 and M.SHOT1), but win it according to M.BFP.

As with the underlying principles of CT, the definition of transitions in Brennan et al. (1987) is not the only way of specifying transitions in CT. The metric M.GJW can be defined on the basis of the definition of transitions in Grosz et al. (1995) as follows:[13]

vi. Revision of M.BFP using only one SHIFT (M.GJW):

[13]Clearly, Sum(SHIFT)=Sum(SMOOTH-SHIFT)+Sum(ROUGH-SHIFT).

– **scoring function**:

  Sum(CONTINUE), Sum(RETAIN), Sum(SHIFT)

– **evaluation method**:

  Sum(CONTINUE)>Sum(RETAIN)>Sum(SHIFT)

However, M.GJW is not expected to return very different results from M.BFP because its evaluation method, like the one of M.BFP, considers the various sums of SHIFTs very late (if ever). For this reason, it is better to incorporate this transition with M.SH and M.SHOT1 as shown in section 3.3.4 above.

3.5 Examining the relation between principles and transitions

In this section, we explore the relationship between principles and transitions in more detail. First, we enhance the table of transitions in Strube and Hahn (1999), by defining the full set of basic transitions using all possible combinations of the three underlying principles.

This results in a large number of basic transitions, which can be subsequently merged with each other according to the number of underlying principles that they violate. These new transitions give rise, in turn, to more metrics of entity coherence. Like the different ways of ranking the sums of violations of the CT principles in section 3.2.2, these additional metrics of entity coherence have not been taken into account in our experimentation, the results of which are reported in subsequent chapters of the thesis.

However, they are included in this chapter as promising directions for empirical future work and as a way of exemplifying the *proliferation of CT-based metrics of entity coherence*. The chapter concludes with a discussion of how this this problem relates to our research aims and the underspecification of CT as discussed in Poesio et al. (2002).

3.5.1 Extending the table of FC transitions

As mentioned in the previous chapter, the table of transitions in FC (Strube and Hahn 1999, Table 20, p.333) is incomplete. More specifically, as Table 2.3 in section 2.2.4.3 of chapter 2 shows, CONTINUE and SMOOTH-SHIFT in FC satisfy CHEAPNESS. Violating CHEAPNESS defines two additional transitions, EXP. CONTINUE and EXP. SMOOTH-SHIFT. Note that CHEAPNESS does not apply to RETAIN and ROUGH-SHIFT.

Table 3.7 presents a more complete table of *basic Centering transitions* using all possible combinations of the three underlying principles (assuming that CONTINUITY holds). In this Table, a

Basic transition	Centering principles			N of violated principles
	COHERENCE	CHEAPNESS	SALIENCE	
CONTINUE	+	+	+	0
EXP. CONTINUE	+	*	+	1
RETAIN	+	+	*	1
EXP. RETAIN	+	*	*	2
SMOOTH-SHIFT	*	+	+	1
EXP. SMOOTH-SHIFT	*	*	+	2
ROUGH-SHIFT	*	+	*	2
EXP. ROUGH-SHIFT	*	*	*	3

Table 3.7: Basic Centering transitions using all combinations of the three Centering principles

Transition	Centering Principles		
	COHERENCE	CHEAPNESS	SALIENCE
ESTABLISHMENT	NOCB(U_{n-1})	+	+
EXP. ESTABLISHMENT	NOCB(U_{n-1})	*	+
RETAIN-ESTABLISHMENT	NOCB(U_{n-1})	+	*
EXP. RETAIN-ESTABLISHMENT	NOCB(U_{n-1})	*	*

Table 3.8: The various types of the ESTABLISHMENT transition

violation of an underlying principle is indicated with a "$*$". Satisfying a principle is indicated with a "$+$". The last column of the Table reports the number of violated principles for each basic transition.

Another thing to notice is that COHERENCE in FC and the standard formulation of CT is not a binary constraint. As Table 2.1 and Table 2.3 in chapter 2 show, COHERENCE holds both when Cb(U_n)=Cb(U_{n-1}) and when U_{n-1} does not have a NOCB. As discussed in section 2.2.4.1 of chapter 2, the additional transition ESTABLISHMENT is often introduced to account for these cases. In Table 3.8 we equate COHERENCE with the existence of a NOCB transition in U_{n-1} and use the remaining two underlying principles to define different sorts of ESTABLISHMENT.[14] Appendix A shows the analysis of examples (3.1) and (3.2) in terms of basic transitions and ESTABLISHMENTs.

[14]The second utterance in a sequence is always taken to be some kind of ESTABLISHMENT as well under this configuration, although U_1 is not classified as a NOCB.

Basic PT transition	Basic Centering transitions	N of violated principles
V0	CONTINUE	0
V1	EXP. CONTINUE, RETAIN, SMOOTH-SHIFT	1
V2	EXP. RETAIN, EXP. SMOOTH-SHIFT, ROUGH-SHIFT	2
V3	EXP. ROUGH-SHIFT	3

Table 3.9: Basic transitions in the Principles and Transitions (PT) formulation of Centering

3.5.2 A new set of transitions

As the previous section shows, the 3 underlying principles of CT result in at least 12 transitions; 12 transitions can be ranked in at least 12! ways (that is, more than 479,000,000 possibilities) as discussed in section 2.2.4.2 of chapter 2. Obviously, an exhaustive exploration of this vast number of possible rankings of basic transitions is impossible. Using principles instead of transitions, as argued by Beaver (2003) and Kibble (2001), does simplify the CT framework significantly, but fails to express possible combinations of the principles in the same utterance.

We believe that an interesting direction for future work within the transition-based versions of CT is to investigate possible mergings of transitions using Tables 3.7 and 3.8. This might express the combination of the underlying principles in the same utterance in a more complete way than in FC and standard CT, avoiding the complications that the set of 12 transitions creates. In the remainder of this section we sketch out one such possibility.

Our novel formulation of CT is called *Principles and Transitions* (PT). In this framework, all underlying principles are of equal importance and the basic Centering transitions in Table 3.7 are merged into basic PT transitions according to the number of principles they violate.

The set of basic PT transitions is shown in Table 3.9. The first column of the table shows the conventional name of the novel PT transition. The second column shows which basic Centering transitions from Table 3.7 are conflated into the corresponding basic PT transition. The third column reports the number of violated principles by the basic transitions. As we mentioned already, the basic Centering transitions that violate the same number of principles are merged into the same basic PT transition although they violate different principles.

Both FC and standard CT prioritise COHERENCE over SALIENCE which means that a RETAIN transition is preferred over a SMOOTH-SHIFT although both transitions violate only one principle in Table 3.7. In PT, RETAIN and SMOOTH-SHIFT belong to the same basic transition V1 because both violate only one underlying principle (albeit a different one).

Moreover, RETAIN and EXP. RETAIN in FC and the standard version of CT are considered to be the same transition although RETAIN violates only one underlying principle and EXP. RETAIN violates

two. In PT, RETAIN belongs to the basic PT transition V1 which is different from the PT transition V2, where EXP. RETAIN is classified, because RETAIN violates only one underlying principle and EXP. RETAIN violates two.

PT is a hybrid between the ideas in Kibble and Power (2000), FC and the standard formulation of CT. As in Kibble and Power (2000) we do not define priorities between the underlying principles of CT. Following FC and the standard formulation of CT, we translate the combination of principles into transitions. Our table of basic transitions is more exhaustive than the ones used so far in the CT literature. To simplify the framework, the basic transitions are conflated into basic PT transitions. Then, a ranking is imposed into the PT transitions according to the total number of principles they violate:

> **Rule 2 in PT**
>
> Transitions which violate less principles are preferred over transitions which violate more principles:
>
> V0>V1>V2>V3

There exist two ways of incorporating the ESTABLISHMENTs in Table 3.8 with basic PT transitions.[15] This gives rise to the extended PT transitions displayed in Table 3.10. In the upper section of the Table, labelled after the scoring function PT-EST-1, the various cases of ESTABLISHMENT behave like CONTINUEs or RETAINs. According to the second scoring function, namely PT-EST-2, an ESTABLISHMENT is considered to be a type of SHIFT.

The scoring functions in Table 3.10 can be used in the definitions of the PT-based metric, the evaluation method of which follows the preference order defined in PT's version of R2:[16]

- PT-based metric of entity coherence (M.PT):
  - **scoring function**:

    Sum(V0), Sum(V1), Sum(V2), Sum(V3)
  - **evaluation method**:

    Sum(V0)>Sum(V1)>Sum(V2)>Sum(V3)

The evaluation method of M.PT works in the same way as the evaluation method of M.BFP in section 3.4.1 with the exception that its input is not sums of standard CT transitions, but sums of PT transitions extended with ESTABLISHMENTs.

[15] Thus, the problem of whether a NOCB transition in U_{n-1} satisfies COHERENCE in U_n is now treated as an open question. This is motivated by the discussion of ESTABLISHMENT in Poesio et al. (2002).

[16] That is, we actually define two PT-based metrics, one for each way of incorporating ESTABLISHMENTS to the definition of basic PT transitions. The translation of examples (3.1) and (3.2) into extended PT transitions and their scores for each PT-based metric are shown in appendix A.

PT-EST-1		
PT transition	Basic Centering transitions & ESTABLISHMENTS	N of violated principles
V0	CONTINUE, ESTABLISHMENT	0
V1	EXP. CONTINUE, RETAIN, SMOOTH-SHIFT, EXP. ESTABLISHMENT, RETAIN-ESTABLISHMENT	1
V2	EXP. RETAIN, EXP. SMOOTH-SHIFT, ROUGH-SHIFT EXP. RETAIN-ESTABLISHMENT	2
V3	EXP. ROUGH-SHIFT	3

PT-EST-2		
PT transition	Basic Centering transitions & ESTABLISHMENTS	N of violated principles
V0	CONTINUE	0
V1	EXP. CONTINUE, RETAIN, SMOOTH-SHIFT, ESTABLISHMENT	1
V2	EXP. RETAIN, EXP. SMOOTH-SHIFT, ROUGH-SHIFT, EXP. ESTABLISHMENT, RETAIN-ESTABLISHMENT	2
V3	EXP. ROUGH-SHIFT, EXP. RETAIN-ESTABLISHMENT	3

Table 3.10: Extending the transitions in the Principles and Transitions (PT) formulation of Centering with ESTABLISHMENTS

3.6 The proliferation of CT-based metrics

This chapter addresses the question whether it is possible to use CT to define metrics that might prove useful for the purposes of text structuring in NLG. Our conclusion is summarised in the following statement:

> **Proliferation of CT-based Metrics**
>
> There exist **many ways** of using CT to define metrics of entity coherence for the purposes of text structuring.

As we have shown in the previous section and in section 3.2.2, CT is open-ended enough for one to propose new metrics which appear to be as plausible as some existing ones from a purely theoretical point of view. Hence, a general methodology for identifying which metrics represent the most promising candidates for text structuring is required, so that at least some of the possible metrics can be compared empirically.

Although ultimately all these metrics need to be subjected to empirical verification, the experiments reported in this thesis employ the eight metrics in Table 3.11 for reasons of practicality.[17]

[17]LS in Table 3.11 stands for preferring the structure with the lowest score in the evaluation method of the metric in question. OT signifies that the metric employs an OT ranking of the sums of violations that the scoring function returns.

Name	Scoring Function	Eval. method
M.NOCB	Sum(NOCB)	LS
M.CHEAP	Sum(CHEAP∗)	LS
M.MIL	Sum(NOCB)+Sum(ROUGH-SHIFT)	LS
M.SH	Sum(NOCB)+Sum(COH∗)	LS
M.KP	Sum(NOCB)+Sum(COH∗)+Sum(CHEAP∗)+Sum(SAL∗)	LS
M.SHOT1	Sum(NOCB), Sum(COH∗)	OT
M.POT1	Sum(NOCB), Sum(COH∗), Sum(CHEAP∗), Sum(SAL∗)	OT
M.BFP	Sum(CONTINUE), Sum(RETAIN), Sum(SMOOTH-SHIFT), Sum(ROUGH-SHIFT)	R2-OT

Table 3.11: The metrics used in our experiments

Priority was given to these metrics because their scoring functions are discussed in the existing CT literature more extensively than novel metrics such as the ones emerging from the alternative POT rankings (see section 3.2.2) and the extended PT transitions (see the previous section) which are introduced in this chapter for the very first time. Despite limiting our empirical investigation to eight metrics, this thesis considers more metrics of entity coherence than any previous work.

3.7 The second research question for text structuring

After having restricted the scope of the thesis to the metrics in Table 3.11, the question that we deal with in the next four chapters of the thesis is the following:

Q2: Which metrics of entity coherence constitute the most promising candidates for text structuring?

The next chapter employs a psycholinguistic study on text acceptability as our initial investigation of (Q2). Chapter 5 presents a novel corpus-based, search-oriented methodology as the main experimental framework under which we attempt to resolve the competition of the metrics. Although this methodology is general enough to be applied to any existing (or possible) CT-based metric, our experiments investigate the performance of the eight metrics specified above. The results of our initial corpus-based experiments are reported in chapter 6 and chapter 7. A modification of the metrics on the basis of an additional constraint on entity coherence is introduced in chapter 8. An alternative evaluation methodology which supplements the one in chapter 5 is discussed in chapter 9.

In a very general sense, trying to answer (Q2) is an effort similar to addressing the underspecification of CT discussed in section 2.3.1 of chapter 2. Hence, this thesis builds upon Poesio et al. (2002) in two ways. First, we identify and attempt to evaluate more scoring functions of entity coherence

The special OT-like evaluation method of M.BFP is named R2-OT.

than the ones employed by Poesio et al. (2002).[18] Second, while the approach in Poesio et al. (2002) does not consider the problem of choice, which is particularly important from an NLG viewpoint as we discussed extensively in the previous chapter, our methodology in chapter 5 is specifically devised for this purpose.

On the other hand, while Poesio et al. (2002) experiment with many different ways of specifying the CT parameters, in our work these parameters are **fixed** according to the needs of text structuring under the assumptions stated in chapter 2, as we discuss in chapter 6 and chapter 7 in more detail. Arguably, each of the employed metrics can be tested against different specifications for CT parameters such as *utterance*, *realisation*, etc. as well.[19]

[18] As we mentioned in section 2.3.2 of chapter 2 and in the beginning of this chapter, Poesio et al. (2002) employ the scoring functions of M.NOCB and M.BFP (as well as M.CHEAP and a version of S.KP) in their experiments. We believe that the other metrics in this chapter can also be seen as evaluating "claims" of CT, like the scoring functions employed by Poesio et al. (2002).

[19] See section 6.6 of chapter 6 for a more specific suggestion.

Chapter 4

A preliminary study on text acceptability

In this chapter, we initiate our empirical work with a psycholinguistic study that aims at testing the different predictions of three metrics of entity coherence using acceptability judgements. We report the problems we encountered in this study, comment on the cost of human-based evaluation and conclude that an alternative methodology is desirable for deciding which metrics represent good candidates for the purposes of text structuring, the results of which can be supplemented by subsequent human-based evaluation on a smaller scale.

4.1 Introduction

The previous chapter motivates our main research question, namely to identify the CT-based metrics of entity coherence which constitute the most promising solutions for text structuring among eight preselected candidates. This chapter initiates our empirical work with a psycholinguistic study that investigates the different predictions of three of these metrics using human acceptability judgements. Like the rest of the experiments reported in this thesis, this study aims to identify promising metrics before their actual implementation in a text structuring component. The following sections describe the experimental design, the predictions and the results of this study.

4.2 Magnitude estimation

The experimental paradigm employed in this study is *Magnitude Estimation* (ME), a technique originating from psychophysics. In ME, the participants estimate the magnitude of physical stimuli by assigning numerical values proportional to a reference stimulus which is called the *modulus*.

Bard et al. (1996) and Cowart (1997) show that linguistic judgements can be elicited in the same way as judgements of sensory stimuli using ME. Following the standard ME procedure, the partic-

ipants are first exposed to a linguistic stimulus that serves as the modulus to which they assign an arbitrary number. Then, they are asked to express the acceptability that they perceive by assigning numbers to a series of linguistic stimuli. Each stimulus is rated in proportion to the modulus, that is, if the participant is presented with a stimulus that is perceived to be three times more acceptable than the modulus she assigns three times the modulus number to it, etc.

Because ME provides gradient data, it is very appropriate for judgements which fall within a continuum. These judgements are difficult to express in the informal scales usually employed in traditional linguistic studies which appear to compress a very wide range of acceptability levels into just a few imprecise categories of grammaticality.

Numerous recent studies such as Keller and Alexopoulou (2001) and Alexopoulou and Keller (2003) show that ME provides fine-grained measurements of linguistic acceptability which are robust enough to yield statistically significant results, while being replicable both within and across speakers. In these experiments, ME is used to judge the acceptability of single sentences or short discourses consisting of up to two sentences. The judgements shed light on the effects of various syntactic phenomena which cannot be modelled using a binary notion of grammaticality.

Further to this, Pearson (2000) shows that ME can be used to estimate the acceptability of longer texts that differ in their coherence. The text structures in Pearson (2000) were generated automatically following a stochastic approach similar to the one in Mellish et al. (1998a). Intuitive measures for “very high” and “very low” entity coherence, remotely related to CT, were used in the scoring function, alongside rhetorical coherence and coherence based on fact types. The structures were realised by hand and the participants were asked to judge the acceptability of the resulting texts using a ME technique.

Our experiment is similar to the one in Pearson (2000) since we are interested in measuring the acceptability of texts consisting of more than two sentences. In our experiment, entity coherence is modelled in terms of three of the metrics that were presented in the previous chapter. The metrics make different predictions with respect to the relative acceptability of the experimental items. The experiment investigates which predictions are best validated by human judgements.

Although ME appears to be a very appropriate experimental paradigm for our purposes, designing a single experiment which accounts for the different predictions of all eight metrics would be extremely complicated if not practically impossible. For this reason this study employs only three of them, namely M.NOCB, M.SHOT1 and M.SH. If the results are encouraging, then a series of similar experiments can be motivated in order to compare the metrics with each other in a systematic way.

4.3 Experimental conditions

Our experimental design uses the six textual variations in Figure 4.1 as the experimental conditions. In utterance (a) of condition (4.1) two entities A and B (in this case the referents of the phrases "this exhibit" and "an amphora") are introduced in the discourse. Utterance (b) provides more information about entity B, while utterance (c) also refers to B and evokes, among others, a third entity C (in our case the referent of the phrase "type A") which is subsequently mentioned together with entity A in utterance (d). The discourse concludes with additional information about A in (e). The textual variations in conditions (4.2) to (4.6) represent different orderings of utterances (b) to (e).

The table at the bottom end of the Figure shows the utterances with NOCB transitions and the violations of COHERENCE (COH$*$) for each condition (COND). The experimental design is based on the fact that M.NOCB, M.SHOT1 and M.SH make different predictions about the relative acceptability of the texts according to the number of NOCBs and COHERENCE violations in each condition.

4.3.1 Predictions

The experimental design is similar to the way Keller and Alexopoulou (2001) investigate whether violations of syntactic constraints affect the relative acceptability of sentences belonging to different conditions. In our set-up, one of the conditions is predicted to be more acceptable than the others according to the way that the violations of entity coherence are combined to a metric. Since the conditions in Figure 4.1 are ranked by the metrics in different ways, the aim of the study is to find out which ranking is confirmed by the experimental data.

Following Keller and Alexopoulou (2001), we make the distinction between *Strict Optimality* and *Relative Suboptimality* with respect to the way that this ranking is defined. According to the Strict Optimality Hypothesis (SOH), one condition will be preferred whereas the remaining conditions will be equally suboptimal. The predictions of each metric under SOH are presented in Table 4.1.

As Table 4.1 shows, both M.NOCB and M.SHOT1 predict that the optimal condition will be (4.1) because it has fewer NOCB transitions that the other conditions. M.SH uses the sum of NOCB transitions and violations of COHERENCE as the measure of incoherence in a structure. According to M.SH the most acceptable condition will be (4.2), because this is the condition that returns the lowest sum. As we have already mentioned, under SOH acceptability is viewed as a binary notion: one optimal candidate is selected whilst all remaining conditions are equally suboptimal.

Table 4.2 shows the predictions of the metrics under the Relative Suboptimality Hypothesis (RSH). Under RSH, the metrics make the same predictions as under SOH with respect to the condition that is ranked best. Further to this, under RSH the metrics make additional predictions about the relative acceptability of the suboptimal conditions.

(4.1) (a) This exhibit is an amphora. (b) Amphoras have an ovoid body and two looped handles, reaching from the shoulders up. (c) They were produced in two major variations: type A and the type with a neck. (d) This exhibit is a type A amphora. (e) It comes from the archaic period.

(4.2) (a) This exhibit is an amphora. (b) Amphoras have an ovoid body and two looped handles, reaching from the shoulders up. (c) They were produced in two major variations: type A and the type with a neck. (e) This exhibit comes from the archaic period. (d) It is a type A amphora.

(4.3) (a) This exhibit is an amphora. (e) It comes from the archaic period. (b) Amphoras have an ovoid body and two looped handles, reaching from the shoulders up. (c) They were produced in two major variations: type A and the type with a neck. (d) This exhibit is a type A amphora.

(4.4) (a) This exhibit is an amphora. (c) Amphoras were produced in two major variations: type A and the type with a neck. (d) This exhibit is a type A amphora. (e) It comes from the archaic period. (b) Amphoras have an ovoid body and two looped handles, reaching from the shoulders up.

(4.5) (a) This exhibit is an amphora. (e) It comes from the archaic period. (c) Amphoras were produced in two major variations: type A and the type with a neck. (b) They have an ovoid body and two looped handles, reaching from the shoulders up. (d) This exhibit is a type A amphora.

(4.6) (a) This exhibit is an amphora. (c) Amphoras were produced in two major variations: type A and the type with a neck. (d) This exhibit is a type A amphora. (b) Amphoras have an ovoid body and two looped handles, reaching from the shoulders up. (e) This exhibit comes from the archaic period.

COND	NOCB	COH∗
(4.1)	-	d, e
(4.2)	e	-
(4.3)	b	d
(4.4)	b	d, e
(4.5)	c, d	-
(4.6)	b, e	d

Figure 4.1: Experimental conditions and their NOCB transitions and COHERENCE violations

	M.NOCB, **M.SHOT1**		**M.SH**	
Rank	COND	NOCB	COND	NOCB+COH∗
1	(4.1)	0	(4.2)	1
2	(4.2), (4.3), (4.4), (4.5), (4.6)	> 0	(4.1), (4.3), (4.4), (4.5), (4.6)	> 1

Table 4.1: Rankings of conditions under the Strict Optimality Hypothesis for each metric

	M.NOCB		M.SH	
Rank	COND	NOCB	COND	NOCB+COH∗
1	(4.1)	0	(4.2)	1
2	(4.2), (4.3), (4.4)	1	(4.1), (4.3), (4.5)	2
3	(4.5), (4.6)	2	(4.4), (4.6)	3

	M.SHOT1		
Rank	COND	NOCB	COH∗
1	(4.1)	0	2
2	(4.2)	1	0
3	(4.3)	1	1
4	(4.4)	1	2
5	(4.5)	2	0
6	(4.6)	2	1

Table 4.2: Rankings of conditions under the Relative Suboptimality Hypothesis for each metric

More specifically, M.NOCB under RSH predicts that conditions (4.2), (4.3) and (4.4) will be equally acceptable when compared to each other, albeit more acceptable than (4.5) and (4.6). This is because all conditions in the second rank have fewer NOCB transitions than the conditions that are ranked third. Conversely, M.SH predicts that the group of conditions that will be ranked second will consist of (4.1), (4.3) and (4.5), whereas the conditions with the highest sum of NOCBs and violations of COHERENCE will be ranked last.

Finally, the violations of COHERENCE play a crucial role in distinguishing between M.NOCB and M.SHOT1 under RSH. Conditions such as (4.2), (4.3) and (4.4) have the same number of NOCBs and are considered equivalent by M.NOCB. Since these conditions differ on the violations of COHERENCE, M.SHOT ranks them relatively to each other in accordance with these violations. The same holds for conditions (4.5) and (4.6).

4.3.2 Testing for significance

Both under SOH and RSH we expect either (4.1) or (4.2) to be the most acceptable condition on average. If this is the case, a planned comparison between the most acceptable condition and the one with the second highest average will be employed. If the test shows that the difference is significant, then the experiment will provide evidence in favour of the metric(s) which makes the corresponding prediction under SOH.

More specifically, if (4.1) is most acceptable and significantly different than the condition with the second average acceptability, then the experiment provides evidence in favour of M.NOCB and M.SHOT1 under SOH. If (4.2) is most acceptable and significantly different than the second condition on average, then we have evidence in favour of M.SH under SOH.

Investigating RSH with standard tests of significance might require a large number of additional pairwise comparisons, most of which cannot be planned in advance. For this reason, the predictions under RSH are tested using the methodology of Keller and Alexopoulou (2001).

First, the average judgements are converted to ranks, ignoring differences that are smaller that one standard error. More specifically, Keller and Alexopoulou (2001, p.333) adopt the following criterion: two means m_1 and m_2, for which $m_1 > m_2$ holds, are considered to be of different rank if **both** m_2 is lower than $m_1 - se_1$ **and** m_1 is higher than $m_2 + se_2$, where se_1 and se_2 are the standard errors associated with m_1 and m_2 respectively. If either of these tests does not hold, m_2 is assigned the same rank as m_1 in the acceptability order. Then, Keller and Alexopoulou (2001) compare the resulting ranking with the grammaticality order predicted under RSH using standard correlation statistics. The degree of correlation between the theoretical and the experimental orders indicates how well the RSH model fits the experimental data.

4.4 Method

4.4.1 Participants

63 native speakers of English participated in the experiment. They were recruited through the Language Experiments Portal[1] and by postings to mailing lists of various academic institutions within and outside the University of Edinburgh. Participation was voluntary and unpaid.

Six participants were automatically excluded by the experimental software used in this study either because they did not complete the experiment, or because they did not provide a valid email address (see section 4.4.3). Additionally, the data of 11 out of the 57 remaining participants contained many outliers and extreme values (see section 4.5.1 for more details). This left judgements from 46 participants for analysis. Of these, 26 were female and 20 male; 6 were left-handed and 40 right-handed. The age of the participants ranged from 18 to 49 years (mean age: 25.4 years).

4.4.2 Materials

As we shall explain in the next section in more detail, the experiment was conducted in three phases: a training phase, a practice phase and the actual experiment. A different set of materials was prepared

[1] http://www.language-experiments.org/

for each experimental phase.

The role of the training materials is to familiarise subjects with the magnitude estimation task by eliciting judgements on line length. A standard set of four horizontal lines used previously in other ME experiments were used for this task.

The linguistic materials for the next two phases were prepared by the author using information derived from the Getty museum webpage and the sample descriptions of archaeological artefacts written for the purposes of the MPIRO project.[2]

The practice phase familiarises the participants with applying magnitude estimation to linguistic stimuli and is not meant to be taken into account in the analysis. The materials for this phase consisted of three texts and a modulus that were distinct from, but representative of, the materials in the main experiment.

Six variations of 12 texts corresponding to the six experimental conditions, i.e. 72 lexicalisations in total, 12 fillers and a second modulus were prepared for the main experiment. Pronominalisation was controlled according to the algorithm in O'Donnell et al. (1998).[3] The length of the 12 lexicalisations in condition (4.1) ranged from 47 to 70 words (mean length: 58 words). In all cases, condition (4.1) consisted of five utterances evoking and subsequently referring to three entities A, B and C according to the pattern presented in section 4.3. Appropriate reorderings of the utterances in condition (4.1) gave rise to the lexicalisations in the other five conditions. Two native speakers of English were asked to check all lexicalisations for disfluencies and language errors.[4]

The fillers are items which do not belong to the experimental conditions and are not taken into account in the analysis. They are presented together with the experimental items in random order and aim at preventing any bias in judgements caused by the participants being able to recognise the experimental conditions.

The standard ME methodology requires the modulus to appear somewhere "in the middle of the acceptability range". A lexicalisation belonging to condition (4.4) was chosen as the modulus in the main experimental phase because we expected that the mean acceptability of (4.4) will appear somewhere between the highest and the lowest average.[5]

After the lexicalisations had been corrected according to the comments of the native speakers, the texts were grouped randomly in six pairs. Then, a Latin square design was used to form 6 blocks

[2] All materials appear online at http://www.iccs.informatics.ed.ac.uk/~nikiforo/thesis-online/ME/ME.items.html

[3] As we mentioned in section 2.4.2 of chapter 2, we interpret this algorithm as pronominalising the CB(U_n) if it is the same as the CP(U_{n-1}).

[4] Many thanks to Amy Isard and Tracy Markusic for undertaking this task.

[5] Looking at Table 4.2 with hindsight leads us to the conclusion that a better choice for the modulus might have been a lexicalisation belonging to condition (4.3), since (4.3) is always predicted to be "in the middle" of the range by all metrics under RSH. By contrast, M.SH under RSH ranks (4.4) and (4.6) as last. Note that the predictions under SOH shown in Table 4.1 are not helpful for identifying which condition is expected to appear in the middle of the range. See section 4.5.4 for more discussion on the difficulty of choosing an appropriate modulus.

(test sets) in order to make sure that each participant is not going to be presented with more than one lexicalisation for each text. Each block consisted of 2 lexicalisations per condition, that is, 12 experimental items, and the 12 fillers.

4.4.3 Procedure

The experiment was conducted remotely over the Internet using WebExp, an interactive software package for administering web-based psychological experiments (Keller et al. 1998).[6] Although a web-based study is likely to attract rapidly a large number of participants, the experimenter exercises less control over the conditions under which the participants undertake the experiment compared to conventional laboratory experiments. WebExp is designed to ensure the authenticity of web-based data by automatically eliminating participants whose identity cannot be verified, who participate more than once, or who respond either too quickly or too slowly. As Keller and Alexopoulou (2001) show, web-based data obtained via WebExp can be as reliable as laboratory data obtained under a similar experimental design as the experiment conducted through WebExp.

Each participant accessed the experiment using her browser. First, the participant had to read a set of instructions.[7] The instructions started by explaining the concept of numerical magnitude estimation of line length. An example reference line, a longer and a shorter line as well as corresponding numerical estimates were provided to illustrate the concept of proportionality.

Then, the participant was told that the acceptability of texts can be judged in the same way as line length. She was instructed that during the experiment she will be asked to give numbers to texts according to how well each text organises the information it consists of. An example of a reference text and a text to be judged for its acceptability in proportion to the reference were provided together with examples of numerical estimates.

At this point it was stressed that there are no "correct" answers and that the participant should base her judgements on her first impression after having read the text carefully and having compared it with the modulus. She was free to use any number she liked, including decimals, except for zero and negative numbers. She was urged to use a wide range of numbers and to distinguish as many degrees of acceptability as possible.

After reading the instructions, the participant clicked on the "Start" button and was presented with a short demographic questionnaire which included name, email address, age, sex, handedness, occupation or subject of study, and first language region. After filling in the questionnaire, she had to go through the training phase where she was first presented with a modulus line which had to be assigned

[6] Many thanks to Frank Keller for practical advice on running WebExp.

[7] The full text of the instructions appears in appendix B and is an adaptation of the instructions normally used in WebExp-based experiments to our purposes.

with an arbitrary number. The modulus remained on the upper half of her screen while she was subsequently presented with another line centered in her browser that had to be judged in comparison to the modulus. After the judgement was given, the current line disappeared from the screen and a new item to be judged was presented. In total, the participant had to provide judgements for all four items in the training set presented one after the other in random order, with a new randomisation being generated each time.

After the training phase with the lines, the participant entered the practice phase in which she was presented with the modulus text and the three texts to be judged in comparison to it. The presentation and response procedure was the same as in the training phase, with texts being displayed instead of lines. As in the training phase, the participant first judged the modulus item which remained on the screen to facilitate the comparison with the next randomly presented text.

After the training phase and the practice phase, the participant entered the actual experiment. At this stage, she was randomly assigned to one of the six blocks of the experimental design. As in the practice phase, the participant first judged the modulus text and then saw the 24 test items of her block in random order, with a new randomisation generated each time. As we explained in the previous section, the modulus and the experimental items in the main phase did not overlap with the ones used in the practice phase.

4.5 Results

The data from the main phase of the experiment were normalised by dividing each numerical judgement by the modulus value that the participant assigned to the reference text. This operation creates a common scale for all participants. Then, the data were transformed to their decadic logarithm, which is a standard practice for ME data. All analyses were conducted on the normalised, log-transformed judgements.

4.5.1 Outliers

The initial exploration of the 684 judgements from the 57 authenticated participants revealed a large number of outliers. This points to one of the problems related to web-based studies, namely the limited control over the experimental situation, which is, however, counterbalanced by the increased number of participants compared to the number of participants recruited in standard laboratory experiments. The upper part of Figure 4.2 shows the boxplots of the six conditions using the complete set of data including the outliers and extreme values as computed by the SPSS statistical software.[8]

[8]The lower boundary of each box in the boxplot represents the 25th percentile and the upper boundary represents the 75th percentile so that fifty percent of the cases have values within the box. The length of the box corresponds to the interquartile range, that is, the difference between the 75th and the 25th percentiles. The horizontal line inside each box

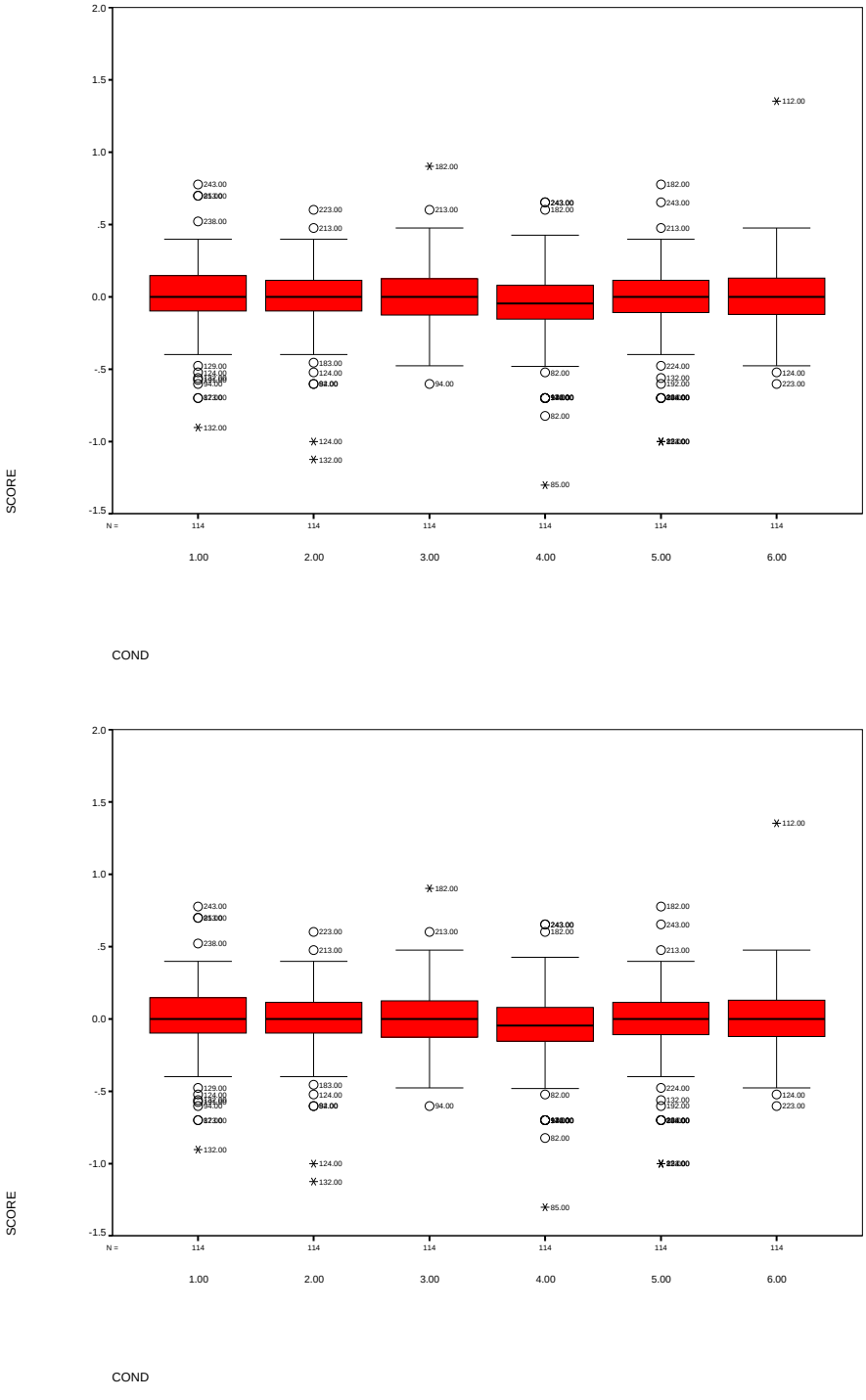

Figure 4.2: Boxplots of conditions with and without outliers

COND	mean	SE
(4.1)	356	160
(4.5)	183	163
(4.3)	51	169
(4.6)	21	194
(4.2)	-8	176
(4.4)	-249	193

Table 4.3: Mean acceptability and standard error (SE) for the experimental conditions

The acceptability judgements of 11 participants were not taken into account because their data mostly consisted of outliers and extreme values. This resulted in 552 datapoints from 46 participants. However, 18 of these datapoints were outliers as well. Because each condition was lexicalised twice for each participant, it was possible to disregard these values and replace them with the value from the other lexicalisation in the same condition.[9] The boxplot of the six conditions using the 534 valid datapoints, cleared from outliers, is displayed in the lower part of Figure 4.2.

4.5.2 Results for strict optimality

Table 4.3 reports the average acceptability and the standard error (SE) for each condition.[10] As the table shows, (4.1) is indeed the most acceptable condition on average. The condition with the second highest mean is (4.5). Condition (4.2) is the second least acceptable, whilst (4.4) is the condition with the lowest mean.

However, the omnibus one-way ANOVA showed no main effect for COND (by subjects: F(5,225)=1.634, p=0.152; by items: F(5,55)=2.250, p=0.062). Hence, this analysis fails to find any significant difference between the six conditions and cannot justify rejecting the null hypothesis.[11]

---

represents the median of the corresponding condition. The boxplot includes two categories of cases with outlying values. Cases that are more than 3 box-lengths from the upper or lower edge of box are called *extreme values* and are denoted with an asterisk. Cases with values between 1.5 and 3 box-lengths from the upper or lower edge of the box are called *outliers* and are designated with a circle. The largest and smallest observed values which are not outliers are also shown with lines drawn from the ends of the box to these values (whiskers). Outliers and extreme values are labelled according to the conventional number assigned by WebExp to the participant who provided the values.

[9] We acknowledge, however, that excluding more than 1/5 of the datapoints from the final analysis compromises the generality of any significant results that might arise from this study.

[10] The numbers in Table 4.3 are in the log-transformed scale, multiplied by 10,000. That is, the mean value 356 corresponds to the value 0.0356 that results from the normalisation procedure. The minus sign indicates that on average the condition was found to be less acceptable than the modulus.

[11] Note that including the outliers in the analysis retains (4.1) as the condition with the highest average and (4.4) as the condition with the lowest average. However, the omnibus ANOVA is again not significant (by subjects: F(5,280)=1.616, p=0.156; by items F(5,55)=1.630, p=0.167).

merged COND	COND	mean	SE
(4.1)		356	160
A3	(4.5)+(4.6)	102	136
A2	(4.2)+(4.3)+(4.4)	-68	118

Table 4.4: Mean acceptability and standard error (SE) for merged conditions

Faced with this inconvenient result, our next step was to try different ways of grouping the conditions together. The most obvious way of doing so was to combine the values from the conditions of the original experimental design into three new conditions according to the predictions of M.NOCB. Table 4.4 shows the means and standard errors of acceptability of the new conditions that this merging results in.

This time the omnibus one-way ANOVA was significant at the 0.05 level (by subjects: F(2,90)=3.480, p=0.035; by items: F(2,22)=3.884, p=0.036). A set of pairwise comparisons were then employed to see whether the predictions of M.NOCB are indeed verified. Note that according to the Bonferroni method the threshold for significance for three pairwise comparisons is reduced from 0.05 to 0.05/3=0.017.

Crucially for M.NOCB and M.SHOT1 under SOH, the planned comparison between (4.1) and A3, that is, the merged condition with the second highest average which contains the combined scores of (4.5) and (4.6), failed to reach significance (by subjects: F(1,45)=1.859, p=0.179; F(1,11)=2.180, p=0.168). The difference between A3 and A2 was also not significant (by subjects: F(1,45)=1.648, p=0.206; by items: F(1,11)=2.374, p=0.152). The only significant difference, in the by subjects analysis, is the one between (4.1) and A2, which is not entirely surprising since A2 contains (4.2) and (4.4), the two least acceptable original conditions (by subjects: F(1,45)=6.838, p=0.012; by items: F(1,11)=6.705, p=0.025, which is higher than the Bonferroni threshold of p=0.017).

The latter result might have been useful, were A2 the condition with the second highest average. However, since A2 is the least acceptable merged condition, this difference does not seem to support any of our initial predictions.

In summary, although (4.1) is the condition with the highest average in all analyses, it does not appear to be significantly different from the condition with the second highest average, contrary to what M.NOCB and M.SHOT1 predict under SOH.

4.5.3 Computing the acceptability order

Computing the acceptability order according to the methodology in Keller and Alexopoulou (2001) was seen originally as a way to investigate RSH without having to resort to a large number of pairwise

COND	mean	SE	mean+SE	mean−SE	$m_2 < m_1 - se_1$	$m_1 > m_2 + se_2$
(4.1)	356	160		196	n.a.	
(4.5)	183	163	346	20	**yes**	**yes**
(4.3)	51	169	220	-118	no	no
(4.6)	21	194	215	-173	no	no
(4.2)	-8	176	168	-184	no	no
(4.4)	-249	193	-56		**yes**	**yes**

Table 4.5: Determining the acceptability order of the experimental conditions

comparisons.

Obviously, RSH can only be examined provided that there is some evidence that SOH (as modelled by at least one of the metrics) holds. This did not emerge from our analysis. However, instead of abandoning the idea of computing the acceptability order altogether, we did follow the methodology of Keller and Alexopoulou (2001) in order to see whether this could reveal any tendencies that might be interesting but not significant enough to survive standard statistical tests.

The second and third columns of Table 4.5 repeat the means and standard errors of the six experimental conditions as in Table 4.3. The next two columns show the two thresholds that Keller and Alexopoulou (2001) use for the conversion of average values to ranks. As we mentioned in section 4.3.2, the first threshold for two subsequent means m_1 and m_2 results from adding m_2 with se_2. The second threshold is the result from the subtraction of se_1 from m_1. The last two columns of Table 4.5 show whether the tests $m_2 < m_1 - se_1$ and $m_1 > m_2 + se_2$ hold. The two cases where the tests do indeed hold are highlighted in bold font.

According to Table 4.5, the acceptability order of the conditions in our experiment is:[12]

- Acceptability order of ME conditions:

[12]The sign $>>$ signifies difference in rank, i.e. (4.1) is more highly ranked than (4.5) which is of equivalent rank as (4.3), (4.6) and (4.2), etc.

(4.1) >> (4.5), (4.3), (4.6), (4.2) >> (4.4)

Although this order does show a tendency for (4.1) to be preferred, the differences between the mean of (4.1) and the boundary set by the mean and the SE of (4.5) and vice versa are very small. In any case, it is difficult to take the acceptability order as computed here as a clear trend in favour of M.NOCB and M.SHOT1, because this order shows a strong tendency for (4.4) to be dispreferred, which is not predicted by any of our experimental hypotheses.[13]

4.5.4 Discussion

In conclusion, an interesting, yet not necessarily generalisable, result of this experiment is a dispreference for condition (4.4) which was not predicted by any of the metrics. A less clear trend for (4.1) to be preferred was also observed. Despite this trend, it is difficult for us to claim that the experiment answers the research questions that motivated it.

We believe that disprefering (4.4) does not have to do with entity coherence, at least in the way that this notion is modelled by the metrics employed in this study. Note that it is not possible to explain the results using any of the other metrics from the previous chapter either, because the way that these metrics differ with respect to the conditions in Figure 4.1 is not as systematic as for the three metrics employed in this experiment.

Trying to determine why (4.4) is dispreferred, we considered additional features that might affect the acceptability of this condition such as "mentioning the shape of the amphora" too late as well as similar characteristics of the other lexicalisations of utterance (b) in Figure 4.1. However, it appears that condition (4.6) which resembles (4.4) in this respect is not penalised as severely as (4.4).

Another implication arises from the fact that the participants identified the lexicalisations in condition (4.4) as less acceptable than the experimental modulus, although the text used as the modulus was thought to correspond to condition (4.4) as well according to our design. Outliers excluded, our dataset contains 90 judgements for the 12 lexicalisations in condition (4.4) most of which are much lower than the score assigned to the modulus, thus giving rise to the negative mean in Table 4.3. Figure 4.3 shows the modulus in comparison with the lexicalisation of condition (4.4) from Figure 4.1.

Subsequent informal interviews with some participants showed that they considered the modulus as "more interesting" than the experimental item in Figure 4.3. Although the participants were specifically instructed to give numbers to each text that "reflect your judgement on the way that the text organises the information it consists of" and not according to some other property, practically

[13] M.SH under RSH does predict that (4.4), together with (4.6), will be the least acceptable conditions. However, it also predicts that (4.2) will be significantly more acceptable than all other conditions including (4.4) which does not hold as we already showed.

Modulus:

This statue was made by Polyclitus. Polyclitus is one of the most important ancient sculptors, together with Phidias and Praxiteles. This statue is wider and better built than the ones of Phidias. It comes from around 440 BC. Polyclitus expresses the spirituality and the anthropocentric attitude of the classical world.

Lexicalisation in (4.4):

This exhibit is an amphora. Amphoras were produced in two major variations: type A and the type with a neck. This exhibit is a type A amphora. It comes from the archaic period. Amphoras have an ovoid body and two looped handles, reaching from the shoulders up.

Figure 4.3: Experimental modulus and lexicalisation in condition (4.4)

it proves that it is very difficult to dissociate unpredictable aspects such as "interestingness" from acceptability judgements.

Although controlling for confounds like this is very hard in any experimental design, this finding brings forward an important point which is specific to ME. Choosing the wrong modulus, as appears to have happened in our experimental set-up, might turn out to be particularly problematic. However, as we mentioned earlier, it is difficult to tell in advance whether an item is an appropriate modulus or not.

4.6 The cost of human-based evaluation

As we mentioned in the beginning of this chapter, because it is practically impossible to come up with an experimental design which accounts for the predictions of all eight candidate metrics at the same time, a systematic way to compare the metrics using psycholinguistic methods before the actual implementation of the text structuring component is to design a series of supplementary perceptual experiments. Designing and preparing the materials for this set of experiments requires a very substantial amount of effort, even if one ignores problems like the ones discussed in this chapter which might render the whole attempt almost fruitless at the end.

On the other hand, the usefulness of psycholinguistic methods is undeniable, thus it is desirable to combine them with the demands of NLG. This general approach has been already followed by Cheng (2002), Rambow et al. (2001) and Bangalore et al. (2000), among others, who validate their quantitative approaches with additional evaluation based on human judgements of quality and understandability. However, to the best of our understanding, Rambow et al. (2001) and Bangalore et al. (2000) evaluate sentence planning choices rather than longer texts, and we feel justified to believe that eliciting human judgements for this task might be easier than evaluating the output of a text structuring

system.

On the other hand, while the study in Cheng (2002) takes place on a very small scale (only 10 participants), to the best of our knowledge, the only recent large-scale attempt to evaluate the output of a text structuring component using acceptability judgements, is represented by Pearson (2000), which in turn shows that EM experiments on text acceptability **are** possible. Note, however, that the items that Pearson (2000) used were assigned with extreme values for "high" and "low" coherence by the scoring function of its genetic algorithm. Thus, eliciting judgements that differ significantly for these items might also have been easier than for items much closer in the acceptability range.

Instead of following the experimental design sketched out in this chapter, one could implement each different metric of entity coherence for the purposes of text structuring and then use their outputs in a series of perceptual experiments now aiming at deciding which metric generates the best structures. However, this kind of evaluation is especially demanding and time consuming as well: For instance, Lester and Porter (1997) report a seven year effort to evaluate a system that generates explanations from large and semantically rich databases. Consequently, it would be desirable to minimise the effort by finding another way to decide in advance which metrics represent good candidates for the purposes of text structuring and restrict the implementation and the user-based evaluation only to (some of) these candidates.

The next chapter presents a corpus-based, search-oriented methodology which adheres to the aim of finding out which metrics of entity coherence represent more suitable candidates for text structuring prior to the actual generation of a text structure. It is then these metrics that are best to implement and evaluate using human judgements. Because, as this chapter has shown, addressing this point through a large-scale psycholinguistic study would be especially demanding, we present the corpus-based evaluation as a sensible shortcut. In this sense, the corpus-based evaluation can be seen as a test-bench that provides a subsequent human-based evaluation with fewer hypotheses to test.

Chapter 5

Corpus-based evaluation: Methodology

In this chapter, we discuss the basic methodological issues of our corpus-based, search-oriented evaluation of different metrics of entity coherence and describe the main features of SEEC, the system that was implemented to carry out our experiments. We show how each corpus instance is used as the *Basis for Comparison* in a search-oriented evaluation which calculates the *classification rate* of each metric and compares their performance. We conclude the chapter with a discussion of our solution to the factorial complexity of the operation that this search entails.

5.1 Motivation

As we discussed in chapter 2, most of the existing corpus-based studies of CT treat each NOCB transition as an absolute measure of incoherence and evaluate the preferences of R2 using the relative frequency of the standard CT transitions. However, this method of evaluation is insufficient, at least from an NLG point of view, because it does not address the problem of choice.

Moreover, in chapter 3, we showed that the fundamental notions of CT can be used to define a very wide range of metrics of entity coherence. Therefore, an adequate corpus-based evaluation of CT should not be restricted simply to the preferences of C1 or R2 but should try to experiment with many theoretically motivated metrics. In the previous chapter, we discussed the problems that emerged from a study on human perception and concluded that an alternative methodology is desirable for deciding which metrics represent good candidates for the purposes of text structuring.

In this chapter, we discuss how our corpus-based, search-oriented evaluation operates on a representation called the *Basis for Comparison* in order to compute a *classification rate* for each metric. This rate is used to compare the metrics to each other. Compared to previous ways of evaluating versions of CT using corpora, our approach is the first one that considers the problem of choice. Note that, as we mentioned in section 3.7 of chapter 3, the methodology described in this chapter is general

enough to apply to any metric that might emerge from CT's proliferation of possible metrics. However, the experimental studies reported in subsequent chapters of the thesis employ only eight of these metrics for practical reasons.

5.2 Issues in corpus-based evaluation

According to Reiter and Dale (2000, p.80), both the genre of interest and the particular domain of application impose constraints on the kind of text structures that it is appropriate to generate. Hence, GNOME-LAB and MPIRO-PROP, the corpora that we use in the studies reported in subsequent chapters of the thesis, are chosen as representatives of the text genre and a particular application domain respectively.

GNOME-LAB, which is used in our main corpus study in chapter 6, is a subset of the collection of texts in the museum section of the GNOME corpus which is made up of texts published on official webpages and books about museum collections (Poesio 2000). This subset consists of all texts in the museum section that are recognised as museum labels.[1]

Subsequent evaluation within the context of a specific application makes use of MPIRO-PROP which consists of sets of coherent sequences of propositions instead of texts. The propositions were derived from the database of the MPIRO system and manually assigned an order to reflect what a domain expert considered to be the most natural ordering of the corresponding sentences in the texts to be generated (Dimitromanolaki and Androutsopoulos 2003).

5.2.1 The assumption of the gold standard

Reiter and Sripada (2002) notice an increasing interest in using corpora of human-authored texts in NLG, especially for knowledge acquisition (e.g. Barzilay and McKeown 2001; Duboue and McKeown 2001; Hardt and Rambow 2001). Most of the papers reviewed by Reiter and Sripada (2002) assume that an approximate evaluation for a system is to compare its output to human texts from the corpus. This is in turn based on the underlying assumption that NLG systems should attempt to generate corpus texts, in other words that corpus texts are a *gold standard* for NLG.

A clear manifestation of this approach is described by Bangalore et al. (2000) who define several intrinsic metrics for quantitative evaluation of their single-sentence realisation module using different ways of calculating the string edit distance between the surface output of their system and the reference corpus string or substrings derived by the dependency tree of the reference string. Similarly, Cheng and Mellish (2000a) and Cheng (2002, Chapter 8) use corpus texts to evaluate the output of a genetic

[1] The museum section of the GNOME corpus includes texts from other genres which have not been taken into account in our study (see section 6.2.1 of chapter 6 for more details).

algorithm which models the interaction between aggregation and text structuring in the ILEX domain. Although the aim of our corpus-based study is not to evaluate the output of a system, but to identify promising metrics before the actual implementation of the text structuring component takes place, the assumption of the gold standard which underlies existing corpus-based evaluations is shared by our approach to a great extent.

Reiter and Sripada (2002) draw from their own experience in using corpora as gold standards to question the underlying assumption that an NLG system should produce texts similar to the corpus texts for two main reasons:

1. Human authors make mistakes, especially when they are writing hastily. NLG systems should not imitate these mistakes.

2. There are substantial variations between individual writers which reduces the effectiveness of corpus-based learning.

Note that the points brought forward by Reiter and Sripada (2002) do not have to do with the coherence of the texts per se, but raise general issues on the quality and use of corpora. In general, we agree with Reiter and Sripada (2002) that for the purposes of our evaluation a smaller corpus of high-quality texts would be more useful than a larger corpus of problematic texts. Further to this, we have reasons to believe that the quality of our corpora has not been severely compromised by the circumstances of authoring. This is because our corpora are either publicly available texts targeting a wide audience (GNOME-LAB) or carefully considered text structures resulting from close consultation with a domain expert (MPIRO-PROP). Therefore, the writers are expected to have paid enough attention in order to avoid sloppiness during authoring, although it is impossible to ensure that the corpora are completely flawless.

Since the texts in GNOME-LAB are written by multiple individuals, some variation between them is unavoidable. Nevertheless, this is a rather desirable property in our opinion, if one wants to avoid overfitting the data. Crucially, there is no way to predict in advance the extent to which the expected variation affects the performance of the metrics of entity coherence in the evaluation task. For this reason, it is desirable to use texts from different authors in order to see whether a metric does indeed reflect general preferences for entity coherence shared by different writers.

Taking an application-specific corpus into account makes it possible to identify which constraints on text structure from the genre of interest apply to a real application such as MPIRO. Since MPIRO-PROP is authored by only one domain expert, the question arises whether the results from MPIRO-PROP are due to an idiosyncratic behaviour of this expert or whether they express more general strategies for ordering the data from the particular application. In order to address this question,

additional orderings from multiple experts are gathered and compared to the orderings of the expert consulted by Dimitromanolaki and Androutsopoulos (2003) in a general way in chapter 9.

As Gaizauskas (1998) notices, evaluation efforts in text generation (and fields closely related to it such as dialogue systems and speech synthesis) have to face the possibility that there exist more than one good output. Although we are aware of the problems related to the assumption that a corpus text is "the best possible text", the purpose of our corpus-based evaluation is to estimate the performance of each metric across different texts in an attempt to overcome any implications caused by potential mistakes, the variation between texts and the possibility of many good outputs. In this way, we believe that we manage to account for these problems to a satisfactory extent.

As we explained at the end of the previous chapter, the main motivation behind our approach is to avoid time-consuming evaluation of many generated outputs by restricting human evaluation to a subset of the theoretically **and** empirically motivated metrics. In this sense, we agree with Reiter and Sripada (2002) that the results of a corpus-based evaluation should be treated as hypotheses which need to be integrated with other types of evaluation. Such a subsidiary evaluation which supplements the main methodology of the current chapter is presented in chapter 9.

5.3 The Basis for Comparison

As we discussed in chapter 2, the representation that CT operates on is the CF list. In a standard CT analysis, the surface utterances in a text are translated into a corresponding sequence of CF lists. Since our metrics are defined in terms of CT, for their corpus-based evaluation to take place, each text in our corpora needs to be represented as a sequence of CF lists in a similar way. This is possible not only for GNOME-LAB, our main corpus, which consists of human texts, but also for MPIRO-PROP, our supplementary corpus, which consists of coherent sequences of propositions instead of texts. In order to be able to use the same basic terminology irrespective of the identity of our corpora, will use the term *corpus instances* to refer to what our corpora consist of, which is either texts or coherent sequences of propositions.

As we mentioned already, each corpus instance in GNOME-LAB corresponds to an annotated text. The translation of the corpus instance as a sequence of CF lists gives rise to a representation that is called the *Basis for Comparison* (BfC). The unordered set of CF lists that the BfC consists of is called the *semantic content* of the BfC.

To explain this terminology with an example, let us turn our attention to the text **torc1** in (5.1), one of the shortest corpus instances in GNOME-LAB. Following the existing annotation of the corpus, each utterance in (5.1) is indexed with its unit-id (e.g. u210 is the unit-id for the first utterance):

(5.1) **torc1 corpus instance**:

u210: 144 (top left) is a torc.

u212: Its present arrangement, twisted into three rings, may be a modern alteration;

u214: it should probably be a single ring, worn around the neck.

u216: The terminals are in the form of goats' heads.

The representation of the utterances in (5.1) as a sequence of CF lists is shown in (5.2):[2]

(5.2) **BfC for torc1**:

u210: `CF(de374, de375)`

u212: `CF(de376, de374, de377)`

u214: `CF(de374, de379)`

u216: `CF(de380, de381, de382)`

The sequence of CF lists in (5.2) is the BfC for (5.1). The unordered set of CF lists in (5.2) is the semantic content of (5.2).

For MPIRO-PROP, the term corpus instance does not refer to a sequence of utterances but to a sequence of propositions. Each BfC in MPIRO-PROP corresponds to the translation of a sequence of propositions as a sequence of CF lists.[3] The semantic content is again the unordered set of CF lists that the BfC consists of.

5.4 Exploring the search space

Given a corpus instance of attested coherence and a method for computing the corresponding BfC, one can use a metric of entity coherence to compare the properties of the BfC with the properties of alternative ways of structuring its semantic content. As we discussed in section 2.5.4 of chapter 2, the idea that underlies this operation is that certain properties of the BfC such as the sums of CT transitions can be thought of as expressing a *preference* for entity coherence only if it is possible to structure the semantic content in ways that deviate from the observations from the BfC. In order to be able to estimate this deviation empirically, first one needs to define a metric of entity coherence that

[2]Since each text in the GNOME corpus has already been annotated for information related to its entity coherence we draw on the methods of Poesio et al. (2002) for the translation of the utterances in (5.1) into CF lists. According to this, the entities that the NPs refer to are denoted by the prefix `de` and a number. For example, `de374` in the CF list of unit u210 corresponds to the entity referred to by the NP "144" in (5.1). Section 6.3 of chapter 6 provides more details on the translation of the corpus instances in GNOME-LAB into BfCs.

[3]The main aspects of this procedure have already been introduced in chapter 2. Section 7.3 of chapter 7 provides additional details.

computes such a property from the BfC. Then, a search-oriented experimental methodology is needed to specify the extent to which this metric distinguishes the BfC from its alternatives.

The *System for Evaluating Entity Coherence* (SEEC) is a program that implements the main stage of our corpus-based, search-oriented experimental methodology. In this section, we discuss how SEEC uses a metric of entity coherence to compare the BfC with alternative ways of structuring the same semantic content. Then, we show how the output of SEEC can be used to compute a performance measure for the metric and how this measure can be used in order to compare many metrics on the basis of their performance across many BfCs.

5.4.1 Comparing permutations using SEEC

In this section, we describe how SEEC uses a metric of entity coherence to compare a BfC with alternative ways of ordering its semantic content by navigating through a large search space of possible structures. Figure 5.1 shows the algorithm that SEEC uses to calculate its outputs. The inputs to SEEC are:

a) a BfC B

b) the semantic content of B, SC_B

c) a metric M

SEEC returns three outputs which correspond to the number of permutations of SC_B that score *Better*, *Equal* and *Worse* than B according to M in the explored search space. The outputs are held as the final values of the variables `NBetter`, `NEqual` and `NWorse` respectively.

The output variables are originally set to 0. Let $\mathbf{S}_b$ be the score that M assigns to B. In each iteration, SEEC creates an ordering of the CF lists that constitute SC_B. A sequence of the CF lists as the result of the `permute` operation is called a *permutation*.[4] Assuming that the sequence of CF lists in (5.2) is B, the sequence of CF lists in (5.3) represents another permutation of the members of the semantic content of (5.2). In (5.3) the CF list of u216 appears between the CF lists of u210 and u212:

(5.3) **A permutation of $SC_{(5.2)}$:**

u210: `CF(de374, de375)`

u216: `CF(de380, de381, de382)`

u212: `CF(de376, de374, de377)`

u214: `CF(de374, de379)`

[4]Note that B can also be produced as a permutation of SC_B.

Given B, SC_B and M:
a. Compute S_b, the score that M assigns to B.
b. `NBetter`=0, `NEqual`=0, `NWorse`=0.
c. i is the number of iterations, $t = 0$.
while $t < i$
do
- `permute` SC_B to get Per.
- Compute S_p, the score that M assigns to Per.
- Compare S_p with S_b and
- Increment `NBetter` or `NEqual` or `NWorse` by 1 accordingly.
- $t = t + 1$
end while
Report the final values of `NBetter`, `NEqual`, `NWorse`.

Figure 5.1: Algorithm for calculating how many permutations of the elements of the set SC_B score Better, Equal and Worse than the permutation represented by B according to metric M

SEEC moves through the large search space of possible sequences of CF lists by permuting the members of SC_B. Each permutation is assigned a score $\mathbf{S}_p$ by M which can be directly compared to $\mathbf{S}_b$ using the evaluation method of M.[5] Then, SEEC increments the count of `NBetter`, `NEqual` or `NWorse` by 1, according to the result of the comparison. When the specified number of iterations *i* is reached, SEEC reports the final values of the three output variables, each corresponding to the size of the set of permutations classified as Better, Equal or Worse than B by M in the explored search space.

The exact size of the search space that is explored by SEEC depends on a) the idiosyncrasies of the genre that specify the *permutation strategy* of SEEC, and b) the cardinality of SC_B which determines the *search strategy* of SEEC.

With respect to the permutation strategy, it was suggested in section 2.5.4 of chapter 2 that there is a standard way to start a description, typically by using a title or a phrase of the type: *"This exhibit is a ..."* This means that instead of permuting CF_1, the first CF list of B, this CF list should always appear in the first position of every permutation.[6] In other words, `permute` should operate on all the members of the semantic content with the exception of CF_1.

SEEC implements this preference as the standard permutation strategy in the descriptive genre. This strategy is labelled `not1` and can be seen as a simple heuristic modeling the interaction of entity

[5] See chapter 3 for more details on how two structures are compared with each other according to their scores for M.
[6] Assuming that the sequence of CF lists in (5.2) is B, CF_1 is the one corresponding to the CF list of unit u210.

coherence with a piece of domain-specific communication knowledge (Kittredge et al. 1991), thus excluding some less plausible permutations from the search space. Hence, we assume that when SC_B serves as the input to text structuring, it is possible to identify which of its members will be the initial CF list of the structure to be generated.[7]

Alternatively, SEEC can apply the `permute` operation on the complete set of CF lists that constitute SC_B. This alternative strategy for creating permutations is labelled `all`, but has not been used in the experiments reported in subsequent chapters of the thesis. The complete set of permutations that the search space consists of according to a given permutation strategy is the set of *valid permutations*.

The exact number of permutations that is explored for B depends on the cardinality of SC_B. If SC_B consists of up to 11 CF lists we use exhaustive enumeration of all valid permutations (EX) as the search strategy.[8] This means that when $|SC_B| = n \leq 11$, the permutation strategy being set to `not1`, SEEC goes through the complete search space of size $i = (n-1)!$ by creating all valid permutations of the members of the semantic content with the exception of CF_1 which always appears in the first position of every permutation.

For a larger semantic content we generate 10 random samples of 1,000,000 permutations. In the latter case, the performance of M on B is estimated by calculating its average performance on the 10 random samples. In section 5.6 we show that by using a large random sample we can estimate reliably the performance of M on the whole population of valid permutations for any large n.

5.5 Computing a performance measure

As we mentioned in the previous section, SEEC calculates the sizes of the sets of permutations scoring Better, Equal or Worse than B according to M in the explored search space. The next question is how it is possible to use this output to determine the suitability of the many CT-based metrics for text structuring as required by (Q2) at the end of chapter 3.

We start investigating this question by discussing how two metrics, M_x and M_y, can be compared to specify which one represents a more promising solution for the purposes of a hypothetical text structuring algorithm that uses SC_B as its input. After this is made clear, we show how this comparison can be applied to more than one BfC and all our available metrics.

5.5.1 Calculating the classification rate

Hypothetical distributions of the permutations in the search space generated by SEEC on a) B, SC_B and M_x and b) B, SC_B and M_y are shown in Figure 5.2. What we want to do initially is to compare M_x

[7] As shown in chapter 6 and chapter 7, this is true for the BfCs from both GNOME-LAB and MPIRO-PROP.

[8] EX is implemented in such a way that a different permutation is created in each iteration. Thus, the number of iterations i equals the number of valid permutations.

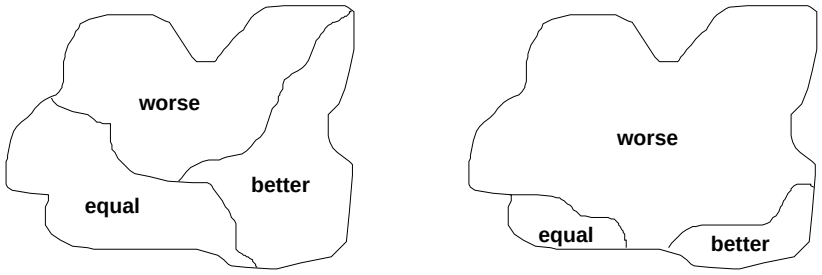

Figure 5.2: Distributions of the search space of possible permutations for metrics M_x and M_y

with M_y according to the portions that the search space is divided into. Such a comparison is called the *individual comparison* of M_x and M_y on B.

Because our ultimate goal is to express the performance of the metrics on search spaces of various sizes defined by any $|SC_B|$, it is best to state the number of permutations classified under each portion of the explored search space in terms of percentages. Hence, Better(M) corresponds to the percentage of the explored search space that scores Better than B according to M, Equal(M) is the percentage of the explored search space that scores Equal to B according to M, etc.

Let us first assume that the search space of size *i* explored for M equals the set of valid permutations of SC_B and that Better(M)= 0%, which means that the set of permutations classified as Better than B by M equals the empty set. In addition, let Equal(M)= $\frac{100}{i}$%, which means that the set of Equal permutations contains just B. If both conditions hold, then M singles out B as the unique best structure in the explored search space. If it is possible to use M to guide a text structuring component that receives SC_B as its input, then M on its own can be seen as a very promising solution for the purposes of structuring SC_B. Under this scenario, B resembles the notion of the gold standard as discussed in section 5.2.1.

Hence, a possible way to resolve the competition between M_x and M_y is to favour the metric that best fits these two requirements, maybe by first considering whether Better(M_y) is smaller than Better(M_x), and then by resorting to the comparison of Equal(M_y) with Equal(M_x) if the percentage of permutations classified in Better is the same for both metrics.

However, what we are mainly interested in is not the individual comparison of M_x and M_y on B per se. Rather, we strive to come up with a way of assessing the performance of the metrics on more than one BfC from a corpus. This is necessary in order to be able to generalise our results safely, accounting for the problems discussed in section 5.2.1. Under this perspective, a more convenient way to account for each individual comparison is by using a single arithmetic measure to express the performance of each metric on the explored search space. This measure can be subsequently used as

the dependent variable in the statistical analysis which compares M_x with M_y across many BfCs.

The dependent variable we employ is called the *classification rate* of a metric M on B. Clearly, there exist several ways of defining the classification rate, the simplest of which might be to compare the metrics according to the percentage of Better only. That is, if Better(M_y) is smaller than Better(M_x), then M_y beats M_x in their individual comparison on B.

However, defining the classification rate solely in terms of the percentage of Better ignores the percentage of permutations classified in Equal. A possible way to account for Equal is by using the sum of the percentage of Better and the percentage of Equal in the definition of the classification rate, thus resolving the individual comparison in favour of M_y, if it returns a lower sum of Better(M_y)+Equal(M_y) than M_x.[9]

Nonetheless, we believe that it is preferable to associate Equal(M) with a weight instead of considering it equivalent to Better(M) in the calculation of the classification rate. The value of the weight for Equal(M) is set to $\frac{1}{2}$. Hence, the classification rate υ of a metric M on B is formally defined as follows:

(5.4) Classification rate

$$\upsilon(M,B) = Better(M) + \frac{Equal(M)}{2}$$

If $\upsilon(M_x,B)$ is the classification rate of M_x on B, and $\upsilon(M_y,B)$ is the classification rate of M_y on B, M_y beats M_x in their individual comparison on B if $\upsilon(M_y,B)$ is smaller than $\upsilon(M_x,B)$. In what follows we explain the motivation behind weighting Equal with $\frac{1}{2}$.

First, let us assume a generation scenario where a permutation of SC_B has a higher chance of being selected as the output of text structuring the better it scores according to M.[10] In that case, the existence of Better structures increases the probability that one of them will be the output instead of B. Even if no Better structures exist, acknowledging a set of Equal structures introduces additional distractors in an attempt to output B. Assuming that the algorithm implements additional biases for the selection of the output when the permutations are assigned the same score for M, the less B violates these constraints the higher are its chances to be favoured over the members of Equal. Conversely, if the additional biases disfavour B, it might end up being the least probable solution for text structuring between the members of Equal. On average, one expects B to sit in the middle of the set of Equal structures when ranked according to the additional biases of the hypothetical text structuring algorithm. Given this scenario, the expected percentage of structures with a higher probability being generated than B is $Better(M) + Equal(M)/2$, i.e. the classification rate of M on B.

[9] Obviously, since Better(M)+Equal(M)+Worse(M)=100%, using Better(M)+Equal(M) is the same as defining the classification rate in terms of Worse(M) and deciding the individual comparison of M_x and M_y on B in favour of the metric with a higher percentage of Worse.

[10] Thanks to Chris Mellish for coming up with this scenario.

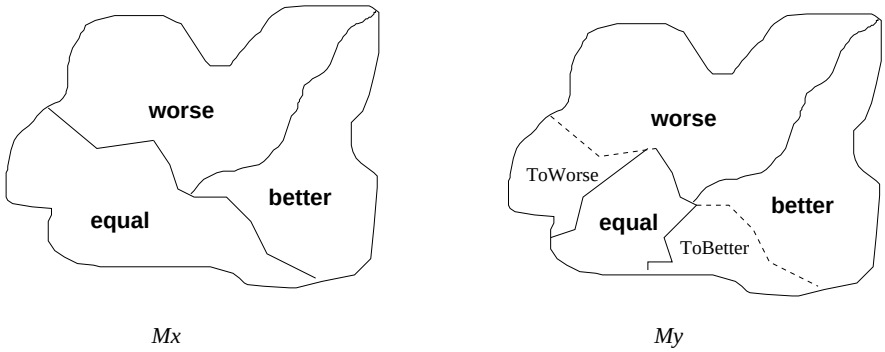

Figure 5.3: Spliting the set of Equal for metric M_x in the distribution of the search space for M_y

In a further attempt to justify the value $\frac{1}{2}$ for the weight of Equal, let us consider Figure 5.3. In this Figure, the members of Equal for M_x are redistributed to either Equal, Better or Worse for M_y, but the members of Better and Worse for M_x are unchanged. This relation is called *SplitEqual* and always arises in the individual comparison of M.NOCB with either M.SHOT1 or M.POT1 when the metrics classify the permutations in the search space in a different way.[11] ToBetter(M_y) is the percentage of permutations that leave Equal of M_x to join Better for M_y. ToWorse(M_y) is the percentage of permutations that leave Equal of M_x and are added to Worse for M_y. SplitEqual(M_y) is defined as the sum of ToBetter(M_y) and ToWorse(M_y).

If the classification rate simply penalises a metric according to Better(M)+Equal(M), M_y needs to classify only one permutation from the Equal of M_x to its own Worse to win the competition with M_x on B.[12] For this reason, we believe that the most objective way of dealing with SplitEqual is to compare ToBetter(M_y) with ToWorse(M_y). If ToBetter(M_y) is larger than ToWorse(M_y), then SplitEqual is divided in a way that favours M_x, but if ToBetter(M_y) is smaller than ToWorse(M_y), then M_y wins the competition with M_x. The idea behind this is that M_y should be thought as doing better than M_x, when it manages to move more than $\frac{1}{2}$ of SplitEqual(M_y) to Worse(M_y), thus reducing Equal(M_y) when compared to Equal(M_x) without increasing Better(M_y) disproportionately. The proof in appendix C shows that when ToWorse(M_y) is higher that half of SplitEqual(M_y), then the classification rate of M_x on B is higher than the classification rate of M_y on B.

Finally, assume that m number of BfCs $B_1, ..., B_m$ from a corpus C and their corresponding semantic contents are used as the subsequent inputs to SEEC, as well as M. A convenient way of summarising the performance of M on the BfCs from C is in terms of the *average classification rate Y* which is

[11] This happens because M.SHOT1 and M.POT1 can return different results from M.NOCB with respect to the comparison of Per with B, only if Per and B have the same number of NOCBs (see section 3.3.2 of chapter 3 for more details).

[12] Similarly, if Better(M) is used in the definition of the classification rate, M_y loses the competition with M_x if a single permutation from the Equal of M_x is reclassified as Better for M_y, even if everything else in SplitEqual is characterised as Worse by M_y.

calculated as follows:

(5.5) Average classification rate

$$Y(M,C) = \frac{\upsilon(M,B_1)+\ldots+\upsilon(M,B_m)}{m}$$

5.5.2 Testing for significance

As we mentioned repeatedly, M_x and M_y should be compared on more than one BfC from a corpus C. Comparing M_x with M_y in terms of their classification rates on the BfCs from C using tests of significance is termed the *pairwise comparison* of M_x with M_y on C.

In our standard experimental analysis, the BfCs $B_1,\ldots,B_m$ from C are treated as the random factor in a repeated measures design since each BfC contributes a score for both metrics. Then, the classification rates for M_x and M_y on the BfCs are compared with each other and significance is tested using the signtest. After calculating the number of BfCs that return a lower classification rate for M_x than for M_y and vice versa, the signtest reports whether the difference in the number of BfCs is significant, that is, whether there are significantly more BfCs with a lower classification rate for M_x than the BfCs with a lower classification rate for M_y (or vice versa).

The signtest was chosen to test significance because it does not carry specific assumptions about population distributions and variance. Considering the small number of BfCs in one of our corpora and their often unequal size, priority is given to a rather conservative statistic like the signtest over its parametric alternatives such as the paired-samples t-test.

5.5.3 Comparing several metrics

The previous section introduced the pairwise comparison of M_x with M_y using the scores from the BfCs in a corpus. Our experiments extend this effort to comparing many metrics of entity coherence, each defined according to a different formulation of CT (see chapter 3 for an overview).

In our analysis, instead of performing all possible pairwise comparisons between the metrics, we define M.NOCB, the simplest metric, as the **baseline** and compare its classification rates with the classification rates of each of the remaining metrics. In this sense, M_x in the previous section stands for M.NOCB, and M_y for one of its competitors. The null hypothesis of each planned pairwise comparison is that the difference between the baseline and M_y will not be significant. If more than one metric is found to significantly outperform the baseline, we conduct pairwise comparisons between them. Note that despite conducting more than one pairwise comparison simultaneously we refrain from further adjusting the overall threshold of significance (e.g. according to the Bonferroni method, typically used for multiple planned comparisons that employ parametric statistics) since it is assumed that

n	$n!$	time
10	3,628,800	4 mins
11	39,916,800	45 mins
12	479,001,600	9 hours and 45 mins
13	6,227,020,800	5 days and 6 hours
14	87,178,291,200	2 months and 14 days
15	1,307,674,368,000	more than 3 years

Table 5.1: Approximate computation time for $n!$ permutations

choosing a conservative statistic such as the signtest already provides substantial protection against the possibility of a type I error.

5.6 Dealing with factorial complexity

One of the most intimidating problems in our search-oriented methodology is the factorial complexity of producing all valid permutations. The second column of Table 5.1 displays the total number of permutations of the members of a set which consists of n elements when n ranges between 10 and 15. The third column of the table shows approximately how long it will take for SEEC to perform the classification task for M.NOCB, the simplest of our metrics, by exhaustively enumerating all possible permutations for a semantic content of cardinality from 10 to 15 CF lists when running on a computer with more than 2GB of real memory. For a semantic content consisting of 10 to 12 propositions we calculated the average of 5 runs on the same input. The last three cells of the third column show an estimation of the computation time for larger inputs. Table 5.1 shows that when n becomes greater than 12, the population of permutations is so large that the operation becomes impractical. Even though using `not1` as the permutation strategy reduces the overall complexity from $n!$ to $(n-1)!$, the gain is too small when compared to the size of the complete search space that remains to be explored.

In our first attempt to overcome this limitation we considered a constraint programming approach to navigate the search space more efficiently. In this context, M.NOCB was implemented in terms of the Oz programming language.[13] Unfortunately, the informal results of this study suggest that although constraint programming appears to be more efficient than exhaustive enumeration for n smaller than 12, search remains impractical for larger search spaces.

The most efficient solution to the problem of factorial complexity was to implement random sampling of permutations (RS) as an alternative search strategy to EX when n is greater than 11. In the

[13] Special thanks to Malte Gabsdil and Peter Dienes for doing the actual implementation. See www.mozart-oz.org for more details on Oz.

remainder of this chapter we provide a theoretical argument and a simple empirical study which show that implementing RS for the purposes of the search task that is carried out by SEEC returns results that are representative of the entire population of permutations.

5.6.1 Sample size

In this section, we use standard statistical notions in order to show that the result returned for a metric M from a random sample of 1,000,000 permutations is representative of the result from the entire population of valid permutations.[14]

To begin with, let us assume that in the whole population of N valid permutations of the members of SC_B, there are b permutations that are classified as Better than or Equal to B according to M. Let the variable X be

- 1 when a randomly selected permutation is classified as Better than or Equal to B,
- 0 otherwise.

The mean value for X is $\mu = b/N$. The variance for X is $\sigma^2 = b/N * (N-b)/N$.[15] The standard deviation σ is equal to the square root of σ^2. Note that the largest value for σ occurs when $\mu = \frac{1}{2}$, in which case σ is also equal to 0.5.

If we take a random sample of n variables like X, the average of the random sample m is equal to: $m = \frac{X_1+...+X_n}{n}$. According to the Central Limit Theorem when the size of the sample is large enough (i.e. $n > 30$) the average m behaves like a normally distributed variable with a mean equal to the population mean μ and a standard error $\sigma_m = \frac{\sigma}{\sqrt{n}}$.

Since m is normally distributed, in order for it to be significantly different from μ at the standard alpha level of $p \leq 0.05$ it has to achieve a z score of at least 1.96.[16] In other words, 95% of the sample means are expected to appear within 1.96 standard errors from the population mean.

Due to the large size of our sample the standard error is $\sigma_m = \sigma/\sqrt{1000000} = \sigma/1000$. Since σ cannot be greater than 0.5, σ_m cannot be greater than 0.0005. This means that we expect the true value of μ to be at most 0.0005*1.96=0.00098 away from m in 95% of our random samples. Due to the very small value of σ_m we are justified to believe that RS is unlikely to return significantly different results from EX for arbitrarily large search spaces. In section 5.6.3 we present a simple empirical study that confirms this claim.

[14] Many thanks to Chris Mellish for developing the basic aspects of this argument. The argument concerns the estimation of the set of Better or Equal permutations. Similar arguments apply to Better, Equal, etc. separately.

[15] This is a special case of the binomial distribution.

[16] A z score of 1.96 corresponds to a probability value of $p = 0.025$ which is the threshold of significance that is required in a two-tailed prediction like the one made here.

n	Replications			
	Within	%	Between	%
12	1032.5	0.10	2087.7	0.10
13	81	0.01	158.6	0.01
14	5.8	0.00	11.5	0.00
15	0.4	0.00	0.8	0.00

Table 5.2: Average number of replicated permutations within and between samples

5.6.2 Implementing random sampling

Our implementation of RS for SEEC is based on Prolog's `random/3` predicate.[17] Each new permutation is built by randomly selecting a CF list from SC_B and placing it to the first available position in the new permutation.[18] Then another CF list is randomly selected between the remaining CF lists in the semantic content and placed in the next available position in the new permutation. The process is repeated until the semantic content is empty.[19]

This implementation allows for permutations to be replicated within and across samples. This means that it is possible for the same permutation to appear more than once within the random sample. In addition, when generating more than one random sample their intersection might not be equal to the empty set. In order to be able to use RS reliably for our purposes, first we need to estimate the average amount of replicated permutations within and between samples.

In order to estimate how many unique permutations of n members of a set are contained within a random sample of 1,000,000 permutations, we varied n between 12 and 15 and generated 10 samples of size 1,000,000 for each n. Then we counted the number of replicated permutations in each sample and calculated their average number for each n. In addition, we tried to estimate how many permutations are replicated between the samples. To do this we counted the number of intersecting permutations in every pair of the 10 samples and calculated their average.

As Table 5.2 shows, when n is equal to 12, the average number of replicated individuals within a sample of 1,000,000 permutations is approximately 1030, that is, almost 99.9% of the permutations in the sample are unique. The average number of intersecting permutations between two samples was less than 2090, which again suggests that for every pair of samples approximately 99.9% of the 2,000,000 permutations are distinct from each other. When n is equal to 13, the average percentage of replicated permutations within and between samples drops to 0.01%. For $n > 13$ the average number

[17] See the Sicstus Prolog manual at http://www.sics.se/sicstus/ for the definition of `random/3`.

[18] This might be the first position in the new permutation when the permutation strategy is `all` or the second position in the new permutation when the permutation strategy is `not1`. When the permutation strategy is `not1`, then the first CF list of the new permutation is CF_1 which is excluded from the random selection from SC_B.

[19] Or left only with CF_1 when the permutation strategy is `not1`.

of replicated permutations is extremely small.

5.6.3 Empirical studies on random sampling

The fact that there are few replicated permutations within a random sample and that different random samples do not consist of the same permutations shows that our implementation of RS does indeed create different random samples of permutations. The next step is to investigate whether using different samples to perform the search task of SEEC returns significantly different calculations of the classification rate for M.

In this section, we report the results from a simple empirical study which shows that different random samples do not behave significantly differently from each other. In addition, when *n* ranges between 10 and 12, in which case calculating $\upsilon(M,B)$ is possible using both RS and EX, the results of RS are not significantly different from the ones returned by EX. This study in addition to the argument brought forward in section 5.6.1 leads us to the conclusion that for the purposes of the operations performed by SEEC one can safely restrict the search space to a few samples of 1,000,000 permutations instead of enumerating exhaustively the complete search space defined by an arbitrarily large *n*.

As a first step in our empirical study on RS we generated 50 random samples of size 1,000,000 by permuting the CF lists of 10 BfCs. 5 random samples were generated for each BfC. The shortest BfCs consisted of 10 CF lists, and the longest of 15 CF lists. Then, we used the output of SEEC to estimate the performance of three of our metrics (namely, M.NOCB, M.SH and M.KP) on each of the samples. The classification rate of the metric on each sample was the dependent variable and the 10 BfCs were treated as the random factor. SAMPLE (5 levels) and METRIC (3 levels) were the two crossing independent variables. Our prediction was that the random variation between the samples would not have any significant effect in the estimation of the average classification rate for each metric.

Indeed, a repeated-measures ANOVA did not find any main effect for the factor SAMPLE on the dependent variable: $F(4,36)=0.15$, $p=0.960$. By contrast, a significant main effect of the factor METRIC was recorded: $F(2,18)=100.05$, $p<0.001$. The interaction between SAMPLE and METRIC was not significant as predicted: $F(8,72)=0.34$, $p=0.948$.

The results of the first empirical study clearly suggest that different samples do not return significantly different results with respect to the performance of the metrics. However, the metrics do differ from each other with respect to their classification rate. In addition, the difference in the performance of the metrics does not interact with the identity of the random sample.

In our second experiment we focused on the problem discussed in section 5.6.1, that is, we tried to estimate how representative of the population value is the performance of a metric M when calculated

using RS instead of EX. 8 BfCs consisting of 10 to 12 CF lists were treated as the random variable.[20] 5 random samples of 1,000,000 permutations were generated for each BfC. Like the previous study, we used the same 3 metrics as one of the two independent variables. We defined the factor SEARCH as our second independent variable with layers RS and EX. The performance of each metric under RS and EX was the dependent variable. The experiment compares the average performance of each metric in the 5 random samples for each BfC with the classification rate returned by enumerating all possible permutations of the CF lists of the same BfC.[21] Our main prediction was that the difference, if any, in the estimation of the classification rate between the two levels of SEARCH will not be significant.

The factors METRIC (3 layers) and SEARCH (RS vs EX) were used as the input to a repeated-measures ANOVA which again revealed a significant main effect of METRIC on the dependent variable: $F(2,14)=46.40$, $p<0.001$. By contrast, the effect of SEARCH and the interaction between METRIC and SEARCH were not significant: $F(1,7)=1.06$, $p=0.337$ and $F(2,14)=0.84$, $p=0.454$ respectively. This confirms the predictions of the theoretical argument in section 5.6.1.

Once again, the results show that the metrics differ with respect to their average performance. In addition, for BfCs that consist of 10 to 12 CF lists, RS does not return significantly different results from EX. Note that since it takes on average around 1.5 mins for SEEC to generate and evaluate a random sample of 1,000,000 permutations the gain in real time is remarkable: As Table 5.1 shows, enumerating exhaustively 12! permutations takes more than 9 hours, whereas the same result can be achieved with a sample of size 1,000,000 which represents less than 0.21% of the entire population of permutations.

Almost all BfCs that we deal with in the experimental studies reported in subsequent chapters consist of less than 12 CF lists, so we had to resort to RS in only a couple of cases. However, the argumentation and the empirical studies in this section make it clear that despite the factorial complexity of the problem that we are confronted with, SEEC is able to estimate the performance of different metrics using semantic contents of arbitrary cardinality.

In the next two chapters we are going to focus on our main research question by investigating the main effect of the factor METRIC more closely. We will be using SEEC in order to perform the corpus-based, search-oriented evaluation aiming to find out which metrics constitute the most suitable candidates for the purposes of text structuring.

[20] These are BfCs for which exhaustive enumeration is practical (see Table 5.1). For BfCs that consist of less than 10 CF lists RS is not sensible since n factorial is smaller than 1,000,000.

[21] The first empirical study suggests that the samples do not differ significantly from each other with respect to the way they estimate the classification rate of the metrics. Therefore, one could use just one of the 5 samples to estimate the classification rate of the metrics rather than their average. However, in our standard experimental methodology we try to explore as much of the search space as possible and minimise the amount of replicated permutations between samples. For this reason we generate more than one random sample and calculate their average performance. We did the same in this empirical study in order to be consistent with our general sampling strategy.

Chapter 6

Experiments on the GNOME corpus

This chapter presents our main corpus-based experimental work on the museum labels of the GNOME corpus. The GNOME corpus provides us with reliable annotation for the main notions that our analysis is based on, namely *units* for the computation of the CF list and *nominal expressions* that introduce entities to the CF list, making it possible to specify the inputs to SEEC automatically.

We begin with the aims of the study, as motivated by the discussion in the previous chapters. Then, we present an overview of the GNOME corpus and the data used in our experiments. The chapter concludes with the discussion of the main results and directions for future work.

6.1 Motivation and aims

In section 2.3.2 of chapter 2, we reviewed the main results of the evaluation of CT by Poesio et al. (2002).[1] In this section, we present the main aims of our research on the GNOME corpus and their relation to the work carried out by Poesio et al. (2002).

As we mentioned in section 2.5.3 of chapter 2, the problem of choice is very important for the purposes of text structuring. However, even though Poesio et al. (2002) present the most methodologically sound evaluation of CT up to now, they do not investigate this problem in detail.

In the previous chapter, we presented an evaluation methodology more appropriate for the purposes of text structuring than the ones previously employed. In this chapter, we describe how we apply this methodology to a subset of the GNOME corpus, using the same tools as Poesio et al. (2002) did for their experiments.[2]

As discussed in chapter 3, our evaluation estimates the performance of a wider range of metrics of

[1] The GNOME corpus has also been used to inform the implementation of general algorithms for referring expression generation (Cheng et al. 2001; Cheng and Mellish 2000b; Henschel et al. 2000; Poesio et al. 1999b).

[2] I am grateful to Massimo Poesio not only for making the data and computational tools of the GNOME corpus available to me, but also for our lengthy discussions and his continuous interest in my work.

entity coherence than ever before. Thus, the main aim of our research is to use data from the GNOME corpus to answer (Q2), as formulated in section 3.7:

Q2: Which metrics of entity coherence constitute the most promising candidates for text structuring?

6.2 The GNOME corpus

The GNOME corpus is divided into three sections: The museum section consists of descriptions of artefacts, argumentative texts about their style and period of creation, etc.[3] The pharmaceutical section is a selection of leaflets providing patients with the legally mandatory information about their medicine.[4] The tutorial dialogues section consists of a subset of the Sherlock corpus (Lesgold et al. 1992).

The GNOME corpus is annotated for a number of features. In what follows we provide an overview of the features that are relevant to our research. Our brief overview of the annotation is a summary of the discussion in Poesio et al. (2002).[5]

Each subcorpus contains about 6,000 NPs. The NPs in the GNOME corpus are called *nominal expressions* (NEs) and are marked up with the <ne> element. The instructions for identifying NE markables derive from those proposed in the MATE scheme for annotating anaphoric relations (Poesio et al. 1999a), which in turn are derived from those proposed by Passoneau (1997) and MUC-7 (Chinchor and Sundheim 1995).

Each NE in the GNOME corpus has a unique id and is annotated for 14 features specifying its syntactic, semantic and discourse properties (Poesio 2000). The attribute which is most relevant to our work is the grammatical function of an NE (gf). The instructions for marking up this feature are derived from those used in the FRAMENET project (Baker et al. 1998). The values for gf include subj, obj, comp (for indirect objects), adjunct (for the argument of PPs modifying VPs), gen (for NEs in determiner position in possessive NEs), etc. Other NE attributes include the NE type (cat), basic syntactic features such as number (num), person (per), and gender (gen), and more abstract features such as animacy (ani), abstract or concrete ontological status (onto), etc.

The NEs in the GNOME corpus are also marked for their anaphoric relations, again using a variant of the MATE scheme. A special <ante> element specifies the id of the anaphor and the type of the anaphoric relation (e.g. identity), whereas an <anchor> tag embedded within the <ante> element indicates the closest antecedent.

[3] The museum subcorpus extends the corpora collected to support ILEX and a related project called SOLE. SOLE extended ILEX with concept-to-speech abilities, using linguistic information to control intonation (Hitzeman et al. 1998).

[4] The leaflets in the pharmaceutical subcorpus are a subset of the collection of all patient leaflets in the UK which was digitised to support the ICONOCLAST project (Scott et al. 1998).

[5] The annotation manual, available from the GNOME project's webpage at http://www.hcrc.ed.ac.uk/~gnome, provides more details.

The NEs in the GNOME corpus are marked up for a number of bridging relations (Clark 1977) in addition to identity (ident). Previous work, particularly in the context of the MUC initiative suggested that while it is fairly easy to mark up identity relations, annotating bridging references is quite hard; this was confirmed by studies such as Poesio and Viera (1998). For this reason, the GNOME corpus is annotated for only three types of relations between objects (and not e.g. anaphoric reference to propositions or events). The relations that are marked up are a subset of those proposed in the "extended relations" version of the MATE scheme and consist of set membership (element), subset (subset) and generalised possession (poss), which includes part-of relations as well as more traditional ownership relations.

In addition to NEs, the GNOME corpus is annotated for all spans of text that can be claimed to update the local focus. This includes sentences (defined on the basis of punctuation as the span of text ending with a full stop, a question mark, or an exclamation point) as well as what is called a discourse unit. Units include clauses (defined as sequences of text containing a verbal complex, all its obligatory arguments and its postverbal adjuncts) as well as other sentence subconstituents that can be claimed to independently update the local focus, such as parentheticals, preposed PPs, the second element of coordinated VPs and layout elements such as titles and list elements.[6]

The following example gives an idea of the annotation in the GNOME corpus at the beginning and the end of unit u227.[7] The surface text reads as follows: "*The drawing of the corner cupboard, or more probably an engraving of it, ...*" The example shows how units like u227 and u228 are marked up for their finiteness. It also shows how various NEs are annotated for their grammatical function (gf). The <ante> element at the end of example (6.1) specifies that ne549 "*it*" is related to ne547 "*the corner cupboard*" via an identity anaphoric relation.

(6.1)

```
<unit finite='finite-yes' id='u227'>

<ne id='ne546' gf='subj'> The drawing of
<ne id='ne547' gf='np-compl'> the corner cupboard
</ne> </ne>

<unit finite='no-finite' id='u228'>, or more probably

<ne id='ne548' gf='no-gf'> an engraving of
```

[6]The instructions for marking up such elements are based on the discussions of clauses in Quirk and Greenbaum (1973) and the proposals for the annotation of discourse units in Marcu (1999).

[7]This is example (15) in Poesio et al. (2002).

```
<ne id='ne549' gf='np-compl'>  it
</ne> </ne>

</unit>,
...
</unit>.

<!-- ne549 ident ne547 -->
<ante current='ne549' rel='ident'>
<anchor antecedent='ne547'>
</anchor> </ante>
```

As we mentioned in section 2.3.2 of chapter 2, one of the objectives in marking up the GNOME corpus was to achieve reliable annotation in order to make corpus-based studies easier to replicate. Consequently, each feature was marked up by at least two annotators. Agreement on the annotation for each feature was checked using the κ statistic (Carletta 1996) on part of the corpus.

Poesio et al. (2002) report that the annotators reached acceptable agreement for most features relevant to the evaluation of CT, with the exception of thematic role which proved too difficult to annotate reliably. Moreover, despite reducing bridging references into only three categories, reaching agreement on this information involved several discussions between the annotators and more than one pass over the corpus.

6.2.1 GNOME-LAB: Museum labels in GNOME

Not all texts, or in our terms *corpus instances* (see section 5.3 of the previous chapter), in the museum section of the GNOME corpus belong to our genre of interest, namely label-like descriptions of artefacts in a museum gallery. In order to restrict the scope of the experiment to the text-type most relevant to our study, a museum label was identified as a short text which describes a concrete entity, typically evoked by the phrase "*This is a ...*" or the title e.g. "*Cabinet*".[8] This informal definition of a museum label is similar to the one in Cheng (2002, p.65).

Our subcorpus, GNOME-LAB, consists of 20 corpus instances that fall under the informal definition of a museum label.[9] Of these, 8 corpus instances are captions of figures from a book about jewellery and have already been part of the SOLE corpus.[10] The 12 remaining labels describe French

[8] It is then this unit that gives rise to CF_1 (see section 5.4.1 of the previous chapter).

[9] Using GNOME's layout annotation, a corpus instance was identified as the segment of text within a <section> tag in the extracts from the SOLE corpus or a tag in the texts from the Getty webpage.

[10] Tait, H., editor (1986). *Seven Thousands Years of Jewellery*. British Museum Publications, London.

artefacts from the 17th and 18th century and come from the webpage of the J. Paul Getty Museum.[11] In the next section, we describe how each corpus instance from the genre of interest was automatically translated into the corresponding inputs to SEEC.

6.3 Computing the inputs to SEEC

The main computational tool in Poesio et al. (2002) is the script which uses GNOME's annotation to compute the CF list of each utterance in a corpus instance according to a particular instantiation of CT. This information is used to calculate the violations of the main claims of CT in the specified configuration.[12] In order to run our experiments, first we employed the script to compute the CF lists according to a certain CT instantiation. Then, the CF lists were translated into a format appropriate for SEEC.[13]

As we mentioned in section 3.7 of chapter 3, although Poesio et al. (2002) experiment with many different ways of computing the CF list using different configurations of CT, our aim is not to fully replicate their experiments. Hence, we did not follow them in invoking many different instantiations of the script. Instead, we used the specification that appeared to be general enough and most relevant to our main research aim.

According to this specification, the CB was defined according to C3 as in standard CT (see section 2.2.2 of chapter 2). All annotated NEs were treated as introducing CFs, but only direct realisation was used for the computation of the CF list, which means that bridging references were not taken into account. The instantiation of the other two main parameters of the theory, namely utterance and CF ranking, are explained in the following sections.[14]

6.3.1 Definition of utterance

In our specification of the script, a CF list is computed for all finite units except for:[15]

- clause complements
- the second element of coordinated VPs
- relative clauses

[11] http://www.getty.edu/museum/

[12] Hence, the scoring function of entity coherence in Poesio et al. (2002) is incorporated in the script that computes a certain instantiation of CT.

[13] Many thanks to James Soutter for writing the program that performs this transformation.

[14] We do not discuss the instantiation of parameters that are not relevant to our study e.g. which NEs should be counted as pronouns for the purposes of R1, etc.

[15] As in Poesio et al. (2002), titles, which are non-finite units, give rise to independent CF lists as well.

Starting from Kameyama (1998), complements have consistently been shown not to have their own CF list. As in Poesio et al. (2002), we took the CFs of complements to be part of the CF list of the main finite unit they belong to. In a construction such as [$_{VP}$ VP_1 and VP_2], VP_2 is recognised as the second element of the coordination. Depending on whether a trace is computed for the empty subject of VP_2 or not, the elements of the coordinated VP either always satisfy CONTINUITY or introduce a potential NOCB. Following Poesio et al. (2002) again, VP_1 and VP_2 give rise to a single CF list.

The script in Poesio et al. (2002) makes it possible to define CF lists for embedded units, such as finite and non-finite relative clauses and appositions.[16] These units are often surrounded by their superordinate clause, as shown in (6.2):

(6.2) [$_{u291}$ The painted reserves, [$_{u292}$ which imitate carved stone reliefs], show mythological scenes ...].

In example (6.2), `u292` is "center-embedded" within `u291`. The main problem with this construction is that there is no obvious way to order `u292` in relation to its superordinate unit `291` for the definition of the BfC.

However, it seems to us that the relation between `u292` and `291` is more relevant to sentence planning rather than text structuring. For this reason, we decided that it is more appropriate to treat relative clauses and appositions in the same way as complements, that is, by considering their CFs as part of the CF list of the main finite unit they are embedded in.[17] Note that Poesio et al. (2002, pp.31-33) reach the same conclusion, using arguments relevant to their own methods of evaluation. When the script is invoked in this way, the CF lists of the specified finite units are readily put in a sequence for the BfC based on the order that the units appear in the surface text.

We named each corpus instance in GNOME-LAB after the entity it describes. For instance, the corpus instance which the units in example (6.2) come from is conventionally called **vase1**. A BfC that results from the specification of utterance as described in this section is given the suffix **-finite**. Thus, in the BfC of **vase1** is **vase1-finite**. We refer to this way of computing the BfCs collectively as **Finite**.

In the following section, we present an adjustment to the Finite computation of the BfC which helps us investigate to some extent the effect of rhetorical coherence on the performance of the metrics employed in the experiment.

[16] A detailed analysis of the different kinds of modifiers and embedded units in the GNOME corpus appears in Cheng (2002, Chapter 4).

[17] Thus, we assume that the embedding of these units precedes text structuring in the NLG architecture that our metrics are tested to be suitable for. Then, it is the larger units that result from embedding which are ordered by the text structuring component.

6.3.2 Computing CF lists for local rhetorical relations

Some of the units that are used for the computation of the CF list are linked to an adjacent finite unit via a local rhetorical relation (RR) explicitly marked with a cue phrase such as “because”, “but”, “although”, etc. Using a cue phrase as the signal for a RR, we identified 19 RRs in 12 corpus instances from GNOME-LAB.[18] These RRs belong mainly to the informational type (Moore and Pollack 1992; Moser and Moore 1996) and are reminiscent of ILEX-like local RS-trees, for instance:

(6.3) [$_{u185}$ Access to the cartonnier’s lower half can only be gained by the doors at the sides, [$_{u186}$ because the table would have blocked the front]].

In all but one case, the finite units that are related with each other via a RR appear within the same sentence which consists only of these units. These are cases of isomorphism between the surface form of the sentence and the RS-tree that the sentence is analysed to.

As we mentioned in section 2.1.4 of chapter 2, the rhetorical information in (6.3) can pose additional constraints in the architecture of an NLG system. For example, RRs like the one in (6.3) are recorded in the database of ILEX as relation nodes. When a relation node is selected as part of the input to text structuring, the system tries to express it by building a local RS-tree. This tree cannot be interrupted by intervening spans in the text structure.

In the Finite computation of the BfC, rhetorically related units give rise to independent CF lists. We are interested in investigating the performance of the metrics when we treat the RS-trees in GNOME-LAB as units that cannot be interrupted by other units. This will give us the opportunity to explore which metrics are more suitable for an NLG architecture that requires the RRs recorded in the database to be expressed locally during the text structuring process. For this reason, we create an additional input to SEEC where the finite units connected via a RR give rise to a single CF list. This way of computing the BfCs is denoted with the suffix **-finite-RR**.

For example, (6.4) shows the CF lists of the units in example (6.3) as computed in the BfC **carto1-finite**. The entities that the NEs refer to are denoted by the prefix `de`.[19]

(6.4)
```
%%% u185
%%% Access to the cartonnier's lower half
%%% can only be gained by the doors at the sides,

CF(de12, de13, de1, de15, de16)
```

[18] Most RRs (15/19) appear in 10 corpus instances from the Getty webpage. The remaining 4 appear in 2 corpus instances from SOLE. Although we acknowledge that cue phrases are not the only hint for a RR, it has been shown that they constitute a very reliable way of detecting one (Knott and Dale 1994).

[19] More details on the referents and the ranking of the CFs are given in the next section.

```
%%% u186
%%% because the table would have
%%% blocked the front.

CF(de9, de18)
```

Because of the isomorphism between the local RS-tree and the sentence boundaries in (6.3), the CF list of the RS-tree can be computed by using "sentence" instead of "finite unit" for the definition of utterance and defining the other parameters of the script in the same way as for the computation of the CF lists in (6.4). The CF list for the sentence that contains `u185` and `u186` is shown in (6.5). This CF list is used as the CF list of the local RS-tree in (6.3) as well.[20]

(6.5)
```
%%% RR: u185-u186
%%% s55
%%% Access to the cartonnier's lower half
%%% can only be gained by the doors at the sides,
%%% because the table would have
%%% blocked the front.

CF(de12, de9, de18, de13, de1, de15, de16)
```

The CF list of `s55` in (6.5) replaces the CF lists of `u185` and `u186` in (6.4) in the computation of the BfC **carto1-finite-RR**. The next section provides more details on the ranking of the CFs in (6.4) and (6.5).

6.3.3 CF ranking

Following most mainstream work on CT for English, we used grammatical function combined with linear order within the unit (`gftherelin`) for CF ranking: In this configuration, the CP, i.e. the first member of the CF list, corresponds to the first NE within the unit that is annotated as a subject for its `gf`.[21]

[20] Despite the isomorphism between RS-trees and sentences in GNOME-LAB, it would be a mistake to consider the relationship between sentences consisting of more than one finite unit and RS-trees as 1:1. We identified 15 sentences consisting of more than one finite unit which are not related to each other via an explicit RR marked with a connective although they appear within the same sentence. These units are represented as subsequent CF lists in both ways of computing the BfC. A more detailed and general study on the lack of isomorphism between document and rhetorical structure, motivated mainly by examples from GNOME's pharmaceutical section appears in Power et al. (2003).

[21] Or as the post-copular NE in a *there-clause*.

In general, Poesio et al. (2002) report that despite the fact that CF ranking is one of the most debated issues within CT, different specifications of CF ranking did not appear to have a significant effect on their evaluation. This result is not very surprising given that most of the different suggestions on CF ranking are based on evidence from languages other than English. There seems to be considerable consensus, however, that as far as English is concerned using `gftherelin` is a very robust way of estimating CF ranking. Hence, we expect that different ranking functions would not have made much of a difference for our study either.

We will now explain how `gftherelin` gives rise to the CF lists in examples (6.4) and (6.5). In (6.4), the subject of `u185` "*Access to the cartonnier's lower half*" denotes the entity `de12`. Using `gftherelin`, `de12` is ranked as the first member of the CF list of `u185`. Similarly, `de9` is ranked as the CP of `u186` because `de9` is the entity denoted by the NE "*the table*" which functions as the surface subject of `u186`. Note that in both cases, the CP can be seen as corresponding to Arg1 if an ILEX-like representation of the predications underlying (6.4) is used.

Sentence `s55` in example (6.5) contains two NEs annotated as subjects. The CP of `s55` is `de12` because the NE "*Access to the cartonnier's lower half*" that denotes `de12` precedes the NE "*the table*" that denotes `de9` within the sentence.[22]

Notice that if `u186` happened to precede `u185` within the sentence, it would have been `de9`, that is, the referent of the subject of `u186`, that would have been ranked as the CP of `s55` by `gftherelin`. This is the main difference between `gftherelin` and `rs-tree`, which is the ranking function introduced in section 2.4.4 of chapter 2 for the computation of the CF list of an ILEX-like local RS-tree.

More specifically, according to `rs-tree`, since `u185` is the nucleus of the local RS-tree within `s55`, its Arg1 `de12` will always be ranked as the CP of `s55` irrespective of the order of `u185` and `u186` within `s55`. Since Poesio et al. (2002) do not account for the existence of local RRs in GNOME-LAB, the nucleus-satellite distinction was not taken into account in the possible specifications of CF ranking by their script.

From the various suggestions within CT, there seems to be one that comes very close to `rs-tree` but has not been tested by Poesio et al. (2002) either. This is the CF ranking in Miltsakaki (2002) which is also based on grammatical function, but makes the same predictions as `rs-tree` about the CP of `s55`. Instead of referring to the underlying rhetorical structure, Miltsakaki (2002) uses notions from traditional grammar to argue that the subject of the main clause in a sentence, in our case `de12`, should always be ranked as the CP.

In conclusion, although we believe that `gftherelin` was the best available choice for CF ranking at the time that our study took place, we need to point out that using it imposed an additional restriction

[22] As we mentioned already, we found only one case of a local RS-tree consisting of two finite units each forming a single sentence. Since it is not possible to use the script to compute the CF list for this RS-tree automatically as in (6.5) above, we computed its CF list by hand, using the surface order of the sentences for the linearisation of CFs with the same `gf`.

on the assumed NLG architecture that the metrics were tested to be suitable for. More specifically, `gftherelin` assumes that the order of the units that an RS-tree consists of is defined **before** the text structuring process. Somehow differently, Miltsakaki (2002) claims that ordering clauses within a sentence is orthogonal to ordering sentences. Finally, in O'Donnell et al. (2001) the ordering of the fact nodes within an RS-tree appears to be dealt by the sentence planning module quite independently of the text structuring process.

6.4 Experimental questions

As the previous sections have discussed, GNOME-LAB was used to compute two of the inputs to SEEC, namely the BfC and its semantic content. The third input was each of the metrics in Table 3.11 of chapter 3.[23] Our experimental work aims at answering (Q2), using the various BfCs from GNOME-LAB as the random factor and their classification rates for each metric of entity coherence as the dependent variable.

Finite was chosen among the many specifications of CT evaluated by Poesio et al. (2002) as the configuration which is most relevant to our research aims. In addition to this, we want to explore which metrics do best with respect to (Q2) when local RRs are taken into account for the computation of the CF lists as happens in Finite-RR, and whether these metrics are different from the metrics that perform best in Finite. A related experimental question, which involves another potential difference between Finite and Finite-RR, is discussed in the following section.

6.4.1 Rhetorical compensation

As section 2.3.2 of chapter 2 reports, Poesio et al. (2002) set out to find the version of CT with as few NOCB transitions as possible,[24] the underlying assumption being that a version of CT with fewer NOCBs is a "better" model of entity coherence than a version with more NOCBs. Because none of the tested versions completely eliminates NOCBs, Poesio et al. (2002) conclude that entity coherence should be supplemented by other coherence inducing factors at the local level and suggest the framework of Knott et al. (2001) as a plausible model for the interaction between rhetorical and entity coherence.

As we explained in section 2.1.3 of chapter 2, Knott et al. (2001) assume that rhetorical and entity coherence are not simultaneous constraints on text structure. In a further attempt to account for the existence of NOCBs in a structure, one could follow their assumption to hypothesise that when a NOCB

[23] As already mentioned in section 5.4.1 of the previous chapter, the permutation strategy used in the experiments was always `not1`. The search strategy was always EX, with the exception of one corpus instance that gives rise to BfCs consisting of more than 11 CF lists both in Finite and in Finite-RR. The search strategy in these cases was RS.

[24] Or the versions of CT where there are significantly fewer NOCBs than other transitions (Poesio et al. 2003).

exists, the units which violate CONTINUITY are related rhetorically. In other words, local RS-trees introduce some sort of *rhetorical compensation* for the NOCBs in a structure.[25] According to this hypothesis, moving from the Finite to the Finite-RR computation of the BfC is expected to reduce the percentage of NOCBs because (most of) the RRs between two adjacent units make up for a potential violation of CONTINUITY.[26]

Indeed, the change from Finite to Finite-RR results in a lower percentage of NOCBs for 8 BfCs.[27] Although Finite-RR has more NOCBs than Finite for 4 BfCs, the average percentage of NOCBs for the 12 BfCs in Finite-RR is 53%, that is 4% lower than in Finite, thus providing some weak evidence in favour of rhetorical compensation overall.

However, investigating the effect of rhetorical relations on entity coherence simply by calculating the percentage of NOCBs for each way of computing the BfC does not account for the fact that the number of NOCBs in a BfC might not estimate its entity coherence reliably (see section 2.5.4 of chapter 2 for more details). As this issue is addressed by our methodology under a specific NLG perspective, the interesting question from our point of view about the interaction between rhetorical and entity coherence is not simply to calculate the percentage of NOCBs as an indication of rhetorical compensation but to examine whether the move from Finite to Finite-RR results in a decreasing classification rate for M.NOCB as well. M.NOCB is singled out as the metric of interest because it uses the number of NOCBs as the sole measure of the incoherence of a structure. Given the observed reduction on the percentage of NOCBs, we are interested to see whether the percentage of permutations which score Better than or Equal to the BfC is also reduced with the move from Finite to Finite-RR. This question is examined in section 6.5.3.

6.5 Results

In this section we report the main results of the experiments on GNOME-LAB. First, we report the average classification rate (Y) of each metric as a way of summarising its performance across the whole corpus. Then, we attempt to answer (Q2) using the BfCs from Finite and Finite-RR in a set of

[25] The term is due to Jon Oberlander.

[26] As example (6.5) shows, moving from Finite to Finite-RR is also expected to make the BfC shorter, i.e. consisting of fewer CF lists, since the CFs that come from finite units that used to contribute independent CF lists to the BfC in Finite now appear together in the same CF list in Finite-RR. Indeed, the 12 BfCs in Finite-RR contain on average 1.58 fewer CF lists when compared to their average number of CF lists in Finite. This in turn means that the set of valid permutations is smaller in Finite-RR than in Finite.

[27] However, in all 8 cases the percentage of the pairs of utterances that violate one or more of the underlying principles of CT, namely COHERENCE, CHEAPNESS or SALIENCE, is increased. That is, Finite-RR has fewer NOCBs than Finite as predicted, but the transitions that are mainly introduced by this change are not CONTINUEs. Poesio et al. (2002) report a similar result with respect to the change from direct to indirect realisation for the computation of the CF list.

Finite				Finite-RR			
Metric	Better	Equal	Y	Metric	Better	Equal	Y
M.NOCB	8.66	22.58	19.95	M.NOCB	9.25	27.96	23.24
M.POT1	15.33	10.02	20.34	M.SHOT1	16.49	19.30	26.15
M.SHOT1	11.43	18.69	20.77	M.POT1	23.65	6.33	26.81
M.MIL	10.82	25.29	23.47	M.MIL	14.79	26.24	27.91
M.SH	15.19	27.72	29.05	M.SH	17.15	27.02	30.66
M.BFP	26.33	13.37	33.01	M.BFP	28.71	9.38	33.39
M.CHEAP	35.56	43.35	57.23	M.KP	39.38	34.99	56.87
M.KP	40.60	35.24	58.22	M.CHEAP	41.98	40.22	62.10
N	20			N	12		

Table 6.1: Average classification rate (Y=Better+Equal/2) in Finite and Finite-RR

pairwise comparisons with the baseline, M.NOCB.[28] Finally, we discuss the differences between the two ways of computing the BfCs and how these relate to the question raised in the previous section.

6.5.1 Average classification rate

Table 6.1 shows the average performance of the eight metrics in the experiment. The metrics are sorted in ascending order according to their average classification rate (Y) for each way of computing the BfC. In addition to this, the Table reports the average percentage of Better and Equal permutations, as well as the number of BfCs (N) in Finite and Finite-RR.

Before entering the discussion of the main questions that motivate this chapter, some preliminary observations are appropriate. First, one needs to keep in mind that on average more than 50% of the recorded transitions in Finite are NOCBs (as Poesio et al. 2002 have already reported) and the same is true for Finite-RR. However, the Y value for M.NOCB in each way of computing the BfC shows that on average the BfC appears very close to the top 20% of alternative permutations when these permutations are ranked according to their probability of being selected as the output of the hypothetical text structuring algorithm in section 5.5.1 of the previous chapter.

This shows that even though the ordering of CF lists in the BfC might not be completely minimising the number of observed NOCB transitions, the BfC tends to be in greater agreement with the preference to avoid NOCBs than most of the alternative orderings. In this sense, it appears that the BfC optimises with respect to the number of potential NOCBs to a certain extent. As in section 2.5.4 of

[28] As section 5.5.2 of the previous chapter explains, a metric M beats the baseline if it has a lower classification rate than the baseline for significantly more BfCs than the BfCs for which the classification rate for the baseline is lower than for M.

Pair	Winner	
	Finite	Finite-RR
M.NOCB vs M.CHEAP	M.NOCB	M.NOCB
M.NOCB vs M.KP	M.NOCB	M.NOCB
M.NOCB vs M.BFP	M.NOCB	ns
M.NOCB vs M.SH	ns	ns
M.NOCB vs M.MIL	ns	ns
M.NOCB vs M.SHOT1	ns	ns
M.NOCB vs M.POT1	ns	ns

Table 6.2: Winners of pairwise comparisons with M.NOCB in Finite and Finite-RR

chapter 2, we maintain that exploring the search space of valid permutations provides more information on the role of entity coherence as a text structuring constraint than simply calculating the number of observed NOCBs in the BfC.

Moving to issues closer to our experimental questions, it is worth noting that the metric with the lowest Y both in Finite and in Finite-RR is the baseline, M.NOCB. Moreover, Y(M.NOCB, Finite) is lower than Y(M.NOCB, Finite-RR), 19.95% versus 23.24% respectively. Indeed, the average classification rate of all metrics except for M.KP is greater in Finite-RR than in Finite.

As we stated in the previous chapter, Y is reported for purely descriptive purposes. The following section presents the results of the statistical analysis which employs the signtest in a set of pairwise comparisons between M.NOCB and each competing metric of entity coherence in order to investigate potential answers to (Q2).

6.5.2 Pairwise comparisons with M.NOCB

Table 6.2 provides an overview of the results of the pairwise comparisons with M.NOCB in Finite and Finite-RR. The first column of the Table identifies the comparison in question, e.g. M.NOCB versus M.CHEAP. The next two columns report which metric is the "winner" of the pairwise comparison, that is, which metric has a lower classification rate than its competitor for significantly more BfCs in each way of computing the BfC. In our example, the baseline M.NOCB has a lower classification rate than M.CHEAP in significantly more BfCs both in Finite and Finite-RR. By contrast, in the comparison of M.NOCB versus M.POT1, neither metric has a lower classification rate for significantly more BfCs than its competitor in either Finite or Finite-RR.

The exact number of BfCs for which the classification rate of M.NOCB is lower than its competitor for each pairwise comparison in Finite is reported in the second column of Table 6.3. For example,

Pair	M.NOCB			p	Sig	Winner
	lower	greater	ties			
M.NOCB vs M.CHEAP	18	2	0	0.000	***	M.NOCB
M.NOCB vs M.KP	16	2	2	0.001	***	M.NOCB
M.NOCB vs M.BFP	12	3	5	0.018	*	M.NOCB
M.NOCB vs M.SH	7	4	9	0.274	ns	
M.NOCB vs M.MIL	5	6	9	0.500	ns	
M.NOCB vs M.SHOT1	4	6	10	0.377	ns	
M.NOCB vs M.POT1	7	10	3	0.315	ns	
N	20					

Table 6.3: Details of pairwise comparisons with M.NOCB in Finite

M.NOCB has a lower classification rate than M.CHEAP for 18 (out of 20) BfCs in Finite. M.CHEAP manages a lower classification rate for only 2 BfCs in Finite, while there are no ties, i.e cases where the classification rate of the two metrics is the same. The p value, rounded in the third decimal point, which is returned by the signtest for the difference in the number of BfCs is reported in the fifth column of the Table.

The last two columns report the achieved level of significance and the winner of the comparison. In this and subsequent similar tables throughout the thesis, three asterisks (***) indicate $p \leq 0.001$, as happens for the comparison of M.NOCB versus M.CHEAP, while two asterisks (**) are used for $0.001 < p \leq 0.01$, and one asterisk (*) for $0.01 < p \leq 0.05$. When the p value is greater than 0.05, the result is reported as not significant (ns) and no winner is reported for the pairwise comparison. Similarly, Table 6.4 reports the numbers of BfCs for each category, the associated p values, the achieved level of significance and the winners of the pairwise comparisons in Finite-RR.

The Tables clearly show that M.NOCB does significantly better than M.CHEAP and M.KP irrespective of the way that the BfC is computed. Most additional comparisons fail to reach significance, a result which might be due to the small sample size used in this study. In general, we acknowledge that this is a serious limitation which raises issues of statistical power and generalisability.[29] However, we also point the attention of the reader to the fact that some of the observed effects **are** strong enough to reject the null hypothesis, despite the small size of the samples. Crucially, all significant differences in the pairwise comparisons are in favour of M.NOCB. In addition to this, none of the other comparisons shows a trend to the opposite direction.

For example, while there are significantly more BfCs with a lower classification rate for M.NOCB

[29] Note that these issues are addressed in the next chapter, where a larger corpus is used to evaluate the same metrics.

Pair	M.NOCB			p	Sig	Winner
	lower	greater	ties			
M.NOCB vs M.CHEAP	10	2	0	0.019	*	M.NOCB
M.NOCB vs M.KP	11	1	0	0.003	**	M.NOCB
M.NOCB vs M.BFP	7	5	0	0.387	ns	
M.NOCB vs M.SH	5	3	4	0.363	ns	
M.NOCB vs M.MIL	4	4	4	0.500	ns	
M.NOCB vs M.SHOT1	2	4	6	0.344	ns	
M.NOCB vs M.POT1	5	6	1	0.500	ns	
N	12					

Table 6.4: Details of pairwise comparisons with M.NOCB in Finite-RR

than for M.BFP in Finite, the result of the comparison of M.NOCB versus M.BFP in Finite-RR is not significant. Keeping the discussion in the previous paragraph in mind, if one applies Occam's logical principle to resolve a pairwise comparison when the signtest fails to reach significance, a simple metric such as M.NOCB that employs only the violations of the prerequisite of CONTINUITY in its scoring function will be given priority over a competitor such as M.BFP that makes use of a more complicated model of entity coherence, especially given that M.NOCB is found to overtake two other potential competitors overwhelmingly in the same sample. Although we are careful enough to admit that lack of significance per se does not provide conclusive evidence against a certain competitor of M.NOCB, we do maintain that the results of the study in GNOME-LAB can be interpreted as indicating that M.NOCB is a baseline hard to overtake for the rest of the employed metrics of entity coherence.

6.5.3 Differences between Finite and Finite-RR

In section 6.5.1 we spotted a general trend for the average classification rate to rise with the move from Finite to Finite-RR. In order to investigate this trend a bit further we conducted signtests on the classification rates of the 12 BfCs in Finite-RR and the corresponding classification rates in Finite for each metric, the details of which are displayed in Table 6.5. The second column of the Table shows the number of BfCs for which the classification rate in Finite-RR is lower than in Finite for a given metric, whereas the third column shows the number of BfCs with a greater classification rate in Finite-RR than in Finite. The last section of the Table reports the associated p value for the difference on the number of BfCs, the achieved level of significance and the winner of the comparison, i.e. the way of computing the BfC which contains significantly more BfCs with a lower classification rate.

Metric	Finite-RR			p	Sig	Winner
	lower	greater	ties			
M.NOCB	3	9	0	0.073	ns	
M.SH	4	8	0	0.194	ns	
M.SHOT1	4	8	0	0.194	ns	
M.POT1	4	8	0	0.194	ns	
M.BFP	5	7	0	0.387	ns	
M.MIL	2	10	0	0.019	*	Finite
M.KP	10	2	0	0.019	*	Finite-RR
M.CHEAP	9	3	0	0.073	ns	
N	12					

Table 6.5: Details of the comparisons of Finite-RR versus Finite for each metric

The Table confirms the already observed trend of the metrics to return a greater classification rate in Finite-RR than in Finite. Although only M.MIL returns significantly more BfCs with a greater classification rate in Finite-RR than in Finite, all other comparisons are in the same direction as for M.MIL, except for the ones involving M.CHEAP and M.KP. In fact, M.KP reports significantly more BfCs with a lower classification rate in Finite-RR than in Finite. However, since both M.CHEAP and M.KP are defeated by M.NOCB in Finite-RR (see Table 6.4) as well as in Finite, this improvement in the classification rate is of little use.

Note that Table 6.5 is relevant to the question raised in section 6.4.1 as well. More specifically, the classification rate of M.NOCB is lower in Finite-RR than in Finite for **only** 3 BfCs. As we have already reported, changing from Finite to Finite-RR reduces the percentage of NOCBs for 8 BfCs. As Table 6.6 shows, the 3 cases where the classification rate of M.NOCB is lower in Finite-RR than in Finite do indeed belong to this category. However, for the remaining 5 BfCs from the same category the reduction on the percentage of NOCBs in the BfC results in the classification rate of M.NOCB being **greater** in Finite-RR than in Finite.[30] This shows that even though a BfC in Finite-RR may often have fewer NOCBs than in Finite, the percentage of permutations that score Better than or Equal to the BfC is not always lower in Finite-RR than in Finite. This situation is similar to the examples discussed in section 2.5.4 of chapter 2, showing once again that using the percentage of NOCBs absolutely to estimate the entity coherence of the BfC is not as informative as going through the search space of valid permutations.

[30]The 4 additional BfCs with a greater classification rate in Finite-RR are the ones where the percentage of NOCBs in the BfC rises with the move from Finite to Finite-RR.

Finite-RR vs Finite		
% of NOCBs	ʊ of M.NOCB	
	lower	greater
lower	3	5
greater	-	4

Table 6.6: Distribution of BfCs according to the difference in a) the percentage of NOCBs and b) the classification rate (ʊ) of M.NOCB between Finite-RR and Finite

6.5.4 Discussion

Assuming the generation scenario in section 5.5.1 of the previous chapter, the results in section 6.5.2 suggest that if one is provided with a semantic content similar to what emerges from the BfCs in GNOME-LAB as the input to the hypothetical text structuring algorithm and has to choose which of the eight candidate metrics to use to guide the algorithm (aiming to arrive at the BfC as the output), the baseline M.NOCB is a better choice than M.KP or M.CHEAP because it returns a smaller percentage of permutations with a higher probability of being selected than the BfC in significantly more cases than the other two metrics.

Moreover, since none of the remaining five metrics manages to overtake M.NOCB, one can provisionally argue that it is the baseline which represents the most suitable candidate for text structuring (between the ones considered) in the genre represented by GNOME-LAB. This holds irrespective of whether local RS-trees are taken into account or not for the computation of the input to the hypothetical algorithm.

The decision whether RRs should be accounted for in this input could be quite independent of which way of computing the BfC returns better results in the corpus-based evaluation of the metrics. For instance, O'Donnell et al. (2001) emphasise the importance of representing RRs in ILEX's database, so from their point of view the best performing metrics in Finite-RR would probably be preferable to the best performing metrics in Finite irrespective of any differences in their performance since entity coherence is in practice inferior to RRs in characterising a descriptive text structure. In this sense, the fact that M.NOCB proves robust across both ways of computing the BfC is an additional point in its favour, at least as far as the comparison with the other candidate metrics goes.

Section 6.5.3 indicates that although the search space of valid permutations becomes smaller when local RS-trees are taken into account, there are proportionally more permutations scoring Better than and Equal to the BfC in Finite-RR than in Finite. This seems to suggest that RRs and entity coherence are not supplementary but conflicting constraints on text structuring since optimising on one does not result in optimising on the other. It also shows that if M.NOCB is used to guide the generation

process (always according to the scenario in section 5.5.1), the chances of the BfC being chosen over its alternatives will probably be reduced when the input is computed according to Finite-RR instead of Finite. However, as we have already mentioned, this cannot be used as the sole criterion to decide which way of computing the BfC best serves as the input to text structuring.

At this point we need to emphasise that even though M.NOCB can be taken to be the most promising candidate, it is still the case that on average around 20% of the permutations appear more likely to be selected than the BfC (see Table 6.1). This shows that M.NOCB needs to be supplemented by other features to improve its performance, although this cannot be achieved with the help of RRs as already discussed.[31]

In general, the difference in the two types of BfC is an issue that we intend to investigate in more detail in the future, as the following section points out. After reviewing the main directions for future work, we conclude the chapter with a summary of the main findings from the investigation of the selected subset of the GNOME corpus.

6.6 Future work

There are two main directions for future work: a) specifically for the GNOME corpus and b) involving more general issues that apply to the thesis as a whole.

As far as the GNOME corpus is concerned, given the importance of bridging references in the evaluation of Poesio et al. (2002), we would like to experiment with a configuration of CT which uses indirect realisation for the computation of the CF list. Moreover, we intend to investigate the difference between the Finite and the Finite-RR way of computing the BfC in more detail, e.g. by identifying factors that might account for the drop in the classification rate more clearly than the percentage of NOCBs. Last but not least, we are looking forward to an extension of the GNOME corpus so that more corpus instances are added in GNOME-LAB, which in turn might enable us to investigate more subtle differences than the ones reported in this chapter.

With respect to the general directions for future work, we remain committed to the employment of more metrics of entity coherence from chapter 3 in order to investigate whether they can yield better results in their pairwise comparison with the baseline than the metrics used in the current experiments. Moreover, we would be interested in investigating the performance of the metrics on a different genre.[32]

[31] Another attempt to supplement M.NOCB with an additional domain-specific constraint on coherence giving rise to similar results as the attempt to use RRs is discussed in chapter 8.1.2.

[32] In addition to the museum labels, there are three texts on the Getty webpage which provide biographical information about three famous craftsmen from the 18th century. Our preliminary investigation of the performance of the metrics on these texts reveals that they show different preferences from the museum labels. However, due to the tiny size of the sample it is not possible to defend the generality of these results which are not reported here. Thus, a more extended study on the

6.7 Summary

In summary, the main result from the study on the specified subset of the GNOME corpus is that none of the seven employed metrics of entity coherence manages to return significantly better results than the baseline M.NOCB. By contrast, M.NOCB, which simply employs the number of NOCB transitions in its scoring function, performs significantly better than M.KP and M.CHEAP given both ways of computing the BfC, namely Finite and Finite-RR. These results indicate that M.NOCB sets a baseline difficult to overtake for any metric that makes use of additional constraints for entity coherence.

Finite-RR computes the CF lists for each BfC by taking the local RRs between adjacent finite units into account. We noticed that in general the performance of the metrics drops as one moves from Finite to Finite-RR. This is the case for M.NOCB as well, although a BfC in Finite-RR has often fewer NOCBs than in Finite. Although this cannot be used as the only criterion to decide which way of computing the BfC best serves as the input to text structuring, it suggests that RRs and entity coherence are conflicting constraints in the overall coherence of the structures from GNOME-LAB.

As we mentioned in section 6.5.4, although the tradeoff between rhetorical and entity coherence is one of the most interesting issues that emerged from this study, we shall not pursue it any further in the remaining chapters. Instead, we will continue to explore potential answers for (Q2) using additional data from a specific application domain.

As mentioned in section 5.2 of the previous chapter, GNOME-LAB represents the genre of interest, namely human-authored museum labels. Additional constraints on text structuring can be imposed by a particular application. The domain of application is represented in our study by the MPIRO system which lacks informative RRs and specifies rather simple preferences for sentence planning. What is more, because the corpus from MPIRO is much larger than the available data in GNOME, it makes it possible to address the issues of power and generalisability raised in this chapter. In the following chapters we describe the experimental efforts within the context of MPIRO and how these relate to the study in GNOME.

behaviour of biographies awaits future work.

Chapter 7

Initial experiments on the MPIRO domain

This chapter presents our initial experiments on data from the MPIRO system which are appropriate for the task performed by SEEC. We begin with the aims of our study and an overview of the available data. Then, we present the results of our experiments and compare them with the results from GNOME-LAB reported in the previous chapter. The chapter concludes with a summary of the main findings from the two investigated datasets.

7.1 Motivation and aims

As mentioned in section 5.2 of chapter 5, GNOME-LAB represents the genre of interest in our study. Additional constraints on text structuring can be imposed by a particular application. The MPIRO system provides us with the opportunity to investigate the performance of the metrics on data from an existing NLG system and was chosen as the application-specific domain of our study. Hence, the aim of this chapter is to use data from MPIRO to answer (Q2), first defined in section 3.7 of chapter 3 and repeated throughout the thesis:

Q2: Which metrics of entity coherence constitute the most promising candidates for text structuring?

Moreover, we investigate whether the findings from MPIRO are similar to the results obtained from GNOME-LAB in order to identify which of these results apply to a real application such as MPIRO.

As we already mentioned in section 2.5 of chapter 2, MPIRO is the multilingual extension of ILEX. At the time that our experiments took place, the main difference between MPIRO and ILEX was that the database of MPIRO did not represent rhetorical relations. Hence, MPIRO is a predominantly entity-based NLG domain which prima facie appears to be very suitable for our purposes. In the next two sections, we present the acquired data in more detail and comment on their portability to the task performed by SEEC.

7.2 Database facts in MPIRO

Our application-specific corpus, MPIRO-PROP, contains 122 ordered sets of propositions (facts) and is a subset of the 880 orderings employed in Dimitromanolaki and Androutsopoulos (2003). Dimitromanolaki and Androutsopoulos (2003) derived the facts from the database of the MPIRO system and assigned them to sets, each set of propositions being treated as a hypothetical input to text structuring. Each set consisted of 6 facts which were manually assigned an order to reflect what a domain expert (EM) considered to be the most natural ordering of the corresponding sentences in the text to be generated.

An example of a set of facts as ordered by EM is shown in (7.1):[1]

(7.1) a. subclass(exhibit1, amphora):
This exhibit is an amphora.

b. creation-period(exhibit1, archaic-period):
This exhibit comes from the archaic period.

c. period-story(archaic-period, entity-4009):
The archaic period ranges from 700 to 480 BC.

d. creation-time(exhibit1, date-894):
This exhibit dates from the early 5th century BC.

e. painting-technique-used(exhibit1, red-figure-technique):
This exhibit was painted using the red figure technique.

f. technique-description(red-figure-technique, entity-2474):
In the red figure technique, the background was painted black and the figures that were predesigned had the natural color of the clay.

The dataset in Dimitromanolaki and Androutsopoulos (2003) was divided into ten disjoint parts, each of which was used for testing while the other 9 parts were used for training. Each time, two standard machine learning (ML) algorithms were applied to the orderings of EM in the training set to specify the most natural ordering of the facts. The information that the ML algorithms primarily used was the fact-type (predicate), i.e. “subclass”, “creation-period”, etc.[2] A set of classifiers informed by the results of ML was then used to generate an order for each set of facts in the testing set. This text

[1] Each fact is accompanied with a context-independent realisation of its content, derived by EXPRIMO, MPIRO’s generation engine. More details on realising facts out of context are given in section 9.2.3 of chapter 9.

[2] Note that some fact-types in Dimitromanolaki and Androutsopoulos (2003) express information on the generic type of the entity that appears as their Arg1. For instance, the generic fact-type “story” is broken down into a set of fact-types such as “period-story”, “painter-story”, etc.

structuring technique did significantly better than two baselines in terms of their average precision when compared to the orderings of EM.

We obtained 225 randomly sampled orderings from the dataset of Dimitromanolaki and Androutsopoulos (2003).[3] In the next section, we explain in more detail how the acquired data were transformed into the inputs to SEEC and why our evaluation had to be restricted to only a subset of them.

7.3 Computing the inputs to SEEC

As we explained in section 2.4.2 of chapter 2 in some detail, we regard the Arg1 of a fact as the CP of the corresponding CF list. Hence, example (7.1) above readily gives rise to the BfC in (7.2).[4] Note that, unlike Dimitromanolaki and Androutsopoulos (2003), the information that SEEC relies on is not the fact-types, but directly the arguments of the facts.

(7.2) a. `CF(exhibit1, amphora)`

b. `CF(exhibit1, archaic-period)`

c. `CF(archaic-period, entity-4009)`

d. `CF(exhibit1, date-894)`

e. `CF(exhibit1, red-figure-technique)`

f. `CF(red-figure-technique, entity-2474)`

In our preliminary experimentation, we found out that a large part of the randomly selected corpus was not informative for our purposes. An example of such an ordering, accompanied by the context-independent realisations of the facts it consists of, is given in (7.3):

(7.3) a. subclass(exhibit41, lekythos):
This exhibit is a lekythos.

b. original-location(exhibit41, attica):
This exhibit originates from Attica.

c. creation-time(exhibit41, date-4475):
This exhibit dates from between 475 and 470 BC.

[3] I am grateful to Aggeliki Dimitromanolaki for providing me with the random samples used in this study and in the experiments reported in chapter 9. As the chapter shows, experimenting with a reasonably large sample makes it easy to identify which data are indeed meaningful for our purposes and investigate potential differences with GNOME-LAB.

[4] Many thanks to David Schlangen for writing the program which translates the acquired data into a format appropriate for SEEC. Note that, like (7.2a), CF_1 always comes from the fact with the "subclass" predicate. Since the database of MPIRO does not record RRs, we did not have to specify CF lists for local RS-trees.

d. painting-technique-used(exhibit41, red-figure-technique):
This exhibit was painted using the red figure technique.

e. exhibit-depicts(exhibit41, entity-4492):
This exhibit depicts an athlete preparing to perform a long jump.

f. current-location(exhibit41, national-archaeological-athens):
This exhibit is currently displayed in the National Archaeological Museum of Athens.

All CF lists in the BfC from (7.3) have the same CP, namely `exhibit41`. When such a BfC is used as the input to SEEC, all metrics return 100% for the percentage of permutations classified in Equal because every permutation of the CF lists is assigned the same score as the BfC. The BfCs which display this behaviour define the class of **AllEq**.

AllEq accounts for 45.78% (103/225) of the BfCs from the randomly sampled orderings. Luckily, more than half of the dataset (122/225, 54.22%) gives rise to BfCs which do not score the same as every permutation of their semantic content. This set of 122 corpus instances defines MPIRO-PROP, the application-specific corpus of our study.[5] Following our standard methodology defined in chapter 5, MPIRO-PROP gave rise to the inputs to SEEC together with the metrics in Table 3.11 of chapter 3.[6]

7.4 Results

In this section, we report the main results of the experiments on MPIRO-PROP. We begin with the average classification rate (Y) of each metric, an often informative summary statistic. Then, we attempt to answer (Q2) in a set of pairwise comparisons with the baseline M.NOCB. The results in each section are discussed in comparison with the corresponding findings in GNOME-LAB. The chapter concludes with a summary of the main findings from the two investigated datasets.

7.4.1 Average classification rate

The fourth column in the first section of Table 7.1 shows the average classification rate of the metrics on the BfCs from MPIRO-PROP. The metrics are sorted according to the returned value of Y from

[5] In general, ordering the CF lists of a BfC from AllEq is considered to be a problem that is tackled very crudely by the employed metrics of entity coherence, e.g. by choosing randomly one ordering. Although section 9.7 of chapter 9 provides some evidence that the ordering task in AllEq is less constrained than in MPIRO-PROP, it is plausible that other preferences, unrelated to entity coherence as modelled by the metrics, do exist in AllEq as well. In this sense, the algorithms in Dimitromanolaki and Androutsopoulos (2003), which make only indirect use of the notion of entity coherence, are much more informed than any metric of entity coherence in the way they order the data from AllEq. However, at this stage we are only interested in comparing the metrics with each other and not with alternative approaches. An experimental design towards that direction is outlined in section 9.6.5 of chapter 9.

[6] As in GNOME-LAB, the permutation strategy used in the experiments was always `not1`. The search strategy was always EX. See section 5.4.1 of chapter 5 for more details on these terms.

MPIRO-PROP				GNOME-LAB			
				Finite		Finite-RR	
Metric	Better	Equal	Y	Metric	Y	Metric	Y
M.BFP	13.76	12.30	19.91	M.NOCB	19.95	M.NOCB	23.24
M.MIL	7.70	24.73	20.06	M.POT1	20.34	M.SHOT1	26.15
M.NOCB	7.75	25.34	20.42	M.SHOT1	20.77	M.POT1	26.81
M.SHOT1	14.13	17.32	22.79	M.MIL	23.47	M.MIL	27.91
M.POT1	19.55	11.48	25.29	M.SH	29.05	M.SH	30.66
M.SH	11.74	44.90	34.18	M.BFP	33.01	M.BFP	33.39
M.KP	34.13	38.04	53.15	M.CHEAP	57.23	M.KP	56.87
M.CHEAP	65.42	31.23	81.04	M.KP	58.22	M.CHEAP	62.10
N	122			N	20	N	12

Table 7.1: Average classification rate (Y=Better+Equal/2) in MPIRO-PROP and GNOME-LAB

lowest to highest. The percentage of Better and Equal permutations for each metric is also reported as well as the number of BfCs (N). The last four columns of Table 7.1 show the performance of the metrics in Finite and Finite-RR in the same order as in Table 6.1 of section 6.5.1.

Whilst M.NOCB is found to have the lowest Y for both BfC versions in GNOME-LAB, it is overtaken slightly in MPIRO-PROP by two metrics, M.BFP and M.MIL, which do not do as well in GNOME-LAB. The relative positions of the other metrics in MPIRO-PROP do not appear to be very different from GNOME-LAB with M.SHOT1 and M.POT1 immediately following M.NOCB, whereas M.KP and M.CHEAP have the greatest classification rates and always appear last in the Table.[7]

In the following section, we report on the pairwise comparisons of the metrics in MPIRO-PROP which investigate a) whether the baseline is beaten by the two metrics with the lowest Y and b) which metrics are overtaken by the baseline in a similar way as in GNOME-LAB.

7.4.2 Pairwise comparisons with M.NOCB

Table 7.2 presents an overview of the winners of the pairwise comparisons with M.NOCB in MPIRO-PROP and GNOME-LAB. Table 7.3 shows the details of the pairwise comparisons in MPIRO-PROP.[8]

[7] Also note that MPIRO-PROP is similar to GNOME-LAB in that on average the BfC appears close to the top 20% of alternative permutations when these permutations are ranked according to their probability of being selected as the output of the hypothetical text structuring algorithm in section 5.5.1 of chapter 5. As we argued in section 6.5.1, this result indicates that although the ordering of CF lists in the BfC might not be completely minimising the observed number of NOCBs it does optimise on the number of potential NOCBs to a certain extent.

[8] See section 6.5.2 of the previous chapter for clarifications on the notation used in the Tables.

Pair	Winner		
	MPIRO-PROP	GNOME-LAB	
		Finite	Finite-RR
M.NOCB vs M.CHEAP	M.NOCB	M.NOCB	M.NOCB
M.NOCB vs M.KP	M.NOCB	M.NOCB	M.NOCB
M.NOCB vs M.BFP	ns	M.NOCB	ns
M.NOCB vs M.SH	M.NOCB	ns	ns
M.NOCB vs M.MIL	M.MIL	ns	ns
M.NOCB vs M.SHOT1	M.NOCB	ns	ns
M.NOCB vs M.POT1	M.NOCB	ns	ns

Table 7.2: Winners of pairwise comparisons with M.NOCB in the two datasets (MPIRO-PROP and GNOME-LAB)

Pair	M.NOCB			p	Sig	Winner
	lower	greater	ties			
M.NOCB vs M.CHEAP	110	12	0	0.000	***	M.NOCB
M.NOCB vs M.KP	103	16	3	0.000	***	M.NOCB
M.NOCB vs M.BFP	41	31	49	0.121	ns	
M.NOCB vs M.SH	100	17	5	0.000	***	M.NOCB
M.NOCB vs M.MIL	0	6	116	0.016	*	M.MIL
M.NOCB vs M.SHOT1	44	14	64	0.000	***	M.NOCB
M.NOCB vs M.POT1	52	11	59	0.000	***	M.NOCB
N	122					

Table 7.3: Details of pairwise comparisons with M.NOCB in MPIRO-PROP

As Table 7.2 reveals, a striking commonality between the results from the two datasets is the very bad performance of M.CHEAP and M.KP compared to the baseline. Indeed, M.NOCB in MPIRO-PROP is even stronger than in GNOME-LAB, since it manages to beat M.SH, M.SHOT1 and M.POT1 in addition to M.CHEAP and M.KP.

Even though Table 7.1 reports Y(M.BFP, MPIRO-PROP) to be somewhat lower than Y(M.NOCB, MPIRO-PROP), M.NOCB is actually found to have a lower classification rate than M.BFP for 41 BfCs in Table 7.3, while M.BFP has a lower classification rate than M.NOCB for only 31 BfCs. Despite the fact that the difference in the number of BfCs is not significant, inability to beat the baseline in the pairwise comparison is not good news for M.BFP, as explained in section 6.5.2 of the previous chapter.[9]

Thus, it seems that the most genuine competitor of M.NOCB in MPIRO-PROP is M.MIL. Note that the two metrics differ only on 6 BfCs, all of which return a greater classification rate for M.NOCB than for M.MIL. In all other cases, the classification rate of the two metrics is the same. In the next section, we investigate the difference between the two metrics in more detail using additional data from GNOME-LAB.

7.4.3 Examining the number of ROUGH-SHIFTs in the BfC

As we mentioned in the previous section, there are 6 BfCs in MPIRO-PROP which return a lower classification rate for M.MIL than for M.NOCB. Since the classification rate of the two metrics is the same in all other cases, the pairwise comparison ends up in favour of M.MIL. An example of the 6 corpus instances, accompanied by the associated standard CT transitions in addition to the realisations of the facts it consists of, is given in (7.4):

(7.4) a. subclass(exhibit26, relief):
This exhibit is a relief.

b. creation-period(exhibit26, classical-period):
This exhibit was created during the classical period.
CONTINUE

c. creation-time(exhibit26, date-5946):
This exhibit dates from circa 470 BC.
CONTINUE

d. location-found(exhibit26, acropolis):

[9] This is particularly true for the relatively large sample in MPIRO-PROP. Note that M.NOCB versus M.BFP is the only pairwise comparison failing to reach significance in MPIRO-PROP.

This exhibit was found in the Acropolis.

CONTINUE

e. current-location(exhibit26, acropolis-museum):
This exhibit is currently displayed in the Acropolis Museum.

CONTINUE

f. region(acropolis-museum, acropolis):
The Acropolis Museum is in the Acropolis.

SMOOTH-SHIFT

Two ROUGH-SHIFTs are created by reversing the order of the last two facts as shown in example (7.5):

(7.5) a. subclass(exhibit26, relief-generic-instance):
This exhibit is a relief.

b. creation-period(exhibit26, classical-period):
This exhibit was created during the classical period.

CONTINUE

c. creation-time(exhibit26, date-5946):
This exhibit dates from circa 470 BC.

CONTINUE

d. location-found(exhibit26, acropolis):
This exhibit was found in the Acropolis.

CONTINUE

f. region(acropolis-museum, acropolis):
The Acropolis Museum is in the Acropolis.

ROUGH-SHIFT

e. current-location(exhibit26, acropolis-museum):
This exhibit is currently displayed in the Acropolis Museum.

ROUGH-SHIFT

Since (7.5) has the same number of NOCBs as the ordering of EM in (7.4), the two orders are rendered equivalent by M.NOCB.[10] By contrast, M.MIL classifies (7.5) as Worse than (7.4) since its scoring function takes into account the sum of ROUGH-SHIFTs in addition to the sum of NOCBs. The 6 BfCs

[10] In this example, the BfCs of (7.4) and (7.5) do not contain any NOCBs. However, 2 of the 6 BfCs in question contain one NOCB.

MPIRO-PROP			
ROUGH-SHIFTs in BfC	M.NOCB		
	lower	greater	ties
none	-	6	116
one or more	-	-	-
Finite			
ROUGH-SHIFTs in BfC	M.NOCB		
	lower	greater	ties
none	-	6	9
one or more	5	-	-
Finite-RR			
ROUGH-SHIFTs in BfC	M.NOCB		
	lower	greater	ties
none	-	4	4
one or more	4	-	-

Table 7.4: Distribution of the BfCs according to a) their number of ROUGH-SHIFTs and b) the result of M.NOCB versus M.MIL in MPIRO-PROP, Finite and Finite-RR

for which M.MIL is found to have a lower classification rate than M.NOCB are all cases like (7.4), where it is possible to reorder the semantic content of the BfC in such a way as to create at least one additional ROUGH-SHIFT without causing more NOCBs. These additional ROUGH-SHIFTs are penalised by M.MIL, but not taken into account by M.NOCB.

As the top section of Table 7.4 shows, MPIRO-PROP does not contain any BfCs with one or more ROUGH-SHIFTs. GNOME-LAB, however, includes BfCs with ROUGH-SHIFTs. Hence, it allows us to explore how this property of the BfC relates with the result of the pairwise comparison of M.NOCB versus M.MIL in more detail.

The remaining sections of Table 7.4 show the distribution of the BfCs in GNOME-LAB according to their number of ROUGH-SHIFTs and the results of the pairwise comparison of M.NOCB versus M.MIL. Crucially, both in Finite and Finite-RR, when the BfC has at least one ROUGH-SHIFT, then the classification rate of M.NOCB is always lower than M.MIL. Like MPIRO-PROP, when the BfC does not contain a ROUGH-SHIFT, then the classification rate of M.NOCB is either the same or greater than the classification rate of M.MIL.

On the one hand, the proportion of BfCs without a ROUGH-SHIFT in GNOME-LAB for which M.MIL beats M.NOCB (6/15 in Finite and 4/8 in Finite-RR) appears to be greater than in MPIRO-

PROP (6/122). On the other hand, the table shows that BfCs with at least one ROUGH-SHIFT do exist outside MPIRO-PROP, which always result in the classification rate for M.NOCB being lower than for M.MIL. In these cases, the penalty imposed for ROUGH-SHIFTs by M.MIL is too great to allow these BfCs to score better than permutations that have the same number of NOCBs but do not contain any ROUGH-SHIFTs.

In conclusion, when a BfC does not have ROUGH-SHIFTs, the classification rate of M.MIL might or might not be lower than the classification rate of M.NOCB. Crucially, when a BfC has at least one ROUGH-SHIFT, then the classification rate of M.MIL is always greater than the classification rate of M.NOCB. Thus, the significantly better performance of M.MIL in its pairwise comparison with M.NOCB is probably due an idiosyncrasy of MPIRO-PROP, namely, the complete lack of BfCs with at least one ROUGH-SHIFT.

7.4.4 Discussion

The results in section 7.4.2 suggest that there will be proportionally more permutations with a higher probability of being selected than the BfC if M.CHEAP or M.KP are used instead of M.NOCB to structure a semantic content similar to what is provided by MPIRO (modulo AllEq) as well as GNOME-LAB under the generation scenario in section 5.5.1 of chapter 5. Thus, M.CHEAP and M.KP can be identified as the least suitable candidates for text structuring (between the ones considered) both in the specific application domain and in the investigated genre.

The results of the pairwise comparisons of M.NOCB with M.SH, M.SHOT1 and M.POT1 in GNOME-LAB are not significant. Although we acknowledge that lack of significance per se does not provide conclusive evidence against the competitors of M.NOCB, the facts that a) the only significant differences found in GNOME-LAB are in favour of M.NOCB and b) the pairwise comparisons which do not achieve significance do not reveal any particular tendency in any direction, at the very least show that M.NOCB is a baseline difficult to beat in GNOME-LAB.

The results from MPIRO-PROP manifest the superiority of M.NOCB in the context of a particular application quite emphatically since it does significantly better than M.SH, M.SHOT1 and M.POT1 in addition to M.KP and M.CHEAP and cannot be overtaken by M.BFP either. Thus, the only metric which manages to beat M.NOCB, albeit marginally, is M.MIL. However, as the previous section shows, this is should be attributed to the absolute lack of BfCs with ROUGH-SHIFTs, a feature specific to MPIRO-PROP.

Given the fact that the data derived from MPIRO were ordered by only one expert, the question that arises at this point is whether the results from MPIRO-PROP reflect general strategies for ordering the application-specific data or whether they are specific to this expert.[11] In chapter 9, we attempt to

[11] Note that the same question holds for the evaluation methodology of Dimitromanolaki and Androutsopoulos (2003) as

answer this question by gathering orderings from other experts and comparing them to the orderings of EM.

Finally, the analysis in section 7.4.3 can also be taken to suggest that, instead of giving general priority to one or more metrics for the purposes of text structuring, which metric is the best candidate to structure a certain semantic content might depend on factors such as whether it is possible to come up with an ordering containing a ROUGH-SHIFT given this input or some relation between the percentage of possible ROUGH-SHIFTs and their expected frequency in the BfC, etc. Although we conducted some preliminary work in this direction, the results are not conclusive enough to be reported in the context of this thesis. Hence, this is another interesting issue which awaits future work as pointed out in chapter 10.

7.5 Summary of chapters 6 and 7

In summary, both MPIRO-PROP and GNOME-LAB show that significantly more BfCs return a greater classification rate for M.CHEAP and M.KP than M.NOCB. Hence, M.CHEAP and M.KP can be identified as the least suitable candidates for text structuring both in the specific application domain and in the genre of interest.

The pairwise comparisons of M.NOCB with the remaining five metrics in GNOME-LAB show that none of them manages to overtake the baseline, a result which is interpreted as favouring M.NOCB over its competitors. The results from MPIRO-PROP manifest the superiority of M.NOCB in the context of a specific application quite emphatically since it does significantly better than most of its competitors with the exception of M.MIL. However, as section 7.4.3 shows, the marginal difference in favour of M.MIL should be attributed to the absolute lack of BfCs with ROUGH-SHIFTs, a feature specific to MPIRO-PROP.

Up to now we have made exclusive use of the experimental methodology from chapter 5 which simply considers the position of the BfC in the explored search space. This has already provided us with a testable hypothesis, namely that the most promising solution for the purposes of text structuring is M.NOCB (or M.MIL specifically for MPIRO-PROP).

Even though using either of these metrics for the purposes of text structuring is motivated by the fact that it is for these candidate solutions that the BfC has best chances of being selected among its alternatives under the generation scenario in section 5.5.1 of chapter 5, this does not exclude permutations other than the BfC from being selected by M.NOCB or M.MIL with equal or even higher probability. Clearly, a closer look at (some of) these structures is necessary in order to get a more complete picture of the orderings favoured by the metrics.

---

well. However, this question is not relevant to the results from GNOME-LAB, the corpus-instances of which were written by different authors.

In the next chapter, the structures that are assigned the best scores by M.NOCB and M.MIL in MPIRO-PROP are investigated more closely. This introduces an additional, possibly domain-specific, constraint on entity coherence which motivates the final set of pairwise comparisons. The thesis concludes with an experiment which extends the study on MPIRO-PROP using orderings produced by more than one expert.

Chapter 8

The role of PageFocus

The previous chapter recognised M.NOCB and M.MIL as the most promising candidates for text structuring in the MPIRO domain and identified the structures for which the two metrics differ. This chapter begins by inspecting the structures that get the best scores by these metrics more closely. This investigation equips the employed metrics with an additional constraint on entity coherence and motivates a new set of pairwise comparisons between the modified metrics. In these comparisons, a number of modified metrics overtake the baseline in MPIRO-PROP, but not in GNOME-LAB. This identifies a number of promising candidates for text structuring in the particular application domain, but shows that M.NOCB remains very robust as far as the genre of interest is concerned.

8.1 Motivation

As we mentioned at the end of the previous chapter, inspecting the set of best scoring permutations for M.MIL and M.NOCB is expected to provide us with a more complete picture of the orderings favoured by the metrics that are hypothesised to be the most promising solutions for the purposes of text structuring according to the experimental methodology of chapter 5.

In this section, the orderings that are assigned the best scores by M.NOCB and M.MIL in MPIRO-PROP are investigated more closely. This introduces an additional constraint on entity coherence called the *PageFocus* which accounts for differences which are not captured by the definition of the metrics in chapter 3.

8.1.1 Computing the BestTable

In order to investigate which permutations are assigned the best score by the evaluation function of a metric M, the algorithm in section 5.4.1 of chapter 5 was modified to output the set of permutations

that are favoured by M. M and a set of CF lists, i.e. the semantic content SC_B of a BfC B, serve as the inputs to the algorithm. The first CF list in B is marked as CF_1 in SC_B.

The algorithm goes through the complete set of valid permutations of SC_B by always placing CF_1 in the first position of a permutation and permuting the remaining facts.[1] The scoring function of M is used to calculate a score for each permutation and the permutations are ranked according to their scores. The output of the algorithm is the set of best scoring permutations according to M denoted as BestTable(M).[2]

8.1.2 The PageFocus constraint on entity coherence

In our preliminary investigations we inspected the BestTable of M.NOCB and M.MIL for the semantic contents of a number of BfCs from MPIRO-PROP. We soon found out that the BestTable of M.NOCB and M.MIL often contains, among others, the following types of permutations:[3]

(8.1)
a. subclass(exhibit1, amphora):
This exhibit is an amphora.

b. painted-by(exhibit1, painter-of-Kleofrades):
This exhibit was decorated by “the painter of Kleofrades”.
CONTINUE

c. painter-story(painter-of-Kleofrades, entity-4049):
“The painter of Kleofrades” used to decorate big vases.
SMOOTH-SHIFT

d. exhibit-depicts(exhibit1, entity-914):
This exhibit depicts a warrior performing splachnoscopy before leaving for the battle.
NOCB

e. current-location(exhibit1, martin-von-wagner-museum):
This exhibit is currently displayed in the Martin von Wagner Museum.
CONTINUE

f. museum-country(martin-von-wagner-museum, germany):
The Martin von Wagner Museum is in Germany.
SMOOTH-SHIFT

[1] This operation corresponds to the permutation strategy `not1` (see section 5.4.1 of chapter 5).

[2] When the permutations which score better than the BfC are assigned with different scores by M, BestTable(M) is a subset of the set of Better permutations the cardinality of which is calculated by SEEC. When no permutation is scoring Better than the BfC, then the BestTable consists of the permutations classified as Equal by SEEC.

[3] As in chapter 7, the facts in the examples are accompanied by their context-independent realisations and the associated standard CT transitions (see section 9.2.3 of chapter 9 for more details).

(8.2) a. subclass(exhibit1, amphora):
This exhibit is an amphora.

b. painted-by(exhibit1, painter-of-Kleofrades):
This exhibit was decorated by “the painter of Kleofrades”.
CONTINUE

d. exhibit-depicts(exhibit1, entity-914):
This exhibit depicts a warrior performing splachnoscopy before leaving for the battle.
CONTINUE

e. current-location(exhibit1, martin-von-wagner-museum):
This exhibit is currently displayed in the Martin von Wagner Museum.
CONTINUE

f. museum-country(martin-von-wagner-museum, germany):
The Martin von Wagner Museum is in Germany.
SMOOTH-SHIFT

c. painter-story(painter-of-Kleofrades, entity-4049):
“The painter of Kleofrades” used to decorate big vases.
NOCB

The ordering assigned to the set of facts by EM is shown in (8.1), while (8.2) is a permutation with the same number of NOCB and ROUGH-SHIFT transitions as (8.1).[4] Crucially, although both entities are discourse-old, having `exhibit1` as the CP in the NOCB utterance (8.1d) seems to be a better strategy than the one followed in (8.2c) where the CP of the NOCB utterance is the `painter-of-Kleofrades`.

O’Donnell et al. (2001) set `exhibit1` as the focal entity of the whole description which always consists of a single page of hypertext.[5] We will subsequently refer to this entity as the *PageFocus* (PF). Note that assigning the PF with a special status also relates to work on CT which discusses how pragmatic constraints such as situational deixis interact with the prominence of the entities in the discourse model (e.g. Walker et al. 1994; Turan 1998; Hoffman 1998).

We define the PF as the visually salient entity which initiates the generation process. In terms of Reiter and Dale (2000, p.81), the PF corresponds to the parameter of the metalevel communicative goal DescribeExhibit(PF), where PF is the artefact to be described. A particular instantiation of PF results in the selection of a given set of facts from the database.

[4] Note that because (8.1) contains fewer CONTINUEs than (8.2), it scores worse than (8.2) according to M.BFP.

[5] The main difference between (8.2) and an ILEX-like resumption is that (8.2c) is not followed by utterances providing additional information about the `painter-of-Kleofrades`. See sections 2.1.3, 2.1.4 and 2.4.4 of chapter 2 for more details on the role of resumption in the text structure assumed by ILEX.

Since both examples contain the same number of NOCBs (and no ROUGH-SHIFTs), neither M.NOCB nor M.MIL can distinguish them from each other. In fact, because the phenomenon in question is more related to the type of the NOCB transition rather than whether a NOCB can be avoided or not, it cannot be accounted for by any metric employed in chapter 3. Hence, in order to differentiate between examples (8.1) and (8.2), we postulate a distinction between two types of NOCB:

- Plain NOCB where, given a pair of utterances violating CONTINUITY, the CP of the second utterance is the PF:

(8.1') c. painter-story(painter-of-Kleofrades, entity-4049):
"The painter of Kleofrades" used to decorate big vases.
SMOOTH-SHIFT

d. exhibit-depicts(exhibit1, entity-914):
This exhibit depicts a warrior performing splachnoscopy before leaving for the battle.
NOCB, CP=PF=exhibit1

- NOCBPF$*$ where, given a pair of utterances violating CONTINUITY, the CP of the second utterance is **not** the PF:

(8.2') f. museum-country(martin-von-wagner-museum, germany):
The Martin von Wagner Museum is in Germany.
SMOOTH-SHIFT

c. painter-story(painter-of-Kleofrades, entity-4049):
"The painter of Kleofrades" used to decorate big vases.
NOCBPF$*$, CP$\neq$PF=exhibit1

Further to this, a NOCBPF$*$ is taken to be a more serious violation of entity coherence than a NOCB. Then, a *PF-modified* set of metrics can be defined by incorporating the definition of NOCBPF$*$ to the scoring functions of the metrics employed in chapter 3. As shown in Table 8.1, the scoring function of each PF-modified metric computes the sum of NOCBPF$*$s independently from the other scores.[6]

The evaluation method of each PF-modified metric first compares a pair of permutations with respect to the sum of NOCBPF$*$s. The permutation with the smallest sum of NOCBPF$*$s wins the competition. Thus, all PF-modified metrics will now prefer example (8.1) over (8.2). If the sum of NOCBPF$*$s is found to be the same, then the permutations are compared using the rest of the scores

[6]Obviously, Sum(NOCBPF$*$) in e.g. PF.NOCB excludes all plain NOCBs. We will refer to M.NOCB from Table 3.11 in chapter 3 as the non-PF *counterpart* of PF.NOCB.

Name	Scoring Function
PF.NOCB	Sum(NOCBPF∗), Sum(NOCB)
PF.CHEAP	Sum(NOCBPF∗), Sum(CHEAP∗)
PF.MIL	Sum(NOCBPF∗), Sum(NOCB)+Sum(ROUGH-SHIFT)
PF.SH	Sum(NOCBPF∗), Sum(NOCB)+Sum(COH∗)
PF.KP	Sum(NOCBPF∗), Sum(NOCB)+Sum(COH∗)+Sum(CHEAP∗)+Sum(SAL∗)
PF.SHOT1	Sum(NOCBPF∗), Sum(NOCB), Sum(COH∗)
PF.POT1	Sum(NOCBPF∗), Sum(NOCB), Sum(COH∗), Sum(CHEAP∗), Sum(SAL∗)
PF.BFP	Sum(NOCBPF∗), Sum(CONTINUE), Sum(RETAIN), Sum(SMOOTH-SHIFT), Sum(ROUGH-SHIFT)

Table 8.1: Scoring functions of the modified metrics which compute the sum of NOCBPF∗s independently from other scores

in the same way as for their non-PF counterparts. The details of this operation have already been discussed in chapter 3.[7]

8.2 Experimental questions

The previous section motivated the PF-modification of the metrics as a way for distinguishing between examples (8.1) and (8.2). In the remainder of the chapter, the PF-modified metrics are engaged into a series of pairwise comparisons to estimate the effect of the PF-modification on the task performed by SEEC.[8]

First, we investigate whether computing the sum of NOCBPF∗s independently from NOCBs in a metric lowers the classification rate compared to its non-PF counterpart. A lower classification rate for a PF-modified metric compared to its non-PF counterpart shows that there are proportionally fewer permutations with a higher probability of being selected than the BfC when the PF-modified metric is used to guide the hypothetical text structuring algorithm in section 5.5.1 of chapter 5. In this case, the PF-modification increases the suitability of the metric for text structuring. We continue with the main experimental question, now reformulated as (Q2'):

Q2': Which PF-modified metrics of entity coherence constitute the most promising candidates for text structuring?

[7] Note that utterance (3.2b) in section 3.1.2 of chapter 3 is a NOCBPF∗. However, the only difference in the way that the PF-modified metrics evaluate the examples in chapter 3 is that PF.CHEAP prefers (3.1) over (3.2), whereas M.CHEAP renders them equivalent.

[8] The permutation strategy and the search strategy in the set of experiments reported in this chapter were the same as the ones specified in chapter 6 and chapter 7 for the BfCs from GNOME-LAB and MPIRO-PROP respectively. The only difference is that instead of using the metrics in Table 3.11 of chapter 3 as the input to SEEC we used their PF-modifications in Table 8.1 above.

PF-modified				non-PF	
Metric	Better	Equal	Y	Metric	Y
PF.KP	6.46	12.40	12.66	M.BFP	19.91
PF.BFP	7.22	11.48	12.96	M.MIL	20.06
PF.SH	6.74	13.85	13.66	M.NOCB	20.42
PF.POT1	8.38	11.46	14.11	M.SHOT1	22.79
PF.MIL	7.90	12.49	14.15	M.POT1	25.29
PF.SHOT1	8.38	11.71	14.24	M.SH	34.18
PF.NOCB	7.96	13.05	14.48	M.KP	53.15
PF.CHEAP	2.73	26.33	15.90	M.CHEAP	81.04
N	122				

Table 8.2: Average classification rate (Y=Better+Equal/2) for PF-modified and non-PF metrics in MPIRO-PROP

This question is answered by comparing the performance of the new baseline PF.NOCB with the rest of the metrics in Table 8.1. Like the previous chapter, the results obtained from GNOME-LAB are compared to the findings from MPIRO-PROP to identify which of characteristics of the genre of interest apply to a real application such as MPIRO.

8.3 Results from MPIRO-PROP

8.3.1 Average classification rate

Table 8.2 shows the average classification rate of the PF-modified metrics sorted in ascending order as well as the percentages of Better and Equal permutations and the number of BfCs (N) in MPIRO-PROP. The last two columns of the Table are a reminder of the Y values of the non-PF counterparts of the metrics (already reported in the fourth column of Table 7.1 in the previous chapter).

A couple of interesting preliminary observations can be made with respect to Table 8.2. First, Y of every metric is lower in PF-modified than in non-PF. In this context, Y(PF.KP, MPIRO-PROP) with 12.66%, much lower than the 53.15% of Y(M.KP, MPIRO-PROP), appears as the lowest Y among the PF-modified metrics.

Second, even though the classification rate of the baseline Y(PF.NOCB, MPIRO-PROP) is lower than Y(M.NOCB, MPIRO-PROP) - 14.48% versus 20.42% respectively, it turns out to be the second greatest Y value, which indicates a number of metrics with the potential of overtaking the baseline in the pairwise comparisons.

Pair			non-PF			p	Sig	Winner
			lower	greater	ties			
M.NOCB	vs	PF.NOCB	3	55	64	0.000	***	PF.NOCB
M.SH	vs	PF.SH	14	105	3	0.000	***	PF.SH
M.SHOT1	vs	PF.SHOT1	3	55	64	0.000	***	PF.SHOT1
M.POT1	vs	PF.POT1	3	55	64	0.000	***	PF.POT1
M.BFP	vs	PF.BFP	2	53	67	0.000	***	PF.BFP
M.MIL	vs	PF.MIL	3	55	64	0.000	***	PF.MIL
M.KP	vs	PF.KP	15	105	2	0.000	***	PF.KP
M.CHEAP	vs	PF.CHEAP	15	105	2	0.000	***	PF.CHEAP
N			122					

Table 8.3: Details of comparisons of each PF-modified metric with its non-PF counterpart in MPIRO-PROP

In the next two subsections, we answer the experimental questions defined in the previous section by employing the signtest to identify significant differences between the number of BfCs that contribute to each average score in MPIRO-PROP.

8.3.2 Differences between PF-modified and non-PF metrics

Table 8.3 shows the details of the comparison of each PF-modified metric with its non-PF counterpart in MPIRO-PROP.[9] As the Table shows, in all cases the non-PF counterpart has a greater classification rate in significantly more BfCs than its PF-modification. Thus, the winner of the comparison is always the PF-modified metric. This in turn means that considering NOCBPF∗s as different from NOCBs helps the metrics improve their performance in MPIRO-PROP. In terms of the generation scenario in section 5.5.1 of chapter 5, using any PF-modified metric instead of its non-PF counterpart to structure a semantic content defined according to MPIRO-PROP is expected to increase the chances of the BfC being selected over its alternatives as the output of the hypothetical algorithm.

8.3.3 Pairwise comparisons with PF.NOCB

After having established that a PF-modified metric returns a lower classification rate for significantly more BfCs than its non-PF counterpart in MPIRO-PROP, we turn our attention to (Q2') above. Table 8.4 shows the details of the comparison of the new baseline PF.NOCB with the rest of the PF-

[9] See section 6.5.2 of chapter 6 for clarifications on the notation used in the Tables.

Pair	PF.NOCB			p	Sig	Winner
	lower	greater	ties			
PF.NOCB vs PF.CHEAP	67	52	3	0.100	ns	
PF.NOCB vs PF.KP	12	27	83	0.013	*	PF.KP
PF.NOCB vs PF.BFP	9	23	90	0.011	*	PF.BFP
PF.NOCB vs PF.SH	13	26	83	0.027	*	PF.SH
PF.NOCB vs PF.MIL	0	6	116	0.016	*	PF.MIL
PF.NOCB vs PF.SHOT1	10	11	101	0.500	ns	
PF.NOCB vs PF.POT1	9	12	101	0.332	ns	
N	122					

Table 8.4: Details of pairwise comparisons with metric PF.NOCB in MPIRO-PROP

modified metrics in MPIRO-PROP.

By contrast to what is reported in the previous chapter for the non-PF metrics, computing NOCBPF∗s independently from NOCBs makes it possible for a number of PF-modified metrics (namely PF.KP, PF.BFP and PF.SH) in addition to PF.MIL to overtake the baseline. First, it is worth noting that the result of the pairwise comparison of PF.NOCB versus PF.MIL is due to the same 6 BfCs which render M.MIL winner over M.NOCB (see Table 7.3 and section 7.4.3 of the previous chapter for more details). Hence, PF.MIL retains the marginal, possibly domain-specific, superiority of its non-PF counterpart over the baseline irrespective of the PF-modification.

Although M.BFP is unable to overtake the baseline (again see Table 7.3 in section 7.4.2 of the previous chapter), its PF-modification manages to do so. However, the metrics that benefit most spectacularly from the PF-modification are PF.KP and PF.SH: Although their non-PF counterparts are badly beaten by M.NOCB, the PF-modifications of these metrics, perhaps somehow surprisingly, manage to overtake the new baseline PF.NOCB in the same way as PF.MIL and PF.BFP.

In accordance to the methodological point in section 5.5.3 of chapter 5, the metrics which overtake the baseline are compared with each other to identify additional differences. The details of these comparisons are displayed in Table 8.5. While most differences are not significant, PF.SH is overtaken by PF.KP and PF.BFP. This shows that although PF.SH beats the baseline, it is quite often doing worse than two of its competitors, which makes it a less favourable candidate solution than them. Hence, the comparisons in Table 8.5 reveal a dispreference for M.SH but no significant preference for any other PF-modified metric. Consequently, PF.KP, PF.BFP and PF.MIL are identified, at this point, as the most promising candidates for text structuring in MPIRO-PROP.[10]

[10] Arguably, PF.MIL could be seen as the simplest metric of the three. Note that the 6 BfCs that differentiate PF.MIL from

MPIRO-PROP						
Pair	PF.SH			p	Sig	Winner
	lower	greater	ties			
PF.SH vs PF.KP	1	20	101	0.000	***	PF.KP
PF.SH vs PF.BFP	6	28	88	0.000	***	PF.BFP

Pair	PF.MIL			p	Sig	Winner
	lower	greater	ties			
PF.MIL vs PF.KP	13	21	88	0.115	ns	
PF.MIL vs PF.BFP	10	17	95	0.124	ns	
PF.MIL vs PF.SH	16	20	86	0.309	ns	

Pair	PF.KP			p	Sig	Winner
	lower	greater	ties			
PF.KP vs PF.BFP	18	11	93	0.133	ns	
N	122					

Table 8.5: Details of comparisons of PF-modified metrics overtaking the baseline with each other

Finite				Finite-RR			
Metric	Better	Equal	Y	Metric	Better	Equal	Y
PF.MIL	14.00	15.78	21.89	PF.NOCB	16.19	21.07	26.73
PF.NOCB	13.77	16.70	22.13	PF.MIL	20.89	15.91	28.85
PF.POT1	18.06	9.20	22.66	PF.SHOT1	23.40	12.76	29.78
PF.SHOT1	15.77	13.89	22.72	PF.SH	23.44	13.27	30.07
PF.SH	15.88	14.25	23.01	PF.POT1	28.41	5.25	31.04
PF.BFP	18.38	11.40	24.08	PF.BFP	28.12	6.29	31.26
PF.KP	19.58	14.95	27.05	PF.CHEAP	23.93	15.68	31.76
PF.CHEAP	22.46	14.15	29.54	PF.KP	25.90	12.92	32.36
N	20			N	12		

Table 8.6: Average classification rate (Y=Better+Equal/2) of PF-modified metrics in GNOME-LAB

8.4 Results from GNOME-LAB

In this section we investigate the performance of the PF-modified metrics in GNOME-LAB and compare it with MPIRO-PROP. We start by reporting the average classification rate of each metric with and without the PF-modification. Then, we present the results of our statistical analysis.

8.4.1 Average classification rate

Table 8.6 shows the Y values as well as the percentage of Better and Equal permutations for the PF-modified metrics and the number of BfCs in the Finite and Finite-RR versions of GNOME-LAB, whilst Table 8.7 repeats the Y values of the non-PF metrics from Table 6.1 in chapter 6. As always, the metrics are ranked in ascending order according to Y.

The situation in GNOME-LAB seems to be rather different from MPIRO-PROP. First, Y(PF.NOCB, Finite) and Y(PF.NOCB, Finite-RR) appear quite high up in Table 8.6. Moreover, as Table 8.7 shows, M.NOCB returns lower values than all PF-modified metrics in both BfC versions. In the remainder of the chapter, these tendencies are examined more closely and compared with the results from MPIRO-PROP.

Finite				Finite-RR			
PF-modified		non-PF		PF-modified		non-PF	
Metric	Y	Metric	Y	Metric	Y	Metric	Y
PF.MIL	21.89	M.NOCB	19.95	PF.NOCB	26.73	M.NOCB	23.24
PF.NOCB	22.13	M.POT1	20.34	PF.MIL	28.85	M.SHOT1	26.15
PF.POT1	22.66	M.SHOT1	20.77	PF.SHOT1	29.78	M.POT1	26.81
PF.SHOT1	22.72	M.MIL	23.47	PF.SH	30.07	M.MIL	27.91
PF.SH	23.01	M.SH	29.05	PF.POT1	31.04	M.SH	30.66
PF.BFP	24.08	M.BFP	33.01	PF.BFP	31.26	M.BFP	33.39
PF.KP	27.05	M.CHEAP	57.23	PF.CHEAP	31.76	M.KP	56.87
PF.CHEAP	29.54	M.KP	58.22	PF.KP	32.36	M.CHEAP	62.10
N	20			N	12		

Table 8.7: Average classification rate (Y) of PF-modified and non-PF metrics in GNOME-LAB

Pair			non-PF			p	Sig	Winner
			lower	greater	ties			
M.NOCB	vs	PF.NOCB	9	2	9	0.033	*	M.NOCB
M.SH	vs	PF.SH	8	6	6	0.395	ns	
M.SHOT1	vs	PF.SHOT1	9	2	9	0.033	*	M.SHOT1
M.POT1	vs	PF.POT1	8	2	10	0.055	ns	
M.BFP	vs	PF.BFP	7	8	5	0.500	ns	
M.MIL	vs	PF.MIL	9	5	6	0.212	ns	
M.KP	vs	PF.KP	4	14	2	0.015	*	PF.KP
M.CHEAP	vs	PF.CHEAP	4	16	0	0.006	**	PF.CHEAP
N			20					

Table 8.8: Details of comparisons of each PF-modified metric with its non-PF counterpart in Finite

Pair			non-PF			p	Sig	Winner
			lower	greater	ties			
M.NOCB	vs	PF.NOCB	6	0	6	0.016	*	M.NOCB
M.SH	vs	PF.SH	5	5	2	0.500	ns	
M.SHOT1	vs	PF.SHOT1	6	0	6	0.016	*	M.SHOT1
M.POT1	vs	PF.POT1	6	0	6	0.016	*	M.POT1
M.BFP	vs	PF.BFP	3	6	3	0.254	ns	
M.MIL	vs	PF.MIL	6	4	2	0.377	ns	
M.KP	vs	PF.KP	3	9	0	0.073	ns	
M.CHEAP	vs	PF.CHEAP	3	9	0	0.073	ns	
N			12					

Table 8.9: Details of comparisons of each PF-modified metric with its non-PF counterpart in Finite-RR

8.4.2 Differences between PF-modified and non-PF metrics

Table 8.8 and Table 8.9 show the details of the comparisons of each PF-modified metric with its non-PF counterpart in the two ways of computing the BfC in GNOME-LAB. Table 8.10 compares the results from GNOME-LAB with the results from MPIRO-PROP.

As Table 8.10 shows, whereas a PF-modified metric always beats its non-PF counterpart in MPIRO-PROP, M.NOCB, M.SHOT1 and M.POT1 have a lower classification rate for significantly more BfCs than their PF-modifications in both BfC versions in GNOME-LAB.[11] This means that considering NOCBPF∗s as different from NOCBs lowers the performance of these metrics in GNOME-LAB, instead of improving it as in MPIRO-PROP. Note that a number of comparisons in GNOME-LAB do not reach significance.

Even though failing to reach significance in some pairwise comparisons might be due the small sample in GNOME-LAB, the fact that M.NOCB beats PF.NOCB in both ways of representing the BfCs is enough to show that the effect of NOCBPF∗s on the performance of the metrics in GNOME-LAB is not the same as in MPIRO-PROP. This result also shows that M.NOCB remains very robust, and should be retained as the baseline in subsequent comparisons with the PF-modified metrics.

The only PF-modified metrics in GNOME-LAB which have a lower classification rate for signif-

PF.NOCB are distributed in such a way that the comparisons of PF.MIL with the other three metrics that overtake PF.NOCB fail to reach significance. However, the tendency against PF.NOCB appears to be inherited to the comparisons with PF.MIL, especially as far as PF.KP and PF.BFP are concerned.

[11] Although the result of the comparison M.POT1 vs PF.POT1 in Finite is not significant (see Table 8.8), it is very close to p=0.05 and in the same direction as the corresponding comparison in Finite-RR which does achieve significance.

Pair			Winner		
			GNOME-LAB		MPIRO-PROP
			Finite	Finite-RR	
M.NOCB	vs	PF.NOCB	M.NOCB	M.NOCB	PF.NOCB
M.SH	vs	PF.SH	ns	ns	PF.SH
M.SHOT1	vs	PF.SHOT1	M.SHOT1	M.SHOT1	PF.SHOT1
M.POT1	vs	PF.POT1	ns	M.POT1	PF.POT1
M.BFP	vs	PF.BFP	ns	ns	PF.BFP
M.MIL	vs	PF.MIL	ns	ns	PF.MIL
M.KP	vs	PF.KP	PF.KP	ns	PF.KP
M.CHEAP	vs	PF.CHEAP	PF.CHEAP	ns	PF.CHEAP

Table 8.10: Winners of comparisons of each PF-modified metric with its non-PF counterpart in GNOME-LAB and MPIRO-PROP

icantly more BfCs than their non-PF versions are PF.KP and PF.CHEAP in Finite.[12] This is perhaps not very surprising given the very bad performance of M.KP and M.CHEAP in GNOME-LAB, as already reported in section 6.5.2 of chapter 6.

The question which now arises is whether PF.KP and PF.CHEAP, the only two metrics that benefit from the PF-modification for significantly more BfCs than their non-PF counterparts in GNOME-LAB, can overtake the retained baseline M.NOCB. Other potential PF-modified competitors of M.NOCB are PF.SH, PF.BFP and PF.MIL which do not appear to benefit from the PF-modification in GNOME-LAB as much as PF.KP and PF.CHEAP, but might still be able to do better than their non-PF counterparts in the pairwise comparison with M.NOCB.[13] These questions are addressed in the next section.

8.4.3 Pairwise comparisons with M.NOCB

So far we have shown that computing NOCBPF∗s independently from NOCBs does not boost the performance of the metrics in GNOME-LAB as clearly as it does in MPIRO-PROP. Moreover, M.NOCB remains a very robust baseline in GNOME-LAB, although some PF-modified metrics (namely PF.CHEAP, PF.KP, PF.BFP, PF.SH, PF.MIL) might still challenge it. Since these metrics, except for PF.CHEAP, were found to overtake the new baseline PF.NOCB in MPIRO-PROP (see Table 8.4

[12] A difference in the same direction, failing to reach significance, is also observed in Finite-RR (see Table 8.9).

[13] PF.SHOT1 and PF.POT1 are not regarded as "potential PF-modified competitors" of M.NOCB because as Table 8.10 shows, they are beaten by their non-PF counterpart (which in turn fails to overtake M.NOCB as reported in section 6.5.2 of chapter 6).

Finite						
Pair	M.NOCB			p	Sig	Winner
	lower	greater	ties			
M.NOCB vs PF.CHEAP	11	6	3	0.116	ns	
M.NOCB vs PF.KP	13	3	4	0.017	*	M.NOCB
M.NOCB vs PF.BFP	11	3	6	0.029	*	M.NOCB
M.NOCB vs PF.SH	11	3	6	0.029	*	M.NOCB
M.NOCB vs PF.MIL	9	5	6	0.212	ns	
N	20					
Finite-RR						
Pair	M.NOCB			p	Sig	Winner
	lower	greater	ties			
M.NOCB vs PF.CHEAP	9	3	0	0.073	ns	
M.NOCB vs PF.KP	10	1	1	0.006	**	M.NOCB
M.NOCB vs PF.BFP	9	3	0	0.073	ns	
M.NOCB vs PF.SH	8	2	2	0.055	ns	
M.NOCB vs PF.MIL	7	3	2	0.172	ns	
N	12					

Table 8.11: Details of comparisons of metric M.NOCB with potential PF-modified competitors in GNOME-LAB

in section 8.3.3) we are interested to see whether GNOME-LAB returns similar results.

Table 8.11 shows the details of the pairwise comparisons of M.NOCB, the retained baseline in GNOME-LAB, with the metrics that were recognised as its potential PF-modified competitors at the end of the previous section. Table 8.12 compares the winners of the pairwise comparisons with M.NOCB in Table 8.11 with the winners of its corresponding comparisons with the non-PF metrics from section 6.5.2 of chapter 6.

As the Tables show, the PF-modification in GNOME-LAB is not enough for the PF-modified metrics to overtake M.NOCB. Clearly, this does not abide by the findings from MPIRO-PROP where a number of PF-modified metrics overtake the baseline (albeit in a larger sample). Crucially, PF.KP, one of the metrics which beats the baseline in MPIRO-PROP, is overtaken by M.NOCB even though it benefits from the PF-modification in GNOME-LAB. Note that all pairwise comparisons which reach significance are in favour of M.NOCB while the remaining comparisons are in the same direction as well. Keeping the limitations imposed by the small sample sizes in mind, these results indicate that the

GNOME-LAB					
PF-modified	Winner		non-PF	Winner	
	Finite	Finite-RR		Finite	Finite-RR
PF.CHEAP	ns	ns	M.CHEAP	M.NOCB	M.NOCB
PF.KP	M.NOCB	M.NOCB	M.KP	M.NOCB	M.NOCB
PF.BFP	M.NOCB	ns	M.BFP	M.NOCB	ns
PF.SH	M.NOCB	ns	M.SH	ns	ns
PF.MIL	ns	ns	M.MIL	ns	ns

Table 8.12: Winners of pairwise comparisons of metric M.NOCB with some PF-modified metrics and their non-PF counterparts

PF-modified metrics are as unable to beat the baseline as their non-PF counterparts in GNOME-LAB, contrary to what happens in MPIRO-LAB.

8.5 Summary and discussion

In summary, this chapter introduces a modification of the metrics in chapter 3 by computing the sum of NOCBPF∗s separately from NOCBs. This yields a significant improvement in the performance of the metrics in MPIRO-PROP, but not in GNOME-LAB. Thus, the PF constraint on entity coherence does not appear to characterise the whole of the investigated genre as represented by GNOME-LAB, but is useful for, albeit specific to, structuring the data from MPIRO-PROP.

In terms of the generation scenario that the metrics are tested to be suitable for, the chances of the BfC to be the output of text structuring are increased when data similar to MPIRO-PROP serve as the input and certain PF-modified metrics guide the hypothetical algorithm. However, if the input is replaced with a semantic content similar to what comes from GNOME-LAB, then the chances of the BfC to be selected are better when the algorithm is driven by M.NOCB instead of the PF-modified metrics. Hence, although a number of promising candidates for text structuring have emerged in the particular application domain, M.NOCB remains very robust as far as the genre of interest is concerned.

Consequently, the case still is that even though M.NOCB appears to be the most promising candidate among its alternatives in the investigated genre, on average around 20% of the permutations are more likely to be selected than the BfC (see Table 8.7). Clearly, M.NOCB needs to be supplemented by other features to improve its performance, although this cannot be achieved with the help of the PF constraint or the computation of RRs in the BfC (see sections 6.5.3 and 6.5.4 of chapter 6 for more

discussion). Hence, although M.NOCB is good starting point to investigate the effect of entity coherence in the genre of interest, the factors that can supplement it to build a more efficient text structuring metric remain unclear to us.

The PF-modification of the metrics resulted from the inspection of the best scoring permutations for the best performing metrics in chapter 7. As far as MPIRO-PROP is concerned, this modification accounts for differences between the BfC and some of the best scoring permutations that previously remained undetected under the definition of the metrics in chapter 3. Note, however, that the experimental methodology used to evaluate the PF-modified metrics in this chapter is the same as before. Thus, the BfC remains the only point of reference and the methodology continues to be agnostic about the felicity of other permutations that might also be favoured by the new best performing metrics. Hence, an even closer look at these permutations is in order before drawing the final conclusions about the suitability of the best PF-modified metrics for text structuring.

Another pending question is whether the results from MPIRO-PROP reflect an ordering strategy solely followed by EM (the expert who ordered the data derived from MPIRO) or a more general strategy for ordering the data from the particular application domain. In order to answer this question, orderings from more than one expert on the same type of data need to be acquired and compared to the orderings of EM.

In the next chapter, we present the final evaluation experiment which first attempts to shed some more light on the last issue by computing the average distance between the orderings of EM and the orderings of other experts on additional data from MPIRO-PROP. Then, the same performance measure used to investigate the differences between the experts is used to evaluate the best performing metrics and the two baselines employed so far in terms of the average distance of their best scoring permutations from the orderings of the experts. This distance is compared to the distances between the orderings of the experts and a) each other and b) a random baseline, which serve as the upper and the lower bound of the evaluation respectively. This study supplements the experimental methodology of chapter 5 and concludes our experimental efforts in the MPIRO domain.

Chapter 9

Using data from more than one expert

A question not addressed so far is whether the results from MPIRO-PROP are specific to EM (the expert who ordered the information derived from MPIRO) or whether they reflect more general strategies for ordering the application-specific data. In order to answer this question in a general way in this chapter, the dataset from MPIRO-PROP is enhanced with orderings provided by more than one expert. Then, the distance between EM and her colleagues is computed and compared to the distance between her colleagues and each other. The results indicate that EM shares a lot of common ground with two of her colleagues in the ordering task, while another "stand-alone" expert who uses strategies not shared by the rest of the experts is identified as well.

The same methodology used to investigate the distance between the experts is used to automatically evaluate the best scoring orderings of some of the the best performing metrics in chapter 8, as well as the two previously employed baselines. More specifically, distances are computed and compared between the orderings of the experts and a) each other, b) the orderings of a random baseline, c) the best scoring orderings of some of the best performing metrics so far and d) the best scoring orderings of the two previously employed baselines. This attempts to account for a number of possible deficiencies of the main methodology employed in the experiments previously reported in the thesis.

The results provide additional evidence in favour of the PF constraint on entity coherence which is shown not to be specific to EM but shared by her colleagues to a great extent. They also indicate that the distance between the orderings of the experts and the best scoring orderings of each metric is significantly smaller than the distance between the orderings of the experts and the orderings of the random baseline. Hence, all metrics are superior to the random baseline.

However, only one PF-modified metric manages to return a distance from the experts which is not significantly different from the distance of the orderings between the experts and each other. Hence, this metric is identified as the one that performs best across all evaluation tasks and can be rendered as the most promising candidate for text structuring in MPIRO-PROP among the ones investigated.

9.1 Motivation and aims

The data, aims and requirements of this study differ from the ones reported in previous chapters of the thesis. To begin with, one open question behind the results from MPIRO-PROP is whether they should be attributed solely to EM or whether they express more general strategies for ordering the data from the particular application. In an attempt to answer this question, more than one expert is presented with additional sets of facts, similar to the ones that the data in MPIRO-PROP consist of, and asked to provide us with an ordering for each set. These orderings are then compared with the orders of EM in a general way using the methodology of Lapata (2003) (see section 9.4 for more details).

More than one expert may give rise to **more than one** corpus instance for each set of facts. Crucially, the "parallel corpus" collected for this study can prove useful for investigating the performance of the metrics even further. Although the experimental methodology of chapter 5 can be extended to account for the data collected in this study, this will still isolate each BfC as the **only** permutation of interest. As we mentioned repeatedly in the previous two chapters, because this methodology considers only the position of the BfC in the search space, it remains agnostic about the felicity of other permutations that might score Equal to or Better than the BfC according to the metric which is assumed to drive the text structuring process. As the metrics are compared on their ability to single out the BfC as the most desirable output, the possibility that there might actually exist additional equally good solutions is also ignored.

In order to account for these possible deficiencies, we investigate **all** best scoring permutations for the best performing metrics so far even more closely than in the previous chapter and compare them with the orderings of the multiple experts.[1] Interestingly, the same dependent variable which allows us to investigate how different the orders of EM are from the orders of her colleagues can also be used to evaluate the metrics of interest using the complete set of best scoring permutations. This not only reveals interesting points about the best scoring orderings for each metric, but also serves as an additional evaluation test in the sense that will be made clearer in section 9.5.

The dependent variable employed in this study is the *distance between two orderings*, defined more precisely in section 9.4. Because consulting more than one expert is a time-consuming procedure, it can easily be done on a small scale but is more difficult to extend to the complete dataset in MPIRO-PROP. For this reason, the study takes place on a smaller set of data and is restricted to fewer metrics than in the last two chapters. The metrics employed in this chapter are PF.BFP, PF.KP and PF.MIL (that is, the three metrics that outperform the PF-modified baseline but do not differ significantly from each other)[2] as well as the two previously used baselines PF.NOCB and M.NOCB. The

[1] Note that, as stated in section 8.1.1 of the previous chapter, these permutations might be a subset of the permutations used to calculate the classification rate.

[2] PF.SH is not taken into account as it is beaten by PF.BFP and PF.KP (see Table 8.5 in section 8.3.3 of chapter 8).

study is designed to explore the specific issues explained in the previous paragraphs, although it can also be seen as specifying a different, more general, evaluation methodology which is valid to apply to all metrics, provided that a larger dataset of multiple orderings is created.

Hence, the primary aim of the main experiment, the results of which are reported in section 9.6, is to estimate the average distance between (the orderings of) the human experts[3] and:

a) each other:

   This information is first used to investigate the difference between the orderings of EM and her colleagues. Then, it used as the upper bound in the analysis of the performance of the metrics.

b) the orderings that are assigned the best scores by the metrics in question:

   By investigating the difference between a) and b) one can estimate how the metrics perform in comparison to the upper bound.

c) a random baseline:

   This information is used as the lower bound in the subsidiary evaluation of the metrics.

9.1.1 Secondary experiment

In addition to the main experiment, the first score, i.e. the distance between the experts and each other, is used in a secondary experiment which investigates the difference between AllEq and MPIRO-PROP, the two datasets identified in section 7.3 of chapter 7. As we already mentioned in that section, the orderings in MPIRO-PROP are awarded with different scores for their entity coherence as estimated by the metrics. By contrast, in AllEq all orderings are completely equivalent with respect to their entity coherence and are not meaningful for our purposes.

Assuming that entity coherence is important for the experts in their search for a good ordering, it is to be expected that the distance between their outputs in the Testitems from MPIRO-PROP will be smaller than the distance in AllEq, since the space of the most entity coherent solutions in AllEq is much wider than in MPIRO-PROP. Hence, the aim of the secondary experiment is to specify whether the distance between the experts and each other computed on data from AllEq is indeed larger than the distance between the experts and each other computed on data from MPIRO-PROP. The results of this study are reported in section 9.7.

[3]Throughout the chapter we often refer to e.g. “the distance between the orderings of the experts” with the phrase “the distance between the experts”, etc. for the sake of brevity.

9.2 Gathering additional data

In addition to the data used in the studies reported in chapter 7 and chapter 8, 16 sets of facts were randomly selected from the dataset of Dimitromanolaki and Androutsopoulos (2003). We will subsequently refer to each unordered set of facts as a *Testitem*. As mentioned in section 7.2 of chapter 7, the facts that each Testitem consists of are ordered according to the instructions of an expert working for the MPIRO project (EM). Being appropriate for the task performed by SEEC, the 16 orderings of EM are similar to the 122 corpus instances in MPIRO-PROP. These data are supplemented with 6 additional randomly selected orderings belonging to AllEq which were used in our secondary experiment.

Following the procedure in section 7.3 of chapter 7, each Testitem gave rise to a set of CF lists (semantic content), which was then used as the input to the algorithm in section 8.1.1 of chapter 8. Using a metric M and each semantic content as its inputs, the algorithm outputs the permutations that score best according to M (BestTable).

As specified in the previous section, the members of the BestTable (BestOrders) were generated for the semantic content of each Testitem using PF.BFP, PF.KP and PF.MIL as well as the two baselines PF.NOCB and M.NOCB as subsequent inputs to the algorithm.

9.2.1 Examining the BestOrders

An examination of the BestOrders generated by the metrics showed that for quite a few Testitems, the metrics output the same set of BestOrders. More specifically, PF.MIL and PF.NOCB output identical BestTables for all 16 Testitems. This is not very surprising given that, as we saw in section 8.3.3 of the previous chapter, the difference between PF.MIL and PF.NOCB is due to a specific type of construction with relatively low frequency in MPIRO-PROP (6/122). For this reason, the dataset of this study cannot be used to investigate the difference between PF.MIL and PF.NOCB in more detail than already done in section 7.4.3 of chapter 7.

Further to this, the remaining three PF-modified metrics (PF.BFP, PF.KP and PF.NOCB) output the same BestTable as M.NOCB for 6 Testitems. In one additional case, the BestOrders are the same for the three PF-modified metrics, but distinct from the BestOrders of M.NOCB. Hence, there are 7 Testitems in total for which the three PF-modified metrics output identical BestTables.

The differences in the BestOrders of the three PF-modified metrics for the remaining 9 Testitems are summarised in Figure 9.1. As the Figure shows, there are 4 Testitems for which PF.KP outputs 2 more BestOrders, denoted as OR_{n+1} and OR_{n+2}, in addition to the BestOrders of PF.BFP and PF.NOCB, which are identical with each other and with the remaining BestOrders in the BestTable of PF.KP.[4] Finally, for the remaining 5 Testitems, PF.NOCB outputs two additional BestOrders, denoted

[4]Note that n might differ from one Testitem to another.

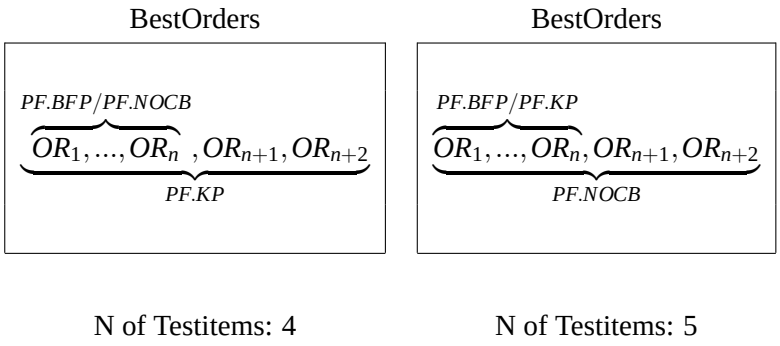

Figure 9.1: Differences in BestOrders for PF.NOCB, PF.BFP and PF.KP

as OR_{n+1} and OR_{n+2}, to the identical BestTables of PF.BFP and PF.KP.

9.2.2 Implementing a random baseline

A random baseline was implemented as the lower bound of the analysis. The random baseline (RB) consists of 10 randomly selected permutations for each Testitem. The permutations are selected irrespective of their scores for the various metrics.

As mentioned in section 9.1, the upper bound of the evaluation is defined by orderings provided by the experts, in our case archaeologists trained in museum labelling. Our methods for preparing the data which were used to consult the experts are described in the next section.

9.2.3 Realising facts as sentences

In order to consult experts who, unlike EM, are not familiar with MPIRO's underlying representation, the facts in each Testitem were realised as sentences with the help of EXPRIMO, MPIRO's generation engine.[5] More specifically, first EXPRIMO was invoked to generate descriptions of the PF in each set of facts.[6] For instance, the PF of the set of facts in example (9.1) is `exhibit36`:

(9.1)
- subclass(exhibit36, kylix)
- creation-period(exhibit36, classical-period)
- period-story(classical-period, entity-8555)
- creation-time(exhibit36, date-4956)
- original-location(exhibit36, attica)

[5] The same methodology was used for the realisation of the sentences which correspond to MPIRO's facts in the examples used throughout the thesis.

[6] We remind the reader that the PF is the artefact to be described, which initiates the generation process. See section 8.1.2 of the previous chapter for more details.

- current-location(exhibit36, museum-of-art-toledo)

Asking EXPRIMO to generate a description for `exhibit36` gives rise to the following text:

> This exhibit is a kylix. It was created during the classical period. The classical period ranges from 480 to 323 BC. It was defined by the rise in the political supremacy of Athens (its "golden age") and the expansion of the Greek world under the rule of Alexander the Great of Macedonia. This kylix dates from circa 480 BC. It originates from Attica and it is currently displayed in the Museum of Art of Toledo.

Then, the author and another computational linguist experienced in MPIRO identified which clause in the generated description corresponds to each fact.[7] The corresponding clauses in this example are:

> *subclass*:[This exhibit is a kylix.] *creation-period*:[It was created during the classical period.] *period-story*:[The classical period ranges from 480 to 323 BC.] ... *creation-time*:[This kylix dates from circa 480 BC.] *original-location*:[It originates from Attica] and *current-location*:[it is currently displayed in the Museum of Art of Toledo.]

After the clauses had been identified, all pronouns were replaced with full noun phrases in order to be presented to the experts out of context. This gives rise to the following set of sentences realising the facts in (9.1):[8]

(9.2)

- subclass(exhibit36, kylix):

  This exhibit is a kylix.

- creation-period(exhibit36, classical-period):

  This exhibit was created during the classical period.

- period-story(classical-period, entity-8555):

  The classical period ranges from 480 to 323 BC.

- creation-time(exhibit36, date-4956):

  This exhibit dates from circa 480 BC.

- original-location(exhibit36, attica):

  This exhibit originates from Attica.

- current-location(exhibit36, museum-of-art-toledo):

  This exhibit is currently displayed in the Museum of Art of Toledo.

[7] When the clauses in a generated description did not realise all facts in a Testitem, then EXPRIMO was invoked again to generate a new description for the same PF so that all facts were eventually realised by a clause in at least one description. I am grateful to Aggeliki Dimitromanolaki for providing me with the Testitems and EM's corpus instances used in this study, invoking EXPRIMO and helping me identify which sentence realises each fact.

[8] Because some facts are occasionally realised by canned text which might span more than one sentence and can be more sophisticated than the output of deep generation, some of the realisations were simplified so that they look similar to the deep generated ones. However, we tried to keep changes to a necessary minimum, in order not to deviate substantially from the surface output of EXPRIMO.

9.3 Interviews with experts

Three experts (E1, E2, E3), one male and two females, between 28 and 45 years of age, all trained in cataloguing and museum labelling, were recruited from the Department of Classics at the University of Edinburgh.[9] Each expert was consulted by the author in a separate interview. First, she was presented with a set of instructions describing the ordering task. The full text of the instructions is given in appendix D and is adapted from the instructions used in Barzilay et al. (2002).[10]

The instructions mention that the expert will be presented with sets of six sentences which come from a computer program that generates descriptions of artefacts in a virtual museum.[11] The first sentence for each set will be given by the experimenter.[12] The task of the expert is to order the remaining five sentences in a coherent text.

When ordering the sentences, the expert was instructed to consider which ones should be together and which should come before another in the text without using hints other than the sentences themselves. She could revise her ordering at any time by moving the sentences around. When she was satisfied with the ordering she produced, she was asked to write next to each sentence its position, and give them to the experimenter in order to perform the same task with the next randomly selected set of sentences.

During the experiment, the expert was encouraged to share any comments or questions she might have with the author. The experiment was followed by an informal interview where the expert commented on the difficulty of the task, the strategies she had followed, etc. In the interview, all experts recognised the task as an interesting and familiar problem which they undertook with enthusiasm.

9.4 Dependent variable

Given an unordered set of sentences and two possible orderings OR1 and OR2, a number of measures can be employed to calculate the distance between the orderings, such as Spearman's r_s, Kendall's τ, etc. Howell (2002, p.309) argues that Kendall's τ has a more straightforward interpretation than Spearman's coefficient:

> If a pair of objects is sampled at random, the probability that two judges will rank these objects in the same order is τ times higher than the probability that they will rank them in the reverse order.

[9] Many thanks to Katerina Kolotourou for her invaluable assistance in recruiting the experts.

[10] These instructions are available online at http://www1.cs.columbia.edu/~noemie/ordering/experiments/

[11] Each sentence was printed on a different filecard. The filecards were presented to the experts in sets according to the design of the experiment. From the sets of sentences presented to the experts, 16 correspond to the Testitems from MPIRO-PROP and 6 to sets of facts from AllEq.

[12] This is the sentence corresponding to the fact with the "subclass" predicate.

	A	B	C	D	E	F
E1	1	2	3	4	5	6
E2	1	5	6	2	3	4

Table 9.1: Example of orders for sentences A to F from two experts E1 and E2

Lapata (2003) uses Kendall's τ in a series of experiments which evaluate the performance of a probabilistic text structuring model in comparison to orderings provided by human judges. Kendall's τ appears to be very appropriate for our purposes as well since this study is very similar to the one reported by Lapata (2003).[13]

Kendall's τ penalises inverse rankings and is sensitive to the fact that some sentences may be ordered next to each other even though their absolute orders might differ. Its calculation is based on the number of *inversions* between the two orderings and is defined in (9.3):

$$\tau = 1 - \frac{2I}{P_N} = 1 - \frac{2I}{N(N-1)/2} \tag{9.3}$$

where P_N stands for the number of pairs of sentences and N is the number of sentences to be ordered.[14] I stands for the number of inversions, that is, the number of adjacent transpositions necessary to bring one order to another. Assume, for example, that the orders produced by two experts E1 and E2 for a Testitem consisting of 6 sentences A, B, C, D, E, F are the ones displayed in Table 9.1.

The sentences in the second row of Table 9.1 are indexed from 1 to 6 according to the order given by E1 and lines are drawn to connect the sentences which are given different positions by E2. The number of inversions I can be calculated by counting the number of intersections between the lines. In this example there are six intersections, and therefore six inversions.

Kendall's τ ranges from -1 (inverse ranks) to 1 (identical ranks). The **higher** the τ value, the **smaller** the distance between the orderings of the two experts. The τ value in our example is 0.20, which indicates that the two orderings are not very close to each other.

9.4.1 Calculating significance

Each τ value can be associated with a z score which indicates whether the orderings of the two experts are significantly close to or away from each other. The formula for calculating the z score for a τ value from Howell (2002) is given in (9.4):

$$z = \tau / \sqrt{\frac{2(2N+5)}{9N(N-1)}} \tag{9.4}$$

[13] Special thanks to Maria Lapata for providing me with the scripts for the computation of τ and appropriate data formating together with her extensive and prompt advice on their use.

[14] In our case N is always equal to 6.

In our example, since τ is equal to 0.20, the associated *z* score is 0.56, which is not significant (p=0.288). This, in turn, means that in this example the orderings of the two experts are not significantly close to each other.[15]

Using the formula in (9.4), one can calculate that the absolute τ value of the orderings of two experts for a given Testitem that consists of 6 sentences has to be 0.696 or greater to achieve significance. In that case, the associated absolute *z* score is at least 1.96, which corresponds to a p value of 0.025, the threshold of significance for a two-tailed prediction.

Since calculating *z* is possible for every single Testitem and a given pair of experts, a useful way for summarising the performance of two experts across many Testitems is to report the percentage of Testitems for which the orders of the experts are significantly close to or away from each other. Indeed, in order to investigate the difference between the Testitems from MPIRO-PROP and the ones from AllEq in section 9.7 we use a descriptive statistic like this, among other means.

However, using the same approach to estimate the performance of a metric runs into two problems. First, since each metric outputs more than one ordering, we need to somehow reward it for an ordering that is significantly close to the expert's order, but penalise it for every ordering that is not. Even if one comes up with a good formula to express this tradeoff, it might still be difficult to say whether the difference with the upper bound, in this case the percentage of Testitems for which the orders of two experts are significantly close, is significant.

A more straightforward analysis which uses directly the complete set of τ values as the dependent variable comes from Lapata (2003). Instead of calculating the percentage of items which achieve a significant *z* score, Lapata (2003) estimates whether the average distance between the orderings of the human judges and each other is significantly different from the average distance between the orderings of the judges and the ones of her stochastic model. In what follows, we will show how the methodology of Lapata (2003) applies to our experimental set-up.

9.4.2 Computing the distance between the experts

As we have already mentioned, for a Testitem TEST1, which is ordered by e.g. EM and E1, τ can be used to compute τ(EM.OR1,E1.OR1,TEST1) as the distance measure between EM.OR1 and E1.OR1, i.e. the orderings of EM and E1 for TEST1:

	EM	E1		
TEST1	EM.OR1	E1.OR1	→	τ(EM.OR1,E1.OR1,TEST1)

First, we calculate a set of τ values for each Testitem in the dataset:

[15] Clearly, this is different from saying that the two orderings are significantly away from each other, which can only be shown when a negative τ value is associated with a significantly low *z* score.

	EM	E1		
TEST1	EM.OR1	E1.OR1	→	τ(EM.OR1,E1.OR1,TEST1)
...	...			...
TESTN	EM.ORN	E1.ORN	→	τ(EM.ORN,E1.ORN,TESTN)

The *average* τ, $T(EM_{E1})$, is the mean of the τ values, and can be used to express the average distance between the (orderings of) EM and the (orderings of) E1 in the complete dataset of cardinality N:[16]

$$T(EM_{E1}) = \frac{\tau(EM.OR1,E1.OR1,TEST1)+...+\tau(EM.ORN,E1.ORN,TESTN)}{N} \tag{9.5}$$

Assume now that two additional experts E2 and E3 give the following set of τ and T values for the distance of their orderings from EM:

	EM	E2	E3			
TEST1	EM.OR1	E2.OR1	E3.OR1	→	τ(EM.OR1,E2.OR1,TEST1)	τ(EM.OR1,E3.OR1,TEST1)
...		...			...	
TESTN	EM.ORN	E2.ORN	E3.ORN	→	τ(EM.ORN,E2.ORN,TESTN)	τ(EM.ORN,E3.ORN,TESTN)
Average:					$T(EM_{E2})$	$T(EM_{E3})$

As in Lapata (2003), significant differences between the average T scores can be investigated with the help of the Tukey test. Provided that an omnibus ANOVA has revealed a significant main effect of the factor DISTANCE, the Tukey test can be used to specify which of the conditions $d_1,...,d_n$ that DISTANCE consists of differ significantly. It uses the set of means $m_1,...,m_n$ (corresponding to conditions $d_1,...,d_n$) and the mean square error of the scores that contribute to these means to calculate a critical difference between any two means. An observed difference between any two means is significant if it exceeds the critical difference.

In our example, the test performs all possible pairwise comparisons between the three means, $T(EM_{E1})$, $T(EM_{E2})$ and $T(EM_{E3})$, to specify whether the average distance between EM and another expert e.g. E1 is significantly different from the average distance between EM and either of her other two colleagues, in this case either E2 or E3.

Since all judges involved in the study are ab initio taken to be of equivalent expertise, in the actual analysis we compute all possible T scores to estimate the distance between each expert and each of her colleagues. That is, we do not simply compute $T(EM_{E1})$, $T(EM_{E2})$ and $T(EM_{E3})$ as above, but also $T(E1_{E2})$, $T(E1_{E3})$ and $T(E2_{E3})$. The 6 T scores give rise to 15 pairwise comparisons, with the Tukey test used to specify which differences are significant.

[16] As already pointed out, we often refer to "the distance between the orderings of EM and the orderings of E1" as "the distance between EM and E1", etc.

9.4.3 Computing the distance between the experts and a metric

Assume now that a metric M outputs K BestOrders, M.OR1.1 ... M.OR1.K, as the members of the BestTable of TEST1. A τ value can be computed for the distance between the order provided by EM and each BestOrder:

	EM	M		
TEST1	EM.OR1	M.OR1.1	→	τ(EM.OR1,M.OR1.1,TEST1)
		...		...
		M.OR1.K	→	τ(EM.OR1,M.OR1.K,TEST1)

The average distance between the order of EM and the BestOrders of M for TEST1, $T(EM_M, TEST1)$, is given by the formula:[17]

$$T(EM_M, TEST1) = \frac{\tau(EM.OR1,M.OR1.1,TEST1)+...+\tau(EM.OR1,M.OR1.K,TEST1)}{K} \quad (9.6)$$

Consequently, the average distance between (the orders of) EM and (the orders of) M, $T(EM_M)$ often referred to as "the distance between EM and M", for N Testitems is calculated as follows:

$$T(EM_M) = \frac{T(EM_M,TEST1)+...+T(EM_M,TESTN)}{N} \quad (9.7)$$

Hence, one can compare the average distance between M and an expert e.g. $T(EM_M)$ with the distance between M and another expert e.g. $T(E1_M)$. Moreover, one can compare e.g. $T(EM_M)$ with $T(EM_{E1})$ to see whether EM stands closer to M than to E1, etc. Clearly, as the remainder of the chapter shows, this sort of analysis can be extended to scores involving M and all available experts.

In the next two sections, we present the analysis of the differences in the average distance between the orders provided by the experts and a) each other, b) the best scoring orders of the four metrics employed in this study (PF.BFP, PF.KP, PF.NOCB, M.NOCB) and c) the orders of RB using the Testitems from MPIRO-PROP. We start by formulating our predictions. Then, we report which predictions were verified.

9.5 Main experiment: Predictions

As Barzilay et al. (2002) report, differences between humans in the way they order sentences are not uncommon. Hence, it is difficult to make very specific predictions with respect to the average distance between the orderings of the experts and each other. On the other hand, one expects the experts to share some common ground in the way they put sentences in order. In this sense, some

[17] The distance between EM and RB for TEST1, $T(EM_{RB}, TEST1)$ is computed as the average of the τ values for the 10 randomly chosen orderings for TEST1.

of the distances between them should be short and not significantly different from each other. Thus, a particularly welcome result for our purposes would be to show that the average distance between EM and her colleagues is short and not significantly different from the distance between most of her colleagues and each other, although it might be the case that some other significant differences in the distances between the experts arise. This will show that EM is not a "stand-alone" expert but deviates from her colleagues to the same extent as they deviate from each other.

Despite the potential differences between the experts, we do expect that the distance between the orders of the experts and each other will be significantly lower than the distance between the orders of the experts and the orders of RB. This is again based on the assumption that even though the experts might not follow identical strategies, they do not operate with complete diversity either. In this sense, for a given pair of experts EM and E1, we predict that $T(EM_{E1})$ will be significantly greater than both $T(EM_{RB})$ and $T(E1_{RB})$, etc.

Since the metrics are found to produce the same BestTable for quite a few Testitems (see section 9.2.1), it is possible that they do not differ significantly from each other with respect to their average distance from the experts. For instance, it might be hard to expect e.g. $T(EM_{PF.BFP})$ to be significantly different from $T(EM_{PF.NOCB})$, since the scores that contribute to both means are identical for 11 out of the 16 Testitems.

Although PF.BFP outperforms PF.NOCB according to the analysis in section 8.3.3 of the previous chapter, the different aims and requirements of this study (pointed out in section 9.1) mean that this result does not need to be replicated here. Rather than comparing the metrics **directly** with each other, this study examines their behaviour with respect to the upper and the lower bound. We anticipate that this will not only reveal interesting points about the best scoring orderings for each metric, but can also allow us to compare the metrics with each other, albeit **indirectly**, in a sense that will be made clearer in the next paragraph.

What is crucial for the analysis is the difference of the T scores that involve the metrics from the upper and the lower bound. That is, although the average distance between a metric and an expert might not differ significantly from the average distance of another metric and the same expert, each T score may or may not differ significantly from the upper and the lower bound of the analysis. For instance, even though $T(EM_{PF.NOCB})$ and $T(EM_{PF.BFP})$ might not be significantly different from each other, it could be the case than one of them is significantly different from e.g. either $T(EM_{E1})$ or $T(EM_{RB})$ and $T(E1_{RB})$ but the other one is not.

We identify the best metrics in this study as the ones whose T scores i) are significantly greater from the T scores involving the experts and RB and ii) do not differ significantly from the T scores that involve the experts and each other.[18]

[18]Criterion (ii) can only be applied provided that the distance between the experts and at least one metric is found to

EM_{E1}:			**	**	**
0.692	EM_{E2}:		**	**	**
	0.717	$E1_{E2}$:	**	**	**
		0.758	EM_{E3}:		
CD at 0.01: 0.338			0.258	$E1_{E3}$:	
CD at 0.05: 0.282				0.300	$E2_{E3}$:
F(5,75)=14.931, p<0.000					0.192

Table 9.2: Comparison of distances between the experts (EM, E1, E2, E3) and each other

As we mentioned in the beginning of the chapter, this experiment is an additional evaluation test for the best performing metrics in the previous studies. This test may help us identify trends and differences that previously remained undetected. Clearly, a metric that has already done well in the previous chapters, gains extra bonus by performing well in an additional test. Hence, as far as the evaluation of the metrics goes, this study attempts to answer the following general question:

Q3: Which of the metrics that were previously identified as outperforming the baseline survive an additional evaluation task?

9.6 Results of main experiment

9.6.1 Distances between the experts and each other

As the first step in our analysis, we computed the 6 average T scores for the 4 experts, namely $T(EM_{E1})$, $T(EM_{E2})$, $T(EM_{E3})$, $T(E1_{E2})$, $T(E1_{E3})$ and $T(E2_{E3})$. Then we performed all comparisons between them using the Tukey test, the results of which are summarised in Table 9.2.

The cells in the Table report the level of significance returned by the Tukey test when the difference between two distances exceeds the critical difference (CD). Significance beyond the 0.05 threshold is reported with one asterisk (*), while significance beyond the 0.01 threshold is reported with two asterisks (**). A cell remains empty when the difference between two distances does not exceed the critical difference. For example, the value of $T(EM_{E1})$ is 0.692 and the value of $T(EM_{E3})$ is 0.258. Since their difference exceeds the CD at the 0.01 threshold, it is reported to be significant beyond that level by the Tukey test.[19]

be significantly lower that the distance between the experts and each other. Then, if the distance between the experts and another metric does not differ significantly from the distance of the experts with each other, the latter metric is among the best performing metrics in the study.

[19]The Table also reports the result of the omnibus ANOVA that uses all scores contributing to the means compared by the Tukey test, which shows a significant main effect of the factor DISTANCE: F(5,75)=14.931, p<0.000. As we explained

The Table shows that the average T scores for the distance between EM and E1 or E2, i.e. $T(EM_{E1})$ and $T(EM_{E2})$, as well as the average T for the distance between E1 and E2, i.e. $T(E1_{E2})$, are significantly greater than the average T scores achieved by E3 and any of the three aforementioned experts, i.e. $T(EM_{E3})$, $T(E1_{E3})$, and $T(E2_{E3})$. Further to this, the differences between $T(EM_{E1})$, $T(EM_{E2})$, and $T(E1_{E2})$ are not significant or approaching significance. The same is true for the differences between $T(EM_{E3})$, $T(E1_{E3})$, and $T(E2_{E3})$.

The comparison of $T(EM_{E1})$ with $T(EM_{E2})$ shows that the distance between the orderings of EM and the orderings of E1 is not significantly different from the distance between her orderings and the orderings of E2. Moreover, neither distance is significantly different from $T(E1_{E2})$, the distance between E1 and E2. This shows that the three experts deviate from each other to more or less the same extent.

Note that these three T values are quite high, which can be taken as an indication that on average the orderings of the three experts are quite close to each other. The fact that the T scores are high and not significantly different from each other suggests that EM, E1 and E2 share quite a lot of common ground in the ordering task. Hence, EM is found to give rise to similar orderings to the ones of E1 and E2, deviating from them only as much as they deviate from each other.

However, when any of the previous distances is compared with a distance that involves the orderings of E3 the difference is significant. In other words, although the orderings of E1 and E2 seem to deviate from each other and the orderings of EM to more or less the same extent, the orderings of E3 stand much further away from all of them. Hence, although there exists a "stand-alone" expert among the ones consulted in our studies, this is not EM but E3.

This finding can be easily explained if we consider that E3 employed domain communication knowledge (Kittredge et al. 1991) as her only constraint for the ordering task. As pointed out in her informal interview, E3 followed the very schematic way shown in Figure 9.2 for ordering the sentences presented to her. This schema gives rise to significantly lower T scores for the distance between her orderings and the orderings of the other experts compared to the distance between the orderings of the other experts and each other.

Although the schema implemented by E3 might indeed be a "way of generating readable text", the analysis in this section shows that the distance between the orderings that this schema outputs and the orderings of the other three experts is significantly greater than the distance between the orderings of the other three experts and each other. In this sense, it is the orderings of E3 that manifest rather peculiar ordering strategies, at least compared to the orderings of EM, E1 and E2. For this reason, the overall distance between the experts and each other, $T(EXP_{EXP})$, is computed without taking into

in section 9.4.2, the Tukey test reveals significant differences between conditions provided that the omnibus ANOVA is significant as well.

	General Schema:
1.	Location
2.	Other
	(e.g. depiction, technique, etc.)
3.	Date
	Schema in Location:
1a.	originates-from
1b.	current-location
	Schema in Date:
3a.	creation-period
3b.	creation-time

Figure 9.2: Schema employed by expert E3 for the ordering task

account the orders of E3:

$$T(EXP_{EXP}) = \frac{T(EM_{E1})+T(EM_{E2})+T(E1_{E2})}{3} = 0.722 \tag{9.8}$$

9.6.2 Distances between the experts and RB

As the upper part of Table 9.3 shows, the distance between any two experts other than E3 is significantly different from their distance from RB beyond the 0.01 threshold. This verifies one of our main predictions since it shows that the distance e.g. between EM and E1 is significantly shorter than the distances between the experts in question and RB.

Only the distances between E3 and another expert, shown in the lower section of Table 9.3, are not significantly different from the distance between E3 and RB. Note that this result does not mean that the orders of E3 are similar to the orders of RB.[20] It simply shows that E3 is roughly as far away from e.g. EM as she is from RB. By contrast, EM stands significantly closer to E1 than to RB, and the same holds for the other distances in the upper part of the Table.

In accordance with the discussion in the previous section, the overall distance between the experts (excluding E3) and RB, $T(EXP_{RB})$, is computed as follows:

$$T(EXP_{RB}) = \frac{T(EM_{RB})+T(E1_{RB})+T(E2_{RB})}{3} = 0.341 \tag{9.9}$$

[20] This could have been argued, had $T(E3_{RB})$ been much closer to 1.

EM_{E1}:			**	**	**
0.692	EM_{E2}:		**	**	**
	0.717	$E1_{E2}$:	**	**	**
		0.758	EM_{RB}:		
CD at 0.01: 0.242			0.323	$E1_{RB}$:	
CD at 0.05: 0.202				0.347	$E2_{RB}$:
F(5,75)=18.762, p<0.000					0.352

EM_{E3}:			
0.258	$E1_{E3}$:		
	0.300	$E2_{E3}$:	
CD at 0.01: 0.219		0.192	$E3_{RB}$:
CD at 0.05: 0.177			0.302
F(3,45)=1.223, p=0.312			

Table 9.3: Comparison of distances between experts (EM, E1, E2, E3) and the random baseline (RB)

9.6.3 Distances between the experts and each metric

In the first phase of our analysis we identified E3 as an "stand-alone" expert standing further away from the other three experts than they stand from each other. We also identified the distance between E3 and each expert as similar to her distance from RB.

In the second phase of our analysis we computed the distance between the best scoring orders of a metric and the orders of each expert. For instance, we computed $T(EM_{PF.BFP})$, $T(E1_{PF.BFP})$, $T(E2_{PF.BFP})$ and $T(E3_{PF.BFP})$, and compared the scores with each other. The results of these comparisons are displayed in Table 9.4.

Because the T scores which involve E3 are always the lowest, they introduce the greatest differences in the comparisons. Hence, for all three PF-modified metrics (PF.BFP, PF.KP and PF.NOCB), the average T between E3 and a metric is significantly lower than at least one of the average T values involving another expert and the same metric.

Consequently, E3 tends to stand further away from the metrics compared to their distance from the other three experts. This result, together with the remarks made in the previous two sections with respect to the distances from E3, give rise to the following set of formulas for calculating the overall distance between the experts (excluding E3) and each metric:

$EM_{PF.BFP}$: 0.604			*
	$E1_{PF.BFP}$: 0.713		**
		$E2_{PF.BFP}$: 0.571	
			$E3_{PF.BFP}$: 0.337

CD at 0.01: 0.303
CD at 0.05: 0.244
F(3,45)=5.992, p=0.002

$EM_{PF.KP}$: 0.546			
	$E1_{PF.KP}$: 0.654		*
		$E2_{PF.KP}$: 0.513	
			$E3_{PF.KP}$: 0.371

CD at 0.01: 0.278
CD at 0.05: 0.224
F(3,45)=3.892, p=0.015

$EM_{PF.NOCB}$: 0.592			*
	$E1_{PF.NOCB}$: 0.679		**
		$E2_{PF.NOCB}$: 0.546	
			$E3_{PF.NOCB}$: 0.354

CD at 0.01: 0.294
CD at 0.05: 0.237
F(3,45)=4.833, p=0.005

$EM_{M.NOCB}$: 0.489			
	$E1_{M.NOCB}$: 0.546		
		$E2_{M.NOCB}$: 0.425	
			$E3_{M.NOCB}$: 0.349

CD at 0.01: 0.245
CD at 0.05: 0.198
F(3,45)=2.635, p=0.061

Table 9.4: Comparison of distances between experts (EM, E1, E2, E3) and each metric (PF.BFP, PF.KP, PF.NOCB, M.NOCB)

EXP_{EXP}:			**	**	**
0.722	$EXP_{PF.BFP}$:			*	**
	0.629	$EXP_{PF.NOCB}$:			**
		0.606	$EXP_{PF.KP}$:		**
CD at 0.01: 0.150			0.571	$EXP_{M.NOCB}$:	*
CD at 0.05: 0.125				0.487	EXP_{RB}:
F(5,75)=19.111, p<0.000					0.341

Table 9.5: Results of the concluding analysis comparing the distance between the experts and each other (EXP_{EXP}) with the distance between the experts and each metric (PF.BFP, PF.NOCB, PF.KP, M.NOCB) and the random baseline (RB)

$$
\begin{aligned}
T(EXP_{PF.BFP}) &= \frac{T(EM_{PF.BFP})+T(E1_{PF.BFP})+T(E2_{PF.BFP})}{3} = 0.629 \\
T(EXP_{PF.KP}) &= \frac{T(EM_{PF.KP})+T(E1_{PF.KP})+T(E2_{PF.KP})}{3} = 0.571 \\
T(EXP_{PF.NOCB}) &= \frac{T(EM_{PF.NOCB})+T(E1_{PF.NOCB})+T(E2_{PF.NOCB})}{3} = 0.606 \\
T(EXP_{M.NOCB}) &= \frac{T(EM_{M.NOCB})+T(E1_{M.NOCB})+T(E2_{M.NOCB})}{3} = 0.487
\end{aligned} \tag{9.10}
$$

In the next section, we present the concluding analysis for the main study which compares the overall distances in (9.8), (9.9) and (9.10) with each other. As we have already mentioned, $T(EXP_{EXP})$ serves as the upper bound of the analysis whereas $T(EXP_{RB})$ is the lower bound. The aim is to specify which scores in (9.10) are significantly greater than $T(EXP_{RB})$, but do not differ significantly from $T(EXP_{EXP})$.

9.6.4 Concluding analysis

The results of the comparisons of the scores in (9.8), (9.9) and (9.10) are shown in Table 9.5. As we have mentioned in section 9.2.1, because the BestTables of the metrics have plenty of orderings in common, most of the scores that involve them are not significantly different from each other, except for $T(EXP_{PF.BFP})$ which is significantly greater than $T(EXP_{M.NOCB})$ at the 0.05 level. Also note that the difference between $T(EXP_{PF.NOCB})$ and $T(EXP_{M.NOCB})$ falls only 0.006 points short of CD at the 0.05 threshold, whereas the difference between $T(EXP_{PF.NOCB})$ and $T(EXP_{EXP})$ is only 0.009 points away from significance.

Crucially, what we are mainly interested in is how the distance between the experts and each metric compares with the distance of the experts from each other, $T(EXP_{EXP})$, and their distance from RB, $T(EXP_{RB})$. This is shown in the first row and the last column of Table 9.5.

As the Table shows, $T(EXP_{RB})$ is significantly lower than $T(EXP_{EXP})$ at the 0.01 level. As in Table 9.3 above, this result shows that randomly assembled orderings are significantly further away

from the orderings of the experts than the orderings of the experts are from each other.

Crucially, $T(EXP_{RB})$ is lower than $T(EXP_{PF.BFP})$, $T(EXP_{PF.NOCB})$ and $T(EXP_{PF.KP})$ as well, at the same level of significance. Notably, even the distance of the experts from M.NOCB, $T(EXP_{M.NOCB})$, is significantly greater than $T(EXP_{RB})$, albeit at the 0.05 level. This result shows that the distance from the experts is significantly reduced when using the best scoring orderings of any metric, even M.NOCB, instead of the orderings of RB. Hence, all metrics score significantly better than RB in this experiment.

However, simply using M.NOCB to output the best scoring orders is not enough to yield a distance from the experts which is comparable to $T(EXP_{EXP})$. Although the PF-modification appears to help towards this direction, $T(EXP_{PF.KP})$ remains significantly lower than $T(EXP_{EXP})$, whereas $T(EXP_{PF.NOCB})$ falls only 0.009 points short of CD at the 0.05 threshold. Hence, PF.BFP appears to be the most robust metric, as the difference between $T(EXP_{PF.BFP})$ and $T(EXP_{EXP})$ is clearly not significant.

The different performance of the three PF-modified metrics can be investigated by taking account of the BestOrders that differentiate them from each other. As we mentioned in section 9.2.1, most BestTables of PF.BFP and PF.NOCB are identical. The only exception to this are the two BestOrders that PF.NOCB adds to the BestTable of PF.BFP (and PF.KP) in 5 Testitems (see Figure 9.1). As these additional BestOrders yield low τ values with the three experts, not only does $T(EXP_{PF.NOCB})$ end up being lower than $T(EXP_{PF.BFP})$ but also falls outside the CD in its difference from $T(EXP_{EXP})$ by just 0.009 points. In this sense, PF.NOCB bears a penalty in its comparison with the upper bound that PF.BFP avoids.

Note that despite this penalty $T(EXP_{PF.NOCB})$ manages to be only 0.006 points away from the CD when compared to $T(EXP_{M.NOCB})$, showing a marginally better performance than its non-PF counterpart. This result shows that the distance from the experts is reduced to a great extent when the best scoring orderings are computed according to PF.NOCB instead of simply M.NOCB. Hence, this experiment provides additional evidence in favour of the PF modification of M.NOCB in MPIROPROP, showing that the PF constraint of entity coherence is not specific to EM but is shared by her colleagues as well.

Interestingly, since PF.KP shares the same BestTable as PF.BFP for the Testitems that penalise PF.NOCB, $T(EXP_{PF.KP})$ should also benefit from the aforementioned Testitems as well. However, the two additional BestOrders in the 4 Testitems for which the BestTable of PF.KP is distinct from the BestTables of PF.BFP and PF.NOCB return low τ values as well. This pushes $T(EXP_{PF.KP})$ down compared to the T scores of the other two PF-modified metrics. Hence, $T(EXP_{PF.KP})$ ends up not simply lower than $T(EXP_{PF.BFP})$ and $T(EXP_{PF.NOCB})$, but clearly significantly further away from $T(EXP_{EXP})$.

(9.11) (a) This exhibit is an amphora. (b) This exhibit depicts a warrior performing splachnoscopy before leaving for the battle. (c) This exhibit was decorated by “the painter of Kleofrades”. (d) The “painter of Kleofrades” used to decorate big vases. (e) This exhibit is currently displayed in the Martin von Wagner Museum. (f) The Martin von Wagner Museum is in Germany.

(9.12) (a) This exhibit is an amphora. (b) This exhibit depicts a warrior performing splachnoscopy before leaving for the battle. (e) This exhibit is currently displayed in the Martin von Wagner Museum. (f) The Martin von Wagner Museum is in Germany. (c) This exhibit was decorated by “the painter of Kleofrades”. (d) The “painter of Kleofrades” used to decorate big vases.

Figure 9.3: Example BestOrders for metric PF.BFP

9.6.5 PF.BFP: A case for variability in text structuring?

The discussion in the previous section reveals that PF.BFP is the metric that not only outperforms the baseline in the previous experiments but does best in the additional evaluation task in the current chapter as well. Hence, PF.BFP is identified as the most promising candidate for text structuring in the MPIRO domain (modulo AllEq) among the ones investigated in the thesis.

The analysis in section 9.6.1 has shown that even though the orderings of most of the experts are similar to each other, they are not always identical. This suggests that, instead of simply replicating the orderings of only one expert, accounting for the *variability* between the experts should be seen as a desideratum for a text structuring algorithm, although the variability between the outputs of the algorithm ought to be proportionate to the limitations set by the experts themselves. PF.BFP does not simply prioritise just a few BestOrders (out of a much larger search space of possible permutations), but also seems to be the metric which allows for enough variability in the BestTable without severely violating these limits.

However, it seems that even PF.BFP might occasionally allow for what appears to be too much variability in its preferred outputs. Such an example is presented in Figure 9.3 which shows the realisations of the two members of the BestTable for one of the Testitems in our study.

BestOrder (9.11) in Figure 9.3 is significantly close to the order of EM and identical to the orders of the other two experts (which are identical to each other in this case).[21] However, BestOrder (9.12) yields much lower τ values. This is because the preferences for entity coherence expressed by PF.BFP

[21] The order of EM is shown in example (8.1) of the previous chapter (Note that the indexation of the utterances in example (8.1) is different from the one in Figure 9.3). Using the indexation in Figure 9.3 as the point of reference, EM places fact (b) between facts (d) and (e). The τ value for the distance between (9.11) and (8.1) is 0.733 (p=0.020).

in this case are supplemented with a preference to place utterances (e) and (f) at the end of the description which is shared by all three experts and penalises BestOrder (9.12). Although the experts are not entirely consistent as far as such preferences are concerned (thus allowing for the observed variation between their orderings), it seems that cases like (9.12) account for the observed, yet not significant, difference between $T(EXP_{EXP})$ and $T(EXP_{PF.BFP})$ in the previous section.

Note that approaches to text structuring such as the one presented by Dimitromanolaki and Androutsopoulos (2003) (see section 7.2 of chapter 7 for a short overview), are not immune to the problems discussed in this chapter either. Because Dimitromanolaki and Androutsopoulos (2003) aim at replicating the orderings of EM, like most other approaches, they do not account for the observed variability between the experts at all.[22] Thus, unless trained on data from multiple experts, these techniques cannot distinguish between the strategies solely used by EM and the strategies shared between more than one expert.

In any case, the ultimate test for the ability of PF.BFP to generate felicitous structures should come from human judgements on the readability of its preferred outputs. Although in chapter 4 we argued that using perceptual experiments to address general questions such as (Q2) first posed in chapter 3 is extremely hard, the discussion in this thesis has resulted into a few experimental questions which are much easier to investigate with the help of psycholinguistic techniques.

One of these questions is whether the BestOrders in Figure 9.3 differ in their readability. In our future work, we intend to ask naive judges to provide us with scores of "goodness" for preferred possible outputs of PF.BFP such as (9.12). These scores will be compared to scores obtained for randomly generated orderings, the outputs of the ML-informed algorithms of Dimitromanolaki and Androutsopoulos (2003) and the orders of the experts consulted in this study. Excluding orderings such as (9.11) from the outputs of PF.BFP, the experiment will investigate whether the additional variability allowed by PF.BFP gives rise to orders which are perceived as less felicitous than the orders of the experts and the outputs of the ML-informed algorithms.

Since the available data derived from MPIRO consist of just a few facts which in turn give rise to texts of predetermined length, it is not certain how well PF.BFP performs when more facts are used as the actual input to text structuring. Karamanis and Manurung (2002) show how a metric of entity coherence can guide a stochastic approach to text structuring when large inputs are provided. In our future work, we intend to use PF.BFP as the evaluation function of the genetic algorithm in Karamanis and Manurung (2002), using the methods discussed in Cheng (2002, Chapter 8) to formally evaluate its performance.

[22] However, as we mentioned from very early on, the algorithms in Dimitromanolaki and Androutsopoulos (2003) are much more informed in the way they order the data from AllEq.

expert pair	Dataset	
	MPIRO-PROP	AllEq
EM_{E1}	68.75	16.67
EM_{E2}	68.75	33.33
$E1_{E2}$	68.75	33.33
N	16	6

Table 9.6: Testitems (%) for which the experts achieve significant τ scores

9.7 Differences between MPIRO-PROP and AllEq

As we mentioned in section 9.1.1, assuming that entity coherence is important for the experts in their search for a good ordering, one expects that the distance between their outputs in the Testitems from MPIRO-PROP will be smaller than the distance in AllEq, since the space of the most entity coherent solutions in AllEq is much wider than in MPIRO-PROP.

The second column of Table 9.6 reports the percentage of τ values which are associated with a significant *z* score for each pair of experts (except for E3) in the Testitems from MPIRO-PROP. The third column of the Table reports the percentage of significant τ values in AllEq.[23]

As the Table shows, the percentage of significant τ values for each pair of experts is 68.75% (11/16). By contrast, the percentage of significant τ values in AllEq does not exceed 33.33% (2/6). As we mentioned in section 9.4.1, reporting the percentage of significant τ values is useful for descriptive purposes. In addition to this, we were interested to see whether the difference in the distance between the experts and each other in the two datasets is significant.

As Table 9.7 shows, the average T between two experts in the Testitems from MPIRO-PROP is always greater than in AllEq. A 3X2 ANOVA with factors PAIR (EM_{E1}, EM_{E2}, $E1_{E2}$) and DATASET (MPIRO-PROP vs AllEq) showed a marginally significant main effect of DATASET: F(1,20)=4.571, p=0.045. The effect of PAIR and the interaction between PAIR and DATASET were not significant: F(2,40)=2.266, p=0.117 and F(2,40)=1.714, p=0.193, respectively.

Hence, it is indeed the case that the experts are significantly closer to each other in MPIRO-PROP than in AllEq. As we mentioned in the beginning of the chapter, it is plausible that the difference in the distance of the experts in the two datasets is due to the larger space of entity coherent solutions that AllEq enables when compared to the more restrictive MPIRO-PROP.

[23] As we mentioned in section 9.4.1, a τ value in the datasets is significant when it is equal to or greater than 0.696. Note that all significant τ values in both datasets are positive, that is, there is no case where the experts are significantly away from each other.

expert pair	Dataset	
	MPIRO-PROP	AllEq
$T(EM_{E1})$	0.692	0.333
$T(EM_{E2})$	0.717	0.578
$T(E1_{E2})$	0.758	0.600
Average	0.722	0.504
N	16	6

Table 9.7: Average T in MPIRO-PROP and AllEq

9.8 Summary and conclusion

A question not addressed until this chapter is whether the results from MPIRO-PROP are specific to EM. In order to answer this question in a general way, the dataset from MPIRO-PROP is enhanced with orderings provided by more than one expert. Then, the distance between EM and her colleagues is computed and compared to the distance between her colleagues and each other. The results indicate that EM shares a lot of common ground with two of her colleagues in the ordering task deviating from them as much as they deviate from each other, while the orderings of a fourth "stand-alone" expert are found to manifest rather peculiar ordering strategies.

The same methodology used to investigate the distance between the experts is used to automatically evaluate the best scoring orderings of some of the best performing metrics so far. This attempts to account for a number of possible deficiencies of the main methodology employed in the experiments previously reported in the thesis. The best scoring permutations of these metrics are isolated and evaluated by comparing their distance from the orderings produced by multiple experts with the distance of the orderings of the experts from each other and their distance from a random baseline RB.[24] The main results of this study are summarised as follows:

First, the distance of the experts from each metric employed in this study, was significantly lower than their distance from RB. This result shows that the distance from the experts is significantly reduced when the best scoring orderings of any metric, even M.NOCB, are used instead of the orderings assembled by RB. Hence, all metrics are superior to RB.

Moreover, the distance from the experts is reduced to a great extent when the best scoring orderings are computed according to PF.NOCB instead of simply M.NOCB. Hence, this experiment provides additional evidence in favour of the PF modification suggested in the previous chapter, showing

[24] As we mentioned in section 9.2.1, PF.MIL had to be excluded at the very early stages of the study, as its BestOrders are identical with the BestOrders of PF.NOCB for all randomly sampled Testitems. Hence, an extension of the study in this chapter to account for PF.MIL is desirable in order to estimate its performance in a more complete way.

that the PF constraint of entity coherence is not specific to EM but is shared by her colleagues as well.

Crucially, $T(EXP_{PF.BFP})$ is the only distance which is not significantly different from the distance of the experts from each other. By contrast, the difference between $T(EXP_{EXP})$ and $T(EXP_{PF.NOCB})$ approaches significance, whereas, $T(EXP_{PF.KP})$ and $T(EXP_{M.NOCB})$ remain significantly lower than $T(EXP_{EXP})$. Hence, PF.BFP, one of the best performing metrics in the previous chapter, yields the best results in this study as well and can be rendered as the most promising candidate for text structuring in MPIRO-PROP among the ones investigated in the thesis.

The portability of PF.BFP to a domain other than MPIRO depends on how similar the new domain is. PF.BFP is the recommended metric for a domain that is very similar to MPIRO (at least between the metrics investigated in this thesis), but if the new domain substantially deviates from MPIRO the evaluation methodology outlined in this thesis might prove crucial for an informed decision: The performance of M.NOCB can be compared to the performance of some of its competitors (possibly excluding solutions such as M.KP and M.CHEAP that are beaten by the baseline both in the genre of interest and the domain of application as in chapters 6 and 7), modifications such as the one discussed in chapter 8 can be introduced and tested, while the best performing metrics can be subjected to an additional evaluation task such as the one presented in chapter 9.

A first step to the portability of PF.BFP to a different application would be to incorporate it to the genetic algorithm of Mellish et al. (1998a) since this application can be seen as being "between" MPIRO and GNOME in the sense that the only additional factor to entity coherence in ILEX is represented by rhetorical relations. Keeping the evaluation features of rhetorical coherence the same, PF.BFP can replace the features of entity coherence and the outputs of the algorithm can be inspected and formally evaluated using human judgements.

Chapter 10

Concluding remarks

This chapter summarises the primary results of the thesis, presents its main contributions and points out possible extensions of our work.

10.1 Contributions

This thesis provides substantial insight into the role of entity coherence as a text structuring constraint. A general methodology for comparing metrics of entity coherence for the purpose of search-based text structuring is introduced and applied to data from two corpora. In a series of empirical studies, the metrics which constitute the most motivated candidates for descriptive text structuring (between the ones investigated) are identified before the actual generation takes place. The evaluation methodology and the results of these studies are useful for any subsequent attempt to generate a descriptive text structure in the context of an application that makes use of the notion of entity coherence.

More specifically, chapter 2 motivates using Centering Theory (CT) to define evaluation metrics of entity coherence for search-based descriptive text structuring. While previous work on NLG has considered CT only in passing, this chapter assesses its potential for this research area, and text structuring in particular, in substantial detail.

Chapter 3 shows how CT's notions can be used to define many different metrics of entity coherence. We argue that CT is open-ended enough for one to propose new metrics that in theory appear as plausible as some existing ones. Hence, a general methodology for identifying which metrics represent more suitable candidates for text structuring is required, so that at least some of the possible metrics can be compared empirically.

After arguing in chapter 4 that resolving the competition between the metrics using psycholinguistic methods requires a complex experimental design in which the confounding factors would be particularly difficult to control for, an alternative methodology for deciding which metrics represent

good candidates for the purposes of NLG is presented in chapter 5. This new corpus-based, search-oriented methodology is general enough to be applied to any possible CT-based metric and can be supplemented by less extended, but generally more costly, human evaluation studies.

The corpus-based studies in the next two chapters apply the novel search-oriented methodology to investigate the performance of eight of the metrics discussed in chapter 3. Despite restricting the empirical investigation to eight metrics for practical reasons, this thesis considers more metrics of entity coherence than any previous work.

Our first study in chapter 6 makes use of GNOME-LAB, a subset of the GNOME corpus representing the genre of interest. The main result of this study is that, none of the other employed metrics of entity coherence manages to return significantly better results that the baseline metric M.NOCB which in fact beats two of its competitors. This chapter also touches on the interaction of entity coherence with rhetorical relations which might pose additional, albeit apparently conflicting, constraints on the generation of a descriptive structure.

The next chapter reports experiments on MPIRO-PROP, an application-specific corpus. These results manifest the superiority of the baseline even more emphatically, as M.NOCB now does significantly better than most of its competitors with the exception of M.MIL which overtakes it. An investigation of the structures that differentiate M.NOCB from M.MIL across both datasets shows that the marginal difference in favour of M.MIL is due to a specific feature of MPIRO-PROP that does not characterise GNOME-LAB.

In chapter 8, we begin inspecting some of the best scoring structures for M.NOCB and M.MIL more closely. This investigation equips the employed metrics with an additional constraint on entity coherence and motivates a new set of pairwise comparisons between the modified metrics. In these comparisons, a number of the modified metrics overtake the baseline in MPIRO-PROP, but not in GNOME-LAB. This identifies a number of promising candidates for text structuring in the particular application domain, but shows that M.NOCB remains very robust as far as the genre of interest is concerned.

All these results indicate that M.NOCB is a good starting point to investigate the effect of entity coherence in general. However, one has to keep in mind that the performance of M.NOCB in the genre of interest is not optimal. Since factors such as rhetorical coherence and the modification of the metrics in chapter 8 do not appear to help, what can supplement M.NOCB to improve its performance in the investigated genre remains unclear to us.

Our experimental efforts in the MPIRO domain are concluded in chapter 9. An alternative methodology which employs the distance between two orderings for the automatic evaluation of the metrics is discussed in an attempt to address a number of unresolved questions in the previous studies. Orderings from more than one expert are collected and used in a subsidiary evaluation which shows that the

metrics are superior to a random baseline. This study also provides additional evidence in favour of the constraint introduced in the previous chapter and identifies PF.BFP as the metric which performs best across all evaluation tasks. Hence, PF.BFP is the most promising candidate for text structuring in the MPIRO domain among the ones investigated in the thesis.

The portability of PF.BFP to a domain other than MPIRO depends on how similar the new domain is. PF.BFP is the recommended metric for a domain that is very similar to MPIRO (at least between the metrics investigated in this thesis), but if the new domain substantially deviates from MPIRO the evaluation methodology outlined in this thesis might prove crucial for an informed decision: The performance of M.NOCB can be compared to the performance of some of its competitors (possibly excluding solutions such as M.KP and M.CHEAP that are beaten by the baseline both in the genre of interest and the domain of application as in chapters 6 and 7), modifications such as the one discussed in chapter 8 can be introduced and tested, while the best performing metrics can be subjected to an additional evaluation task such as the one presented in chapter 9.

10.2 Possible extensions

Throughout the thesis we pointed out to a number of ways in which one can build upon the work presented in each chapter. In this section we comment on the directions that seem to be most interesting to us. We begin with extensions that come very close to the work reported in the thesis and conclude with suggestions that extend the scope of the thesis significantly.

10.2.1 Experimenting with more metrics

Although in chapter 3 we identified a large set of possible CT-based metrics, the empirical investigation in subsequent chapters is restricted to a handful of them. In our future work, we intend to turn our attention to metrics that employ the alternative POT rankings of section 3.2.2 and the extended PT transitions in Table 3.10 of chapter 3. The methodology of chapter 5 can be used to investigate whether any of these metrics overtakes the baseline in GNOME-LAB and how much they benefit from the PF-modification in MPIRO-PROP. Their BestOrders can then be compared to the combined human data using the methods of chapter 9.

10.2.2 Extending GNOME-LAB

One of the aims of the researchers working on the GNOME corpus is to extend its current size so that a corpus large enough to be used for standard reference (a kind of “semantic treebank”) is built. This is particularly welcome for our purposes as well, since it will make it possible to include more corpus instances in GNOME-LAB, which in turn might enable us to investigate more subtle differences than

the ones observed in chapter 6 and could shed more light on the features that can supplement M.NOCB as well. Adding more biographical texts to the existing ones on the Getty webpage will also be very helpful as it will allow us to investigate the performance of the metrics on a related, yet distinct, genre.

Two other directions of future work with respect to GNOME-LAB were mentioned in section 6.6 of chapter 6. First, given the importance of bridging references in the evaluation of Poesio et al. (2002), we would like to experiment with a configuration of CT which uses indirect realisation for the computation of the CF list. Finally, we intend to investigate the difference between the Finite and the Finite-RR way of computing the BfC in more detail than we had the opportunity to do in this thesis, e.g. by identifying factors that might account for the drop in the classification rate more clearly than the percentage of NOCBs.

10.2.3 Future work in the MPIRO domain

As the ultimate test for a text structuring method is the readability of the structures it favours, in section 9.6.5 of chapter 9 we outlined an experimental design to compare the BestOrders of PF.BFP with the orders of the consulted experts and the output of the ML-informed algorithm of Dimitromanolaki and Androutsopoulos (2003) on the basis of elicited human judgements.

Since the available data derived from MPIRO consist of just a few facts which in turn give rise to texts of predetermined length, it is not certain how well PF.BFP performs when more facts are used as the hypothetical input to text structuring. Karamanis and Manurung (2002) show how a metric of entity coherence can guide a stochastic approach to text structuring when large inputs are provided. In our future work, we intend to use PF.BFP as the evaluation function of the genetic algorithm in Karamanis and Manurung (2002), using the methods discussed in Cheng (2002, Chapter 8) to formally evaluate its performance.

A first step to the portability of PF.BFP to a different application would be to incorporate it to the genetic algorithm of Mellish et al. (1998a) since this application can be seen as being “between” MPIRO and GNOME in the sense that the only additional factor to entity coherence is represented by rhetorical relations. Keeping the evaluation features of rhetorical coherence the same, PF.BFP can replace the features of entity coherence and the outputs of the algorithm can be inspected and formally evaluated.

Finally, as PF.MIL had to be excluded from our investigation in section 9.6 chapter 9 due to the scarcity of the data that manifest its difference from PF.NOCB, an extension of the study in chapter 9 to account for PF.MIL is desirable in order to estimate its performance in a more complete way.

10.2.4 Choosing between more complex metrics

As we clarified from very early on, the argumentation throughout the thesis and the inputs to our experiments were devised in such a way as to ignore possible interactions between text structuring and decisions such as content determination, segmentation, aggregation, etc.

Although we are not aware of a model of discourse structure which accounts for these interactions in enough detail, it is possible that specific phenomena can be captured by certain general heuristics. These preferences can supplement the metrics of entity coherence giving rise to larger evaluation modules as already suggested by Kibble and Power (2000). Each of these modules can then be employed by SEEC to specify which are the best candidates for an integrated generation system such as the ones presented e.g. in Cheng (2002) or Manurung (2003) that account for constraints interacting at different levels in the pipeline architecture.

In addition to this, one can allow for more flexibility in the way that SEEC computes the permutations of the BfC so that e.g. different aggregation decisions are manifested. Although the main side-effect of such modifications is that the space of possible permutations grows even larger, the discussion at the end of chapter 5 suggests that using large random samples might be able to overcome this problem without having to enumerate all possibilities exhaustively.

10.2.5 Computing input characteristics

Instead of giving general priority to one or more metrics for the purposes of text structuring, choosing the best metric to structure a specific semantic content that serves as the input to text structuring might depend on certain characteristics of the input. This was already mentioned in section 7.4.4 of chapter 7, where the percentage of computable ROUGH-SHIFTs given a certain semantic content was suggested as one possibility. Provided that such features are identified, a general methodology needs to be developed so that a given metric M is chosen to structure an input SC_B when SC_B exhibits a certain feature α which is easy to compute.

10.2.6 Psycholinguistic plausibility

Although this thesis does not make any specific claims about the psycholinguistic plausibility of the method used to evaluate the employed metrics, investigating this issue in more detail is another interesting direction for furture work. Alongside the perceptual experiments outlined in section 10.2.3, exploring a more psycholinguistically plausible search space of possible permutations can be introduced for the computation of the classification rate and the comparison of the performance of the metrics.

Appendix A

Examples of basic CT and extended PT transitions

In this appendix, we first show the analysis of examples (3.1) and (3.2) in terms of basic transitions and ESTABLISHMENTs, introduced in Table 3.7 and Table 3.8 of chapter 3 respectively. Then, we show how these examples score according to the two definitions of the extended PT transitions introduced in Table 3.10.

The NOCBs, basic transitions and ESTABLISHMENTs in examples (3.1) and (3.2), now repeated as (A.1) and (A.2), are as follows:

(A.1) a. This exhibit is an amphora.

`CF(exhibit1, amphora)`

b. Amphoras have an ovoid body and two looped handles, reaching from the shoulders up.

`CF(amphora, entity-3908),`

`CB=amphora,` EXP. ESTABLISHMENT

c. Amphoras were produced in two major variations: type A and the type with a neck.

`CF(amphora, typeA, type-neck)`

`CB=amphora,` CONTINUE

d. This exhibit is a type A amphora.

`CF(exhibit1, typeA)`

`CB=typeA,` EXP. ROUGH-SHIFT

e. This exhibit comes from the archaic period.

`CF(exhibit1, archaic-period)`

`CB=exhibit1,` SMOOTH-SHIFT

f. This exhibit was painted using the red figure technique.

`CF(exhibit1, red-figure-technique)`

`CB=exhibit1,` CONTINUE

(A.2) a. This exhibit is an amphora.

`CF(exhibit1, amphora)`

c. Amphoras were produced in two major variations: type A and the type with a neck.

`CF(amphora, typeA, type-neck)`

`CB=amphora,` EXP. ESTABLISHMENT

d. This exhibit is a type A amphora.

`CF(exhibit1, typeA)`

`CB=typeA,` EXP. ROUGH-SHIFT

b. Amphoras have an ovoid body and two looped handles, reaching from the shoulders up.

`CF(amphora, entity-3908)`

NOCB

e. This exhibit comes from the archaic period.

`CF(exhibit1, archaic-period)`

NOCB

f. This exhibit was painted using the red figure technique.

`CF(exhibit1, red-figure-technique)`

`CB=exhibit1,` ESTABLISHMENT

Without taking NOCBs into account, the basic transitions can be translated into extended PT transitions in two ways, depending on the way that ESTABLISHMENTS are incorporated to the definition of basic PT transitions (see Table 3.10 of chapter 3). The translation of the basic transitions in (A.1) and (A.2) into extended PT transitions according to the two configurations of Table 3.10 are:

	Transition in (A.1)	PT-EST-1	PT-EST-2
a.	-	-	-
b.	EXP. ESTABLISHMENT	V1	V2
c.	CONTINUE	V0	V0
d.	EXP. ROUGH-SHIFT	V3	V3
e.	SMOOTH-SHIFT	V1	V1
f.	CONTINUE	V0	V0

	Transition in (A.2)	PT-EST-1	PT-EST-2
a.	-	-	-
c.	EXP. ESTABLISHMENT	V1	V2
d.	EXP. ROUGH-SHIFT	V3	V3
b.	NOCB	-	-
e.	NOCB	-	-
f.	ESTABLISHMENT	V0	V1

Note how the EXP. ESTABLISHMENTs in (A.1b) and (A.2c), the second utterance of each example, are classified as V1 by the scoring function PT-EST-1 but as V2 by PT-EST-2. Moreover (A.2f), which follows a NOCB transition but itself contains a CB, is classified as the PT transition V0 by PT-EST-1, but as V1 by PT-EST-2. The extended PT transitions in each example are summarised as follows:

	PT-EST-1			
Text	V0	V1	V2	V3
(A.1)	c, f	b, e	-	d
(A.2)	f	c	-	d

	PT-EST-2			
Text	V0	V1	V2	V3
(A.1)	c, f	e	b	d
(A.2)	-	f	c	d

Note that example (A.1) is the structure that has more instances of the most preferred transition V0 in both configurations. Hence, the evaluation method of M.PT as defined at the end of section 3.5.2 of chapter 3 prefers (A.1) over (A.2) irrespective of whether PT-EST-1 or PT-EST-2 is used as the scoring function.

Appendix B

Instructions to participants in ME experiment

In this appendix, we present the instructions to the participants in the experiment discussed in chapter 4. As the experiment was web-based, each participant accessed the experiment using her browser. The first page she had to access included the instructions in a similar format as below.

Experiment on Text Acceptability

Thanks for taking part in this experiment!

To take part in this experiment you need to be a native speaker of English. If English is not your first language you could check The Psycholinguistic Experiment page for another experiment that suits you best as a possible participant.
Please read the instructions carefully before starting. Do not hesitate to contact the experimenter in case you have any questions or comments concerning this experiment.
If you experience any problems with our experimental software, please consult our Technical Problems Page.

Personal Details

As part of this experiment, we have to collect a small amount of personal information, which we ask you to enter in the Personal Details window below. *This information will be treated confidential, and will not be made available to a third party. None of the responses collected in this experiment will be*

associated with your name in any way. If you have any questions about this practice, please contact the experimenter.
Please be careful to fill in the Personal Details questionnaire correctly, as otherwise we will have to discard your responses.
We ask you to supply the following information:

- your name and email address;
- your age and sex;
- whether you are right or left handed (based on the hand you prefer to use for writing);
- the academic subject you study or have studied (or your current occupation in case you haven't attended university);
- under 'Region', please specify the place (city, region/state/province, country) where you have learned your first language.

Instructions

Part 1: Judging Line Length

Before doing the main part of the experiment, you will do a short task involving judging line length. A series of lines of different length will be presented on the screen. Your task is to estimate how long they seem by assigning numbers to them. *You are supposed to make your estimates relative to the first line you will see, your* **reference** *line. Give it any number that seems appropriate to you, bearing in mind that some of the lines will be* **longer** *than the reference and some will be* **shorter**. Click on "Continue" once you've decided on the reference number.
After you have judged the reference line, assign a number to each following line so that it represents **how long the line is in proportion to the reference**. The **longer** it is compared to the reference, the **larger** the number you will use; the **shorter** it is compared to the reference, the **smaller** the number you will use. So if you feel that a line is **twice as long as** the reference, give it a number **twice** the reference number; if it's **three times shorter** than the reference, provide a number three times smaller than the reference number. Hit RETURN after you've assigned each number.
So, if the reference is this line, you might give it the number 100:

———————————————-

If you have to judge this line, you might assign it 170:

And this one might be 25:

There is no limit to the range of numbers you may use. You may use whole numbers or decimals, but you cannot use zero or negative numbers. If you assigned the reference line the number 1, you might want to call the second one 1.7, and the last one 0.25 in order to express the same relations. Or you could use 10 for the reference line, 17 for the second one, and 2.5 for the last line. Just try to make each number match the length of the line as you see it.

Parts 2 and 3: Judging Texts

In Part 1 of the experiment you used numbers to estimate the length of lines on the screen. In Parts 2 and 3 you will use numbers to judge the *acceptability* of some English texts in the same way.

Those texts are short descriptions of archaeological exhibits that might remind you of the labels that are used in educational websites and virtual museums. There are more and less successful ways of describing such an exhibit and the numbers you will give to each text should reflect your judgement on the way that the text organises the information it consists of.

During the experiment you will see a series of short texts presented one at a time on the screen. Each text is different. Some will seem to organise the information in a good way, but others will not. Your task is to judge how good or bad each text is by assigning a number to it.

As with the lines in Part 1, you will first see a *reference* text, and you can use any number that seems appropriate to you for this reference. For each text after the reference, you will assign a number to show how good or bad that text is **in proportion to the reference text**.

For example, supposing that the texts describe objects from the Italian Renaissance and you are presented with the following reference text:

> (1) This exhibit is the portrait of “Mona Lisa”. It is kept in Louvre. Mona Lisa married in 1495 the well-known nobleman, Francesco del Giocondo, and thus came to be known as “La Gioconda”. The Louvre is the largest museum in France. This exhibit depicts Mona Lisa dressed in the Florentine fashion of her day and seated in a visionary, mountainous landscape.

you would probably give it a rather low number. (You are free to decide what ‘low’ or ‘high’ means in this context.) Supposing now that you read the following text:

> (2) This exhibit is the statue of David. David was chosen as the represented figure because his legend reflects the power and determination of Republican Florence. This exhibit is made of gigantic marble and is 4.34m tall. It is a creation of Michaelangelo. Michelangelo began work on it in 1501, and by 1504 the sculpture was in place outside the Palazzo Vecchio.

What you need to do is to compare text (2) with text (1), the reference text. If text (2) seemed **10 times better** than the reference, you'd give it a number **10 times** the number you gave to the reference. If it seemed **half as good as** the reference, you'd give it a number **half** the number you gave to the reference.

You can use any range of positive numbers that you like, including decimal numbers. *There is no upper or lower limit to the numbers you can use, except that you cannot use zero or negative numbers*.

Try to use a wide range of numbers and to distinguish as many degrees of acceptability as possible.

There are no 'correct' answers, so whatever seems right to you is a valid response. Most of the times you will have to read the reference text again for your judgement, but remember that we are interested in your first impressions, so please don't take too much time to think about each text: try to make up your mind quickly, spending less than a minute on each text.

Procedure

To participate in the experiment, please press the "Start" button.

First you will have to fill in the Personal Details questionnaire as explained above. Then you will be able to take part in the experiment.

The experiment will consist of the following 3 parts:

- Training session: judging 4 lines
- Practice session: judging 3 texts
- Experiment session: judging 24 texts

In each part you will see the reference item in the experiment window. Please enter your reference number and then press the "Continue" button. Now the test items will appear one after the other in the experiment window. Please type your judgement in the box below each item and hit RETURN in order to see the next item.

The experiment will take 15 to 20 minutes. After the experiment is completed you will receive an email confirmation of your participation.

Please keep in mind:

- Use any number you like for the reference text.
- Judge each text in proportion to the reference, that is, compare the reference text with the text that you are currently presented with.
- Use any positive numbers which you think are appropriate.
- Use high numbers for ‘good’ texts, low numbers for ‘bad’ texts and intermediate numbers for texts which are intermediate in acceptability.
- Try to use a wide range of numbers and to distinguish as many degrees of acceptability as possible.
- Try to make up your mind quickly, basing your judgements on your first impressions.

Appendix C

Weighting Equal for the classification rate

In this appendix, we go back to the SplitEqual configuration which arises in the individual comparison of M.NOCB with either M.SHOT1 or M.POT1 and was graphically represented in Figure 5.3 of chapter 5. Using the notation of section 5.5.1 of chapter 5, we show that when ToWorse(M_y) is higher that half of SplitEqual(M_y), then the classification rate of M_x on B is higher than the classification rate of M_y on B and vice versa:

$$ToWorse(M_y) > \frac{SplitEqual(M_y)}{2} \Leftrightarrow \upsilon(M_x, B) > \upsilon(M_y, B)$$

This can be used to support using $\frac{1}{2}$ as the value of the weight for the percentage of Equal in the definition of the classification rate in equation (5.4) of chapter 5.

Equations

1. $ToBetter(M_y) = Better(M_y) - Better(M_x)$
2. $ToWorse(M_y) = Equal(M_x) - Equal(M_y) - ToBetter(M_y) \stackrel{1}{=}$
   $ToWorse(M_y) = Equal(M_x) - Equal(M_y) - Better(M_y) + Better(M_x)$
3. $SplitEqual(M_y) = Equal(M_x) - Equal(M_y)$
4. $\upsilon(M_x, B) = Better(M_x) + \frac{Equal(M_x)}{2}$
   $\upsilon(M_y, B) = Better(M_y) + \frac{Equal(M_y)}{2}$

Proof

$$ToWorse(M_y) > \frac{SplitEqual(M_y)}{2} \stackrel{2+3}{\Leftrightarrow}$$

$$Equal(M_x) - Equal(M_y) - Better(M_y) + Better(M_x) > \frac{Equal(M_x) - Equal(M_y)}{2} \Leftrightarrow$$

$$Better(M_x) + \frac{Equal(M_x)}{2} > Better(M_y) + \frac{Equal(M_y)}{2} \Leftrightarrow$$

$$\upsilon(M_x, B) > \upsilon(M_y, B)$$

Appendix D

Instructions to experts

In this appendix, we present the instructions to the experts who were consulted for the purposes of the study in chapter 9.

Experiment on Sentence Ordering

Goal of the experiment

I investigate how sentences are ordered within a short text. For this reason, I ask humans to perform the same task. Then, I will compare the human orderings with the ones generated by my program.

The Task

The task you will be doing should take approximately 45 minutes.

I will provide you with a few sets of six sentences. I will give you the first sentence for each set and ask you to order the remaining five. The sentences come from a computer program that generates descriptions of artefacts in a virtual museum.

Your task is to order them in a coherent text.

When ordering the sentences, try to look at which ones should be together and which should come before another in the text; do not try to use hints other than the sentences themselves.

For each set: Read the first sentence. Then, read the unordered sentences and select the sentence that should follow the first sentence. Then, read the remaining sentences and select the one that should

be placed next. **You can revise the ordering at any time by moving the sentences around**. When you are satisfied with the ordering you produced write next to each sentence its position, starting with number 2 for the second sentence. Then, give me the sentences and perform the same task with the next set.

Because I do not want you to take into account explicit references while ordering the sentences, the texts do not contain any pronouns. For example, each text starts with the phrase “This exhibit is”. Instead of *it*, you will see the phrase *this exhibit* in every sentence that you need to order and refers to the same object as the first sentence.

Fell free to ask me any questions at any time during the experiment.

Thanks for your participation!

Bibliography

Alexopoulou, T. and Keller, F. (2003). Linguistic complexity, locality and resumption. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, Somerville, MA. Cascadilla Press.

Baker, C., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of ACL/COLING 1998*. Montreal, Canada.

Bangalore, S., Rambow, O., and Whittaker, S. (2000). Evaluation metrics for generation. In *Proceedings of INLG 2000*, pages 1–8, Israel.

Bard, E., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Barzilay, R., Elhadad, N., and McKeown, K. (2002). Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Barzilay, R. and McKeown, K. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of EACL/ACL 2001*, pages 50–57, Toulouse, France.

Beaver, D. (2003). The optimization of discourse anaphora. *Linguistics and Philosophy*. To appear.

Birner, B. J. (1998). Recency effects in English inversion. In Walker et al. (1998b), pages 309–326.

Bouayad-Agha, N., Power, R., and Scott, D. (2000). Can text structure be incompatible with rhetorical structure? In *Proceedings of INLG 2000*, pages 194–200, Israel.

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

Brennan, S. E. (1998). Centering as a psychological resource for achieving joint reference in spontaneous discourse. In Walker et al. (1998b), pages 227–250.

Brennan, S. E., Friedman [Walker], M. A., and Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of ACL 1987*, pages 155–162, Stanford, California.

Byron, D. K. and Stent, A. J. (1998). A preliminary model of centering in dialog. In *Proceedings of ACL/COLING 1998: Student Session*, Montreal, Canada.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chafe, W. (1976). Giveness, contrastiveness, definiteness, subjects, topics and point of view. In Li, C., editor, *Subject and Topic*, pages 25–76. Academic Press, New York.

Cheng, H. (2002). *Modelling Aggregation Motivated Interactions in Descriptive Text Generation*. PhD thesis, Division of Informatics, University of Edinburgh.

Cheng, H. and Mellish, C. (2000a). Capturing the interaction between aggregation and text planning in two generation systems. In *Proceedings of INLG 2000*, pages 186–194, Israel.

Cheng, H. and Mellish, C. (2000b). An empirical analysis of constructing non-restrictive NP modifiers to express semantic relations. In *Proceedings of INLG 2000*, Israel.

Cheng, H., Poesio, M., Henschel, R., and Mellish, C. (2001). Corpus-based NP modifier generation. In *Proceedings of NAACL 2001*, Pittsburgh, US.

Chinchor, N. A. and Sundheim, B. (1995). Message understanding conference (MUC) tests of discourse processing. In *Proceedings of the AIAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 21–26, Stanford, US.

Clark, H. H. (1977). Bridging. In Johnson-Laird, P. N. and Wason, P. C., editors, *Thinking: Readings in Cognitive Science*, pages 9–27. Cambridge University Press.

Cote, S. (1998). Ranking forward-looking centers. In Walker et al. (1998b), pages 55–71.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage, Thousand Oaks, California.

Dale, R. (1992). *Generating Referring Expressions*. MIT Press, MA/London.

De Beaugrande, R. and Dressler, W. (1981). *Introduction to Text Linguistics*. Longman, London.

Di Eugenio, B. (1990). Centering theory and the Italian pronominal system. In *Proceedings of COLING 1990*, pages 270–275, Helsinki, Finland.

Di Eugenio, B. (1996). The discourse functions of Italian subjects: A centering approach. In *Proceedings of COLING 1996*, pages 352–357, Copenhagen, Denmark.

Di Eugenio, B. (1998). Centering in Italian. In Walker et al. (1998b), pages 115–137.

Dimitriadis, A. (1996). When pro-drop languages don't: Overt pronominal subjects and pragmatic inference. In *Proceedings of CLS 32*, pages 33–47.

Dimitromanolaki, A. and Androutsopoulos, I. (2003). Learning to order facts for discourse planning in natural language generation. In *Proceedings of the 9th European Workshop on Natural Language Generation*, Budapest, Hungary.

Duboue, P. and McKeown, K. (2001). Empirically estimating order constraints for content planning in generation. In *Proceedings of EACL/ACL 2001*, pages 172–179, Toulouse, France.

Duboue, P. and McKeown, K. (2002). Content planner construction via evolutionary algorithms and a corpus-based fitness function. In *Proceedings of INLG 2002*, pages 89–96, Harriman, NY, USA.

Gaizauskas, R. (1998). Evaluation in language and speech technology. *Computer Speech and Language*, 12(4):249–262.

Givon, T., editor (1983). *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. John Benjamins.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names and the centering of attention in discourse. *Cognitive Science*, 17:311–347.

Grosz, B. J. and Gordon, P. C. (1999). Conceptions of limited attention and the discourse focus. *Computational Linguistics*, 25(4):617–624.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of ACL 1983*, pages 44–50, Cambridge, Mass.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, B. J. and Sidner, C. L. (1998). Lost intuitions and forgotten intentions. In Walker et al. (1998b), pages 39–53.

Grosz, B. J. and Ziv, Y. (1998). Centering global focus and right dislocation. In Walker et al. (1998b), pages 293–308.

Gundel, J. (1998). Centering theory and the giveness hierarchy: Towards a synthesis. In Walker et al. (1998b), pages 183–198.

Gundel, J., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Hardt, D. and Rambow, O. (2001). Generation of VP-ellipsis: A corpus-based approach. In *Proceedings of EACL/ACL 2001*, pages 282–289, Toulouse, France.

Henschel, R., Cheng, H., and Poesio, M. (2000). Pronominalisation revisited. In *Proceedings of COLING 2000*, pages 306–312. Germany.

Hitzeman, J., Black, A., Taylor, P., Mellish, C., and Oberlander, J. (1998). On the use of automatically generated discourse-level information in a concept-to-speech synthesis system. In *Proceedings of ICSLP 1998*, Australia.

Hitzeman, J., Mellish, C., and Oberlander, J. (1997). Dynamic generation of museum web pages: The intelligent labelling explorer. *Journal of Archives and Museum Informatics*, 11:107–115.

Hitzeman, J. and Poesio, M. (1998). Long-distance pronominalisation and global focus. In *Proceedings of ACL/COLING 1998*, volume 1, pages 550–556. Montreal, Canada.

Hobbs, J. (1978). Resolving pronoun references. *Lingua*, 44:311–338.

Hoffman, B. (1998). Word order, information structure, and centering in Turkish. In Walker et al. (1998b), pages 251–271.

Horn, L. R. (1986). Presupposition, theme and variations. *Chicago Linguistic Society*, 22:168–192.

Hovy, E. (1988). Planning coherent multisentential text. In *Proceedings of ACL 1988*, pages 163–169.

Hovy, E. (1990). Unresolved issues in paragraph planning. In Dale, R., Mellish, C., and Zock, M., editors, *Current Research in Natural Language Generation*, pages 17–45. Academic Press.

Hovy, E. (1991). Approaches to the planning of coherent text. In Paris, C. L., Swartout, W. R., and Mann, W. C., editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 83–102. Kluwer.

Hovy, E. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence: Special Issue on Natural Language Processing*, 63(1-2):341–386.

Hovy, E. and Maier, E. (1995). Parsimonious or profligate: How many and which discourse structure relations. Unpublished manuscript.

Hovy, E. and McCoy, K. F. (1989). Focusing your RST: A step toward generating coherent multisentential text. In *Proceedings of COGSCI 1989*, pages 667–674.

Howell, D. C. (2002). *Statistical Methods for Psychology*. Duxbury, Pacific Grove, CA, 5th edition.

Hudson-D'Zmura, S. (1998). Control and event structure: The view from the center. In Walker et al. (1998b), pages 71–88.

Hudson-D'Zmura, S. and Tanenhaus, M. K. (1998). Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment. In Walker et al. (1998b), pages 199–226.

Hudson-D'Zmura, S., Tanenhaus, M. K., and Dell, G. (1986). The effect of the discourse center on the local coherence of discourse. In *Proceedings of COGSCI 1986*, pages 96–101.

Hurewitz, F. (1998). A quantitative look at discourse coherence. In Walker et al. (1998b), pages 273–291.

Iida, M. (1998). Discourse coherence and shifting centers in Japanese texts. In Walker et al. (1998b), pages 161–180.

Isard, A., Oberlander, J., Androutsopoulos, I., and Matheson, C. (2003). Speaking the users' languages. *IEEE Intelligent Systems Magazine*, 18(1):40–45.

Kaiser, E. (2000). Pronouns and demonstratives in Finish: Indicators of referent salience. In *Proceedings of DAARC 2000*, pages 20–27.

Kameyama, M. (1985). *Zero Anaphora: The Case of Japanese*. PhD thesis, Stanford University.

Kameyama, M. (1988). Japanese zero pronominal binding: Where syntax and discourse meet. In Poser, W., editor, *Papers from the Second International Workshop on Japanese Syntax*, pages 47–74. (Stanford, CSLI). Also available as University of Pennsylvania Technical Report No. MS-CIS-86-60.

Kameyama, M. (1998). Intrasentential centering: A case study. In Walker et al. (1998b), pages 89–122.

Kameyama, M., Passoneau, R. J., and Poesio, M. (1993). Temporal centering. In *Proceedings of ACL 1993*, pages 70–77.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Kan, M.-Y. and McKeown, K. (2002). Corpus-trained text generation for summarization. In *Proceedings of INLG 2002*, pages 1–8, Harriman, NY, USA.

Karamanis, N. (2001). A centering-based algorithm for the generation of the animate subject in Greek. In *Studies in Greek Linguistics, Proceedings of the 22nd Annual Meeting*, Aristotle University of Thessaloniki.

Karamanis, N. and Manurung, H. M. (2002). Stochastic text structuring using the principle of continuity. In *Proceedings of INLG 2002*, pages 81–88, Harriman, NY, USA.

Kehler, A. (1997). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, 23(3):467–475.

Keller, F. and Alexopoulou, T. (2001). Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, 79:301–372.

Keller, F., Corley, M., Corley, S., Konieczny, L., and Todirascu, A. (1998). WebExp: A Java toolbox for web-based psychological experiments. Technical report, University of Edinburgh.

Kibble, R. (1999). Cb or not Cb? Centering theory applied to NLG. In *Proceedings of ACL 1999 Workshop: Reference and Discourse Structure*, pages 72–81, London.

Kibble, R. (2001). A reformulation of rule 2 of centering theory. *Computational Linguistics*, 27(4):579–587.

Kibble, R. and Power, R. (2000). An integrated framework for text planning and pronominalisation. In *Proceedings of INLG 2000*, pages 77–84, Israel.

Kim, H., Cho, J.-M., and Seo, J. (1999). Anaphora resolution using an extended centering algorithm in a multi-modal dialogue system. In *Proceedings of ACL 1999 Workshop: Reference and Discourse Structure*, pages 21–28, London.

Kintsch, W. and van Dijk, T. (1978). Towards a model of discourse comprehension and production. *Psychological Review*, 85:363–394.

Kittredge, R., Korelsky, T., and Rambow, O. (1991). On the need for domain communication knowledge. *Computational Intelligence*, 7:305–314.

Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In Sanders, T., Schilperoord, J., and Spooren, W., editors, *Text Representation: Linguistic and Psycholinguistic Aspects*, chapter 7, pages 181–196. John Benjamins, Amsterdam.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL 2003*, Saporo, Japan.

Lecoeuche, R., Mellish, C., Barry, C., and Robertson, D. (1998). User-system dialogues and the notion of focus. *The Knowledge Engineering Review*, 13(4):381–408.

Lesgold, A., Lajoie, S., Bunzo, M., and Eggan, G. (1992). Sherlock: A coached practice enviroment for an electronics troubleshooting job. In Larkin, J. and Chabay, R., editors, *Computer assisted instruction and intelligent tutoring systems: Shared issues and complemtary approaches*, pages 201–238. Erlbaum, Hillsdale, NJ.

Lester, J. and Porter, B. (1997). Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, 23(1):65–102.

Lyons, J. (1981). *Language, Meaning and Context*. Fontana, London.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organisation. Technical Report RR-87-190, University of Southern California / Information Sciences Institute.

Manurung, H. M. (2003). *An Evolutionary Algorithm Approach to Poetry Generation*. PhD thesis, Division of Informatics, University of Edinburgh.

Marcu, D. (1997). *The Rhetorical Parsing, Summarisation and Generation of Natural Language Texts*. PhD thesis, Department of Computer Science, University of Toronto.

Marcu, D. (1999). Instructions for manually annotating the discourse structures of texts. Unpublished manuscript.

Marcu, D. (2000). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Matsui, T. (1999). Approaches to Japanese zero pronouns: Centering and relevance. In *Proceedings of ACL 1999 Workshop: Reference and Discourse Structure*, pages 11–20, London.

McCoy, K. F. and Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description. In *Proceedings of ACL 1999 Workshop: Reference and Discourse Structure*, pages 63–71, London.

McKeown, K. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Studies in Natural Language Processing. Cambridge University Press.

Mellish, C. and Dale, R. (1998). Evaluation in the context of natural language generation. *Computer Speech and Language*, 12(4):349–373.

Mellish, C., Knott, A., Oberlander, J., and O'Donnell, M. (1998a). Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on NLG*, pages 98–107, Niagara-on-the-Lake, Ontario, Canada.

Mellish, C., O'Donnell, M., Oberlander, J., and Knott, A. (1998b). An architecture for opportunistic text generation. In *Proceedings of the 9th International Workshop on NLG*, pages 28–37, Niagara-on-the-Lake, Ontario, Canada.

Miltsakaki, E. (2002). Towards an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics*, 28(3):319–355.

Miltsakaki, E. and Kukich, K. (2000a). Automated evaluation of coherence in student essays. In *Proceedings of LREC 2000 Workshop: Language Resources and Tools in Educational Applications*.

Miltsakaki, E. and Kukich, K. (2000b). The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of ACL 2000*.

Mooney, D. J., Carberry, S., and McCoy, K. F. (1991). Capturing high-level structure of naturally occuring, extended explanations using bottom-up strategies. *Computational Intelligence*, 7:334–356.

Moore, J. D. and Paris, C. L. (1993). Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–695.

Moore, J. D. and Pollack, M. E. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Moser, M. and Moore, J. D. (1996). Towards a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

O'Donnell, M., Cheng, H., and Hitzeman, J. (1998). Integrating referring and informing in NP planning. In *Proceedings of ACL 1998 Workshop: Computational Treatment of Nominals*, pages 46–55, Montreal, Canada.

O'Donnell, M., Mellish, C., Oberlander, J., and Knott, A. (2001). Ilex: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7(3):225–250.

Passoneau, R. J. (1993). Getting and keeping the center of attention. In Bates, M. and Weischedel, R., editors, *Challenges in Natural Language Processing*, pages 179–227. Cambridge University Press.

Passoneau, R. J. (1997). Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.

Passoneau, R. J. (1998). Interaction of discourse structure with explicitness of discourse anaphoric phrases. In Walker et al. (1998b), pages 327–358.

Pearson, M. (2000). Determining a measure of textual coherence. MA dissertation, University of Edinburgh.

Poesio, M. (2000). Annotating a corpus to develop and evaluate discourse entity realization algorithms: Issues and preliminary results. In *Proceedings of LREC 2000*, pages 211–218. Athens, Greece.

Poesio, M., Bruneseaux, F., and Romary, L. (1999a). The MATE scheme for coreference in dialogues in multiple languages. In *Proceedings of ACL 1999 Workshop: Standards and Tools for Discourse Tagging*, pages 65–74.

Poesio, M., Cheng, H., Hitzeman, J., Kibble, R., and Stevenson, R. (2000). Specifying the parameters of centering theory: A corpus-based evaluation using text from application-oriented domains. In *Proceedings of ACL 2000*.

Poesio, M., Cheng, H., Hitzeman, J., Stevenson, R., and Di Eugenio, B. (2002). A corpus-based evaluation of centering theory. Technical Report CSM-369, Department of Computer Science, University of Essex.

Poesio, M., Henschel, R., Hitzeman, J., and Kibble, R. (1999b). Statistical generation: A first report. In *Proceedings of ESSLLI 1999 Workshop: The generation of Nominal Expressions*.

Poesio, M. and Stevenson, R. (2003). *Salience: Computational Models and Psychological Evidence*. Cambridge University Press. To appear.

Poesio, M., Stevenson, R., Di Eugenio, B., and Hitzeman, J. (2003). Centering: a parametric theory and its instantiations. *Computational Linguistics*. To appear.

Poesio, M. and Viera, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Power, R., Scott, D., and Bouayad-Agha, N. (2003). Document structure. *Computational Linguistics*, 29(2):221–260.

Prasad, R. (2000). A corpus study of zero pronouns in Hindi: An account based on centering transition preferences. In *Proceedings of DAARC 2000*, pages 66–71.

Prasad, R. and Strube, M. (2000). Discourse salience and pronoun resolution in Hindi. In *Penn Working Papers in Linguistics*, volume 6, pages 189–208.

Prince, A. and Smolensky, P. (1997). Optimality: from neural networks to universal grammar. *Science*, 275:1604–1610.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In Cole, P., editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness and information status. In Mann, W. C. and Thompson, S. A., editors, *Discourse description: Diverse linguistic analyses of a fund-raising text*, pages 295–325. John Benjamins, Amsterdam.

Prince, E. F. (1994). Subject prodrop in Yiddish. In Bosch, P. and van der Sandt, R., editors, *Focus and Natural Language Processing*, Intonation and Syntax: Working Papers of the IBM Institute for Logic and Linguistics, pages 159–174.

Prince, E. F. (1999). How not to mark topics: Topicalization in English and Yiddish. Unpublished manuscript.

Quirk, R. and Greenbaum, S. (1973). *A University Grammar of English*. Longman, Harlow, Essex.

Rambow, O. (1993). Pragmatic aspects of scrambling and topicalization in German. In *Workshop on Centering Theory in Naturally Occuring Discourse*, pages 20–28. Institute of Research in Cognitive Science, University of Pennsylvania.

Rambow, O., Rogati, M., and Walker, M. A. (2001). Evaluating a trainable sentence planner for a spoken dialogue travel system. In *Proceedings of EACL/ACL 2001*, pages 426–433, Toulouse, France.

Reinhart, T. (1981). Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27:53–94.

Reiter, E. (1994). Has a consensus NL generation architecture appeared, and is it psycholinguistically plausable? In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 163–170.

Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Journal of Natural Language Engineering*, 3:57–87.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Reiter, E. and Sripada, S. (2002). Should corpora texts be gold standards for NLG? In *Proceedings of INLG 2002*, pages 97–104, Harriman, NY, USA.

Roberts, C. (1998). The place of centering in a general theory of anaphora resolution. In Walker et al. (1998b), pages 359–399.

Scott, D. and de Souza, C. S. (1990). Getting the message across in RST-based text generation. In Dale, R., Mellish, C., and Zock, M., editors, *Current Research in Natural Language Generation*, pages 47–74. Academic Press.

Scott, D., Power, R., and Evans, R. (1998). Generation as a solution to its own problem. In *Proceedings of the 9th International Workshop on NLG*, pages 256–265, Niagara-on-the-Lake, Ontario, Canada.

Sibun, P. (1992). Generating text without trees. *Computational Intelligence*, 8(1):102–122.

Sidner, C. L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English*. PhD thesis, AI Laboratory/MIT, Cambridge, MA. Also available as Technical Report No. AI-TR-537.

Stevenson, R., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

Stevenson, R., Knott, A., Oberlander, J., and McDonald, S. (2000). Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. *Language and Cognitive Processes*, 15(3):225–262.

Strube, M. (1998). Never look back: An alternative to centering. In *Proceedings of ACL/COLING 1998*, volume 2, pages 1251–1257. Montreal, Canada.

Strube, M. and Hahn, U. (1996). Functional centering. In *Proceedings of ACL 1996*, pages 270–227. Santa Cruz, California.

Strube, M. and Hahn, U. (1999). Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Suri, L. A. and 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.

Taboada, M. (2002). Centering and pronominal reference: In dialogue, in Spanish. In *Proceedings of EDILOG 2002*, pages 177–184.

Tetreault, J. R. (1999). Analysis of syntax-based pronoun resolution methods. In *Proceedings of ACL 1999: Student Session*, University of Maryland, US.

Tetreault, J. R. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.

Turan, U. D. (1995). *Null vs Overt Subjects in Turkish Discourse: A Centering Analysis*. PhD thesis, University of Pennsylvania.

Turan, U. D. (1998). Ranking forward-looking centers in Turkish: Universal and language specific properties. In Walker et al. (1998b), pages 139–160.

Walker, M. A. (1989). Evaluating discourse processing algorithms. In *Proceedings of ACL 1989*, pages 251–260, Vancouver, Canada.

Walker, M. A. (1996). Limited attention and discourse structure. *Computational Linguistics*, 22(2):255–264.

Walker, M. A. (1998). Centering, anaphora resolution and discourse structure. In Walker et al. (1998b), pages 401–436.

Walker, M. A., Iida, M., and Cote, S. (1990). Centering in Japanese discourse. In *Proceedings of COLING 1990*, pages 1–8, Helsinki.

Walker, M. A., Iida, M., and Cote, S. (1994). Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

Walker, M. A., Joshi, A. K., and Prince, E. F. (1998a). Centering in naturally occuring discourse: An overview. In Walker et al. (1998b), pages 1–30.

Walker, M. A., Joshi, A. K., and Prince, E. F., editors (1998b). *Centering Theory in Discourse*. Clarendon Press, Oxford.

Walker, M. A. and Prince, E. F. (1995). A bilateral approach to givenness: A hearer-status algorithm and a centering algorithm. In Fretheim, T. and Gundel, J., editors, *Reference and Referent Accessibility*, pages 291–306. John Benjamins.

Webber, B. L. (1978). *A Formal Approach to Discourse Anaphora*. PhD thesis, Harvard University.