Dana Dannélls

Multilingual text generation from structured formal representations

Data linguistica
<http://www.svenska.gu.se/publikationer/data-linguistica/>

Editor: Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg

23 • 2012

Dana Dannélls

Multilingual text generation from structured formal representations

Gothenburg 2012

Data linguistica 23
ISBN 978-91-87850-48-6
ISSN 0347-948X

Printed in Sweden by
Ineko AB Göteborg 2012

Typeset in LaTeX 2ε by the author

Cover design by Kjell Edgren, Informat.se

Front cover illustration:
The right piece in the right place
by Kristian Dannélls ©

Author photo on back cover by Kristina Holmlid

ABSTRACT

This thesis aims to identify the optimal ways in which natural language generation techniques can be brought to bear upon the problem of processing a structured body of information in order to devise a coherent presentation of text content in multiple languages.

We investigate how chains of referential expressions are realized in English, Swedish and Hebrew, and suggest several coreference strategies that can be used to generate coherent descriptions about paintings. The suggested strategies focus on the need to produce paragraph-sized written natural language descriptions from formal structured representations presented in the Semantic Web.

We account for principles of coreference by introducing a new modularized approach to automatically generate chains of referential expressions from ontologies. We demonstrate the feasibility of the approach by implementing a system where a Semantic Web domain ontology serves as the background knowledge representation and where the language-specific coreference strategies are incorporated. The system uses both the principles of discourse structures and coreference strategies to guide the generation process. We show how the system successfully generates coherent, well-formed descriptions in multiple languages.

SAMMANFATTNING

Denna doktorsavhandling i språkvetenskaplig databehandling handlar om automatisk flerspråkig generering av beskrivande texter om museiföremål – närmare bestämt konstverk – från formella beskrivningar av den typ som utvecklats för den semantiska webben. Språkgenereringssystem som bygger på teknologier för den semantiska webben, t.ex. formella ontologier, ställer höga krav både på de språkspecifika genereringsprocesserna och på effektiv anpassning till olika mottagares behov av såväl de genererade texternas struktur som den information som förmedlas.

För att ett genereringssystem ska kunna producera föremålsbeskrivningar automatiskt på flera språk måste systemet ha information om hur sådana föremålsbeskrivningar kan och brukar realiseras syntaktiskt och semantiskt i varje språk. Om systemet dessutom ska generera sammanhängande beskrivningar på mer än ett språk, måste det ha kunskaper om de lingvistiska särdrag som bidrar till att beskrivningarna uppfattas som sammanhängande. I denna avhandling undersöks språkteknologiska metoder och teorier för att förbättra automatisk flerspråkig generering av sammanhängande texter i en avgränsad domän.

De övergripande syftena med den forskning som presenteras i avhandlingen är (1) att empiriskt undersöka hur museiföremålsbeskrivningar formuleras på de tre undersökta språken, samt (2) att omsätta resultatet av den empiriska studien i ett prototypsystem för flerspråkig generering av beskrivande texter om konstföremål. Vi utforskar de principer ett genereringssystem kan utgå ifrån för att automatiskt generera sammanhängande beskrivningar på tre språk: engelska, svenska och hebreiska. Avhandlingens fokus ligger på utforskandet av koreferensmekanismer och koreferensstrategier i de tre språken. De forskningsresultat som presenteras kommer att vara användbara för vidare utveckling av olika applikationer som har till syfte att förmedla information språkligt via t.ex. ett grafiskt gränssnitt.

I den här avhandlingen presenterar vi en kvantitativ och kvalitativ analytisk studie av domänspecifika korpusar på svenska, engelska och hebreiska, så kallade jämförbara korpusar. Varje korpus innehåller

föremålsbeskrivningar från museisamlingar och har samlats in specifikt för den aktuella studien, eftersom ingen sådan korpus såvitt bekant existerade tidigare. Vi har undersökt hur dessa föremålsbeskrivningar struktureras på de tre språken, i synnerhet hur koreferens realiseras syntaktiskt och semantiskt i vart och ett av språken. Undersökningen omfattar de syntaktiska realisationstyperna *pronominell anafor* och *full NP-anafor*, samt följande lexikalisk-semantiska relationer mellan anafor och antecedent: *högre hyperonym*, *direkt hyperonym* och *synonym*.

Vi visar att det finns både gemensamma och språkspecifika drag i de koreferensstrategier som används i de tre undersökta språken, åtminstone vad gäller den domän och den texttyp som undersökts. Genom undersökningen har språkspecifika koreferensstrategier kunnat formuleras. Dessa strategier har sedan implementerats i ett flerspråkigt genereringssystem som genererar beskrivande texter om konstverk från formella ontologiska beskrivningar av den typ som utvecklats för den semantiska webben. Genereringsystemets utveckling bygger på en modulär metod för att effektivt realisera ontologins innehåll på flera språk.

Vi genomför en utvärdering av språkstrategierna genom två undersökningar. Resultaten av utvärderingarna visar att trots att förutsättningarna för att konstruera sammanhängande texter varierar från språk till språk, kan välformade sammanhängande texter produceras även med hjälp av andra språkstrategier, specifika för något av de andra två språken, vilket antyder att skillnaderna mellan språkstrategierna snarast handlar om preferenser. Vi visar att en modulär metod lämpar sig väl för flerspråkig textgenerering från ontologier.

Acknowledgements

The compilation of this thesis would have not been possible without the help and support of my two supervisors. I wish to thank my supervisor, Lars Borin, who directed me throughout this wonderful journey, for interesting discussions on all aspects related to language technology, for giving me a broader perspective in computational linguistics and increasing my genuine interest in language. His involvement has been central in making this research come into existence. I wish to express my gratitude to my co-supervisor, Aarne Ranta who has contributed both directly and indirectly to bring this research forth, and for helping in shaping the final version of this thesis.

Thanks to Barbara Gawronska who was my co-supervisor during the first half of this thesis and who showed a lot of interest in the work.

My deepest thanks go out to Richard Power for taking on the examiner role at the final seminar and for giving this thesis a focus; Robert Dale for his guidelines and interesting discussions about my work during its different phases; Ehud Reiter, for commenting on parts of this thesis and providing valuable insights about the experiments and evaluations. I am grateful to many other people within the natural language generation community whom I met along the way and who made this research grow.

I wish to thank GSLT, the national graduate school of language technology for funding this research and for contributing with a broad academic environment. In particular I wish to thank Robin Cooper for making it all happen. Thanks to Robert Adesam for all his help with computer related problems. Many thanks to other knowledgeable individuals who were involved in GSLT during my years as a PhD student and who helped in one way or another. I wish to extend my thanks to Torbjörn Lager. I would probably not have commenced my PhD studies in language technology without him suggesting so.

Over the last years, the Centre for Language Technology (CLT) in Gothenburg has replaced the role of GSLT. I wish to thank CLT and all of its members for contributing to a high quality research environment. It is an honor to be a part of such an organization which continuously

encourages and supports research collaborations on both national and international levels.

It has been a privilege to be a part of Språkbanken during these years where many people were central to my research. I wish to extend my warmest thanks to Maria Toporowska Gronostaj who has been a close colleague for many years and who provided valuable insights from the field of lexicography on any requested occasion. I wish to thank Rudolf Rydstedt, Leif-Jöran Olsson, Markus Forsberg, Dimitrios Kokkinakis, Sofie Johansson Kokkinakis, Karin Friberg Heppin and Karin Warmenius for being great colleagues during all these years. I also wish to thank other distinguished colleagues who have been around during part of these years (in rough chronological order): Annika Kjellandsson, Lilja Øvrelid, Katarina Heimann Mühlenbock, Taraka Rama, Marin Kaså, Elena Volodina, Yvonne Adesam, Martha D. Brandt, Richard Johansson, Gerlof Bouma, and Kaarlo Volionmaa. Grateful thanks to everyone else at Språkbanken with whom I have interacted over the years.

I wish to thank all the people from the department of Swedish, in particular the teachers and researchers who contributed to this dissertation through lively seminars and lunch discussions. A special thanks to Elisabeth Engdahl who has been a source of inspiration and for her constant support and engagement right from the beginning until the very end. I also wish to thank the researchers from the institute of Swedish as a second language who I shared the same hall with during the past two years for providing a cheerful and pleasant working environment.

Part of the work presented in this thesis was done under the EU project, MOLTO at Chalmers university of technology, from which many colleagues have been helpful over the years. I wish to thank (in chronological order) Bengt Nordström, Peter Ljunglöf, Krasimir Angelov, Ramona Enache, Olga Caprotti, Thomas Hallgren, and John J. Camilleri. Also thanks to Mariana Damova from Ontotext for her distinguished research collaboration.

During these years as a PhD student, I had contact with many knowledgeable people from the museum sector. Many thanks to Marie Björk from the Gotenburg City Museum for showing me around and helping me with all relevant questions on museum data and metadata. Also thanks to Carina Sjöholm for her encouraging collaboration with the university. I wish to thank Martin Doerr for providing valuable comments on some parts of the work presented in this thesis.

I wish to thank Geoffrey Shippey for his draft readings over the years and for providing valuable input into all aspect related to English

language at all times. Thanks to Reut Tsarfaty for being a stimulating researcher and for providing valuable comments on the data annotation and analysis. I wish to thank Alex Lovinger for helping in editing the thesis and to Elad Michael Schiller for his editorial comments. Thanks to Karin Cavallin who has been helpful in miscellaneous ways. Also thanks to Benjamin Lyngfelt and Kristina Holmlid for their help while I was finishing up the thesis. Many thanks to all the anonymous reviewers who read and commented on various parts of the work.

I am grateful to many, many other people whom I met during these years while I was working towards the compilation of this thesis, but whom I have not mentioned here. It has in fact been a great journey thanks to all of these people.

Finally, I wish to thank Karin Bachar-Kaz for being such a good friend. I wish to extend my thanks and appreciation to Eva Rosenberg for her valuable encouragement during my years in Sweden and to both Malin and Amanda Dannélls for being helpful and understandable during this time. Many thanks to Robert Daniels Skagborn. Thank you, my parents, Rachel and Israel Itzhak Deutsch for showing me the right way and always support my decisions. Thank you my brothers, Yariv and Tom and my dearest sister, Maayan. My final thanks go to my husband, Kristian for his endless love and continuous support and to the love of my life, my daughter, Elinor Leah.

Dana Dannélls
Gothenburg, December 6th, 2012

CONTENTS

Abstract i

Sammanfattning iii

Acknowledgements v

1 Introduction 1
1.1 Research questions 2
1.2 Key contributions 2
1.3 Choice of languages and domain 3
1.4 Guide to remaining chapters 3

2 Background 7
2.1 Multilingual natural language generation (MLG) 7
2.1.1 Text generation from Semantic Web ontologies . . . 10
2.1.2 Coreference in text generation 15
2.2 Ontologies and the Semantic Web 19
2.2.1 Semantic Web (SW) 20
2.2.2 Description Logics (DL) 24
2.3 Computational lexical-semantic resources 25
2.3.1 Princeton WordNet 25
2.3.2 SALDO 26
2.3.3 MultiWordNet 27
2.3.4 FrameNets 27
2.4 The Grammatical Framework (GF) 28
2.4.1 Multilingual language generation in GF 29
2.4.2 Multilingual grammar example 31

3 Data collection and analysis 37
3.1 The corpus data 37
3.2 Data annotation and analysis 38
3.2.1 Syntactic processing 39
3.2.2 Semantic processing 42

3.2.3 Referential Expressions (RE) . . . 47
3.2.4 Combining semantic, syntactic and RE . . . 50
3.3 The results of the analysis . . . 56
3.3.1 Syntactic structures . . . 56
3.3.2 Discourse patterns . . . 58
3.3.3 Coreference strategies . . . 59
3.3.4 Patterns of discourse and choice of RE . . . 61
3.4 Summary . . . 62
3.4.1 Limitation of the study . . . 62
3.4.2 Implications of the study . . . 62
3.4.3 Conclusions . . . 63
3.4.4 Future work . . . 64

4 The MLG domain application 65
4.1 Overview of the system . . . 65
4.2 The application ontology . . . 65
4.2.1 The construction of the ontology . . . 67
4.2.2 Taxonomy and terminology specifications . . . 67
4.3 The abstract and concrete syntaxes . . . 71
4.3.1 The abstract syntax . . . 71
4.3.2 The concrete syntaxes . . . 73
4.4 A generation example . . . 83
4.5 Experiments and evaluation . . . 85
4.5.1 Experiment 1 . . . 85
4.5.2 Experiment 2 . . . 87
4.6 Discussion . . . 96

5 Summary and conclusions 99
5.1 Summary . . . 99
5.2 MLG using coreference strategies . . . 100
5.3 MLG from structured knowledge representations . . . 100
5.4 Future directions . . . 101

I Generating tailored texts in the context of the Semantic Web 103

6 A system architecture for conveying historical knowledge 105
6.1 Introduction . . . 105
6.2 The system architecture . . . 106
6.2.1 Pragmatic and Memory Phase . . . 106
6.2.2 Knowledge Phase . . . 107

6.2.3 Generation Phase 108
6.3 Initial results 108
6.4 Conclusion 109

7 Generating tailored texts for museum exhibits 111
7.1 Introduction 111
7.2 Background 112
7.2.1 Generating from an ontology 113
7.2.2 The CIDOC-CRM ontology 113
7.2.3 The Grammatical Framework (GF) 114
7.3 Generating from the ontology 115
7.3.1 The abstract representation 116
7.3.2 The concrete representation 117
7.3.3 The authoring environment 117
7.4 Conclusions and future work 119

II Generating cultural content through discourse strategies 121

8 The value of weights in automatically generated text structures 123
8.1 Introduction 123
8.2 Background 124
8.2.1 Semantic web ontologies 124
8.2.2 Planning the text structure from Web ontologies . . 125
8.2.3 Tailoring the content and form of the text 126
8.3 Methodology 126
8.3.1 Conveying semantic information 126
8.3.2 Tailoring the ontology content 128
8.4 Implementation 128
8.4.1 The generation machinery 128
8.4.2 Stepwise text planning 129
8.5 Evaluation 131
8.5.1 The domain ontology 131
8.5.2 Adjusting the domain properties 131
8.5.3 Experiment and result 132
8.6 Discussion 134
8.7 Conclusion and future work 135

9 Discourse generation from formal specifications using GF 137
9.1 Introduction 137

9.2 Global and local text structure . . . 138
9.3 The realities of a domain specific ontology . . . 139
9.4 From formal specifications to coherent representation . . . 141
9.4.1 Linking statements to lexical units . . . 141
9.4.2 Template specifications . . . 143
9.4.3 A discourse schema . . . 143
9.5 Domain dependent grammar-based generation . . . 144
9.6 Conclusion . . . 150

III MLG generation from SW ontologies 151

10 The production of documents from ontologies 153
10.1 Introduction . . . 153
10.2 Background . . . 154
10.2.1 Generating from ontologies . . . 154
10.2.2 Opportunities and challenges . . . 155
10.3 The domain ontology model . . . 156
10.3.1 Population and maintenance . . . 156
10.3.2 The ontology terminology . . . 157
10.4 Realization of a concept in the ontology . . . 157
10.4.1 A concept representation . . . 157
10.4.2 Surface realization . . . 159
10.5 Conclusion and future work . . . 160

11 A framework for improved access to museum databases 161
11.1 Introduction . . . 161
11.2 The ontologies and museum data . . . 162
11.2.1 The CIDOC-CRM . . . 162
11.2.2 The Swedish Open Cultural Heritage (SOCH) . . . 163
11.2.3 The Painting ontology . . . 163
11.2.4 Proton . . . 164
11.2.5 The Gothenburg City Museum (GCM) database . . . 165
11.2.6 DBpedia . . . 166
11.3 Integrating and accessing museum data . . . 166
11.3.1 Integration for flexible computing . . . 166
11.3.2 Accessing Museum Linked Data . . . 168
11.3.3 The Museum Reason-able View . . . 169
11.4 Ontologies verbalization . . . 171
11.4.1 The Grammatical Framework (GF) . . . 171
11.4.2 Translation of the Museum Reason-able View to GF 171

11.5 Related Work 174
11.6 Conclusions 175

IV FrameNet in the context of the Semantic Web and multilingual natural language generation 177

12 Applying semantic frame theory to automate templates generation 179
12.1 Introduction 179
12.1.1 Semantic frames 180
12.1.2 The language generation module 181
12.1.3 The knowledge representation 181
12.2 From ontology statements to template specifications 181
12.2.1 Lexical units' determination and frame identification 182
12.2.2 Matching the ontology concepts with frame elements 183
12.2.3 Semantic and syntactic knowledge extraction 184
12.3 Testing the method 185
12.4 Discussion and related work 185
12.5 Conclusions 187

13 Toward language independent methodology for generating descriptions 189
13.1 Introduction 189
13.2 Data collection and text analysis 190
13.2.1 Corpus data 190
13.2.2 Semantic analysis 191
13.2.3 Syntactic analysis 191
13.3 Framenets 192
13.3.1 The Berkeley FrameNet 192
13.3.2 The Swedish FrameNet 194
13.4 Multilingual language generation of museum object descriptions 195
13.4.1 The language generator tool 195
13.4.2 Linguistic realisations from framenets 197
13.5 Summary 198

V Coherent multilingual generation from the SW 199

14 Multilingual online generation from SW ontologies 201
14.1 Introduction 201

14.2 The motivation and goals 202
14.3 The Museum Reason-able View 202
14.3.1 Integrating museum data 203
14.3.2 Accessing museum linked data 204
14.4 Natural language generation 204
14.4.1 Translation of the Museum Reason-able View to GF 205
14.4.2 Discourse structures 207
14.4.3 Generation results 208
14.5 Summary and future work 208

15 On generating coherent multilingual descriptions from SW 211
15.1 Introduction 211
15.2 Related work 213
15.3 Data collection, annotations and analysis 213
15.3.1 Material 213
15.3.2 Syntactic annotation 214
15.3.3 Semantic annotation 214
15.3.4 Referential expressions annotation 215
15.3.5 Data analysis and results 216
15.3.6 The results of the analysis 219
15.4 Generating referential chains from Web ontology 220
15.4.1 Experimental data 220
15.4.2 The generation grammar 220
15.4.3 Experiments and results 222
15.5 Conclusions and future work 224

References 224

A Appendix: PoS tag sets 245
A.1 English 245
A.2 Swedish 246
A.3 Hebrew 247

B Appendix: Dependency category sets 249
B.1 English 249
B.2 Swedish 250
B.3 Hebrew 252

C Appendix: Semantic categories set 253

D Appendix: Hebrew character sets 255
D.1 Transliteration and transcription letters 255

E Appendix: The RGL categories and functions 257
E.1 Categories 257
E.2 Functions 257

1 Introduction

In the light of the substantial growth of digital content availability in large structured Web ontology standards, and today's increasingly widespread use of smart phones and small electronic devices, there is a growing need for new natural language processing (NLP) technologies that will facilitate the search and enhance accessibility to this vast amount of information in different languages automatically. One discipline of NLP that is particularity interesting in this endeavour is called Multilingual Natural Language Generation (MLG).

MLG is concerned with producing different types of information in multiple languages from some knowledge representation automatically. It uses the solutions and algorithms developed within Natural Language Generation (NLG) applications to efficiently process data and adapt the presentation of text content to a specific readership by, for example, producing paragraph-sized texts or reducing linguistic complexity in syntax and vocabulary.

Since the beginning of the twenty-first century natural language generation applications have been shifted towards Semantic Web technology. The Semantic Web offers processable structured formal representation language standards which bring several benefits to many institutes and applications on a world wide scale. Generating multilingual natural language from the representation standards offered by the Semantic Web is a relatively new research area and so far there has been little emphasis on how to exploit these existing standards in order to devise coherent multilingual texts.

This thesis is about generating written multilingual coherent, well-formed descriptions from Semantic Web representation standards by adapting linguistic knowledge and employing computational language resources. One particular aspect of coherence addressed in this thesis is the language-specific use of linguistic devices for signalling coreference, i.e. that several linguistic expressions refer to the same entity. It

is shown that there exist general principles that govern coherence in different languages and that multilingual language generators targeted towards the Semantic Web can benefit from them to efficiently produce a coherent text in multiple languages.

1.1 Research questions

The primary concern of this thesis is to work out a multilingual generation methodology that exploits the expressive power of language by adapting linguistic knowledge to produce coherent content from structured formal representations in a particular domain. We address this via the following questions:

1. How are referential forms in English, Swedish and Hebrew realized in a single domain?
2. How can a multilingual language generator access a structured formal representation, such as a Semantic Web ontology to produce well-formed chains of referential forms?

1.2 Key contributions

This thesis has two main contributions: empirical and engineering. The empirical contribution of this thesis is the comparison of coreference strategies in English, Swedish and Hebrew on the basis of three lexical-semantic relations in the domain of cultural heritage. Our investigation shows there are differences in the way chains of referential expressions are realized depending on the language considered. The linguistic knowledge gained from the empirical study brings a better understanding about how to guide coherent written discourse generation in each language.

The engineering contribution of this thesis is in presenting a text generation application which efficiently generates well-formed referential chains when manipulating non-linguistic structured representation standards using Semantic Web technology and by employing a modularized approach. The application was implemented in the framework of MOLTO to generate paragraph-sized multilingual artwork descriptions.[1]

[1] http://www.molto-project.eu/

1.3 Choice of languages and domain

The research presented here focuses on three languages: English, Swedish and Modern Hebrew (MH). English belongs to the West Germanic sub-branch of the Germanic branch of the Indo-European language family. It is a well-studied, high-resource language, spoken as first or second language by more than one billion people. It is a predominantly analytical language with a small amount of inflectional morphology and fixed word order. Swedish belongs to the North Germanic sub-branch of Germanic. It has a moderate amount of fusional and agglutinating inflectional morphology and mainly fixed word order (although less fixed than English). The language has about nine million speakers. Modern Hebrew is a Semitic language spoken by about seven million people. The language has a non-concatenative core (inflectional and derivational) morphology based on consonantal roots, combined with a system of agglutinative prefixes and suffixes. The word order is free. The Hebrew alphabet uses the Hebrew script *alefbet* and is written from right to left. Hebrew is the author's native language and is integrated here to gain important insights that will hopefully be applicable to other major related languages such as Arabic.

The domain this thesis explores is the cultural heritage (CH). What makes the CH domain particularly suitable to explore is the accessibility to well-developed structured representation standards, which although are not structured for either natural language generation or natural language processing, introduce a wide typology of labels to allow recording a mixture of data from different cultural collections. Because a large number of heterogeneous digital collections and other cultural heritage material are accessible through these standards, the requirements imposed on the traditional methods for presenting collections of historical and cultural data in multiple languages are increasing.

1.4 Guide to remaining chapters

This thesis consists of two major sections: the first section contains five chapters, the second section contains five parts.

Chapter 2: Background provides the background knowledge and related work on multilingual natural language generation. We elaborate the notions of coreference and Semantic Wseb ontologies. We describe the computational lexical resources and the grammatical formalism GF, which is employed in this work.

Chapter 3: Data collection and analysis describes the primary data we collected from the cultural heritage domain in order to acquire linguistic knowledge about how coreference is realized in a discourse. It specifies how the data was processed and analyzed. We further describe the results of the analysis and summarize the domain-dependent discourse patterns and the language-specific coreference strategies that follow on from the analysis.

Chapter 4: The MLG domain application presents the application ontology and text generation system. We provide a detailed description of how coreference strategies are modularized in the system and demonstrate how it successfully generate coherent descriptions in all three languages. This chapter also describes the experiments that were carried out to test whether language-specific coreference strategies enhance the output of a multilingual language generator.

Chapter 5: Conclusion summarizes the thesis's main contributions and provides pointers to other research directions that are interesting to explore further.

The remaining chapters of this thesis encompass a selected set of peer-reviewed publications. The typography and layout of the publications have been adapted to adhere to the stylesheet of this thesis, but content-wise they remain unchanged from the original papers. They are structured into five parts:

Part I: Generating tailored texts in the context of the Semantic Web introduces a system for generating object descriptions in the context of the Semantic Web and explores how this system can be adapted to generate text contents to a specific readership.

Part II: Generating cultural content through discourse strategies demonstrates how to generate comprehensible multilingual texts from formal representations by embodying discourse strategies in GF.

Part III: Multilingual language generation from SW ontologies addresses some of the difficulties that are involved in managing and accessing Semantic Web data in order to support reader and listener preferences.

Part IV: FrameNet in the context of the Semantic Web and Multilingual Language Generation investigates how semantic and syntactic information such as that provided in a framenet can contribute to multilingual text generation.

Part V: Coherent multilingual generation from the SW deals with multilingual Web and Web applications that employ Semantic Web ontologies for generating coherent multilingual natural language descriptions about museum objects.

Three of the 10 published papers reproduced in part I–V are co-

authored. In these, the contributions of the present author are as follows:

In chapter 11 (Dannélls et al. 2011), the author contributed with the implementation and description of the painting ontology; the analysis and description of the Gothenburg City Museum database; part of the grammar implementation; writing and editing the paper.

In chapter 13 (Dannélls and Borin 2012), the author contributed with the semantic and the syntactic analyses; the grammar implementation; writing the paper.

In chapter 14 (Dannélls et al. 2012), the author contributed with the translation of the Museum Reason-able View to GF; the ideas about optimizing the grammar with discourse structures; writing and editing the paper.

2 BACKGROUND

This chapter presents some background knowledge on multilingual natural language generation, Semantic Web ontologies, the semantic-lexical resources, and on the grammatical formalism GF, which is employed in this work.

2.1 Multilingual natural language generation (MLG)

Natural Language Generation is the field concerned with building computer software systems, which can map from some underlying, non-linguistic representation of information into a linguistic presentation of that information, whether textual or spoken. The main tasks involved in the process of NLG are to determine what information to extract from some Knowledge Representation (KR) system, impose a suitable order on the elements of this information and make linguistic choices to express this information in natural language that humans understand (Reiter and Dale 2000).

Researches often characterize NLG as a sub-field of artificial intelligence (AI) and computational linguistics (CL). This is not surprising because one of the principle emphasis of natural language generation is to employ AI solutions such as developing intelligent systems, which are capable of making clever decisions based on observations about human language abilities (Paris, Swartout and Mann 1991). The computational linguist aspect of this field is to take advantage of existing machine readable language resources and linguistic knowledge to produce unambiguous natural language that meets the communicative goals of different users depending on their age, language of preference, level of expertise, knowledge of the world, etc.

NLG is considered the inverse of Natural Language Understanding (NLU) (Jurafsky and Martin 2008). Because NLU starts from linguistic output and NLG from non-linguistics one the problems each of

these fields must deal with are very different, although they both try to resolve similar tasks such as summarisation and simplification of texts (Sripada et al. 2003; Murray, Carenini and Ng 2010; Siddharthan 2011). Disambiguation is one distinguishing problem in these endeavors. For example, while NLU must resolve anaphoric references, i.e. finding the entity of a reference in the previous discourse, NLG needs to make linguistic choices to produce unambiguous references of entities mentioned in the discourse.

A further specialization of NLG is Multilingual Language Generation (MLG); the discipline that approaches text production in multiple languages. Many researchers consider MLG as an alternative approach to machine translation (MT) with the capacity of yielding high-quality output texts (Power and Scott 1998). This is because MLG has the advantage of starting from some kind of a knowledge representation system and thereby avoids disambiguation difficulties which often arisen when generating from some source natural language. For example, in the WYSIWYM generation system (Power, Scott and Evans 1998), the generator switches between languages and avoids ambiguities by keeping the semantic meaning of the expression, for example: *generate(proc1, english, feedback)*, *generate(proc1, french, feedback)*.

Most of the applied NLG applications claim to follow a three stage, one-way pipe line model comprised of separate modules (Mellish et al. 2006). The three-module chain architecture, as illustrated in figure 1 has been devised by Reiter and Dale (2000). This widely accepted view of the generation processes is also adopted in this thesis.

As figure 1 portrays, the task of generating a text comprises three sub-tasks: (1) selecting the information the text should convey depending on the purpose of the text to be generated; (2) deciding how to order this information to allow linguistic realization in the target language; (3) choosing the linguistic structures to communicate this infomation to the user based on his/her knowledge. A vast number of computational approaches have been suggested for dealing with each of these tasks, some of which have been particularly influential in the context of the Semantic Web (Wilcock 2003; Chiarcos and Stede 2004; Bontcheva and Wilks 2004; Bontcheva 2005; Isard 2007; Mellish and Sun 2006a, b; Mellish and Pan 2008; Kelly, Copestake and Karamanis 2009; Power 2010; Mellish 2010).

NLG applications are usually built from a computer user perspective, more specifically, the target audience to which the text will be generated. Early work on NLG focused on building applications that are targeted towards domain experts (Goldberg, Driedger and Kittredge

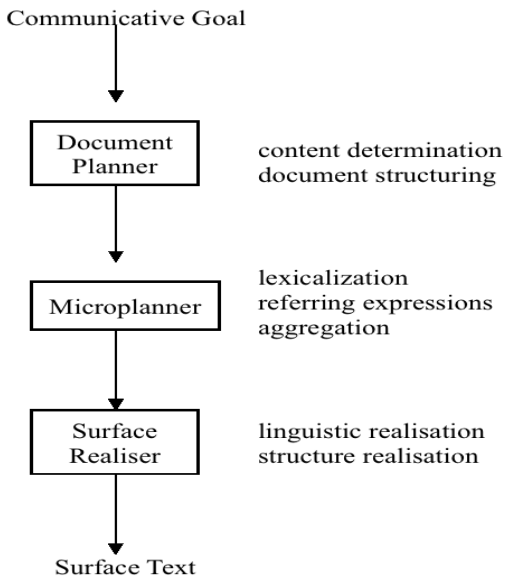


Figure 1: NLG pipeline architecture according to Reiter and Dale (2000).

1994; Power, Scott and Evans 1998) and layman computer users (Reiter, Robertson and Osman 2003). The major difference between these groups is manifested in terminology, syntax and the level of details in the generated output. In this thesis we are mainly concerned with layman user requirements. We rely on the principles drawn from previous studies of cultural heritage (Komsell and Melén 2007; Clough, Marlow and Ireson 2008).

Until the beginning of the 21th century the form of the internal data representations provided as input to a language generator, i.e. the information about the domain, varied from one source to another. Local relational databases have been typical inputs to language generators (Dale et al. 1998; Dannélls 2010a). However, along with the appearance of Semantic Web languages things have started to change. Today there exist formal representation standards (section 2.2) that are becoming increasingly attractive for NLG and in particular for MLG especially because they provide common formalism to generate from, regardless of the domain (Hielkema, Mellish and Edwards 2008).

In a way, the high-level KR provided by Semantic Web technologies

is very similar to other high-level KR employed by early generation systems. For example, the data models employed by Cahill et al. (2001) and Mann (1983) are comprised of similar components to the ones we find in Semantic Web ontologies, i.e. they contain entities, attributes, relationships and classes organized in a hierarchical taxonomy. The distinguished characteristic between these representations is the language formalism used for storing data.

The Semantic Web standard representations that have been explored during the last decade are in the form of *triples* (section 2.2.1). An example of a data representation in this form is:

```
<owl:Thing rdf:about="&painting;Guernica">
 <rdf:type rdf:resource="&painting;OilPainting"/>
 <createdBy rdf:resource="&painting;PabloPicasso"/>
<owl:Thing/>
```

A common term in NLG for describing this type of specification that characterizes the domain is *message*. More specifically, it is a specification of the information that has to be communicated to the hearer/reader and may correspond to a word, a phrase or a sentence in natural language. In the context of the Semantic Web, a message corresponds to an ontology statement, or a set of statements. In the above example there are two statements: one indicating the type of the object *<Guernica rdf:type OilPainting>*, and one indicating the creator of the object *<Guernica createdBy PabloPicasso>*.

2.1.1 Text generation from Semantic Web ontologies

Generating natural language from Semantic Web ontologies implies finding a way to bridge Semantic Web data structures, such as formal ontologies expressed in Resource Description Framework (RDF) or Web Ontology Language (OWL) (section 2.2.1), with coherent (but ontologically unstructured) texts written by humans, see chapter 10 in this thesis. Meteer (1990) argues that generation components as a whole should follow two central principles: (1) expressibility, i.e. the input representation should always allow realization in natural language, and (2) efficiency, i.e. the algorithm itself must be linear. These principles apply in particular to systems that are targeted towards the Semantic Web.

During the last decade there has been an increasing interest in developing natural language generators that support Semantic Web on-

tology languages such as OWL (Schwitter and Tilbrook 2004; Mellish and Sun 2006a; Mellish and Pan 2008; De Coi et al. 2009; Williams, Third and Power 2011). This increase appears to be motivated by the potential information access to distributed ontology models, the high level semantic specification and the 'common-ground' input representation to generate from.[2]

Wilcock (2003) and Wilcock and Jokinen (2003) presented an XML-based NLG and show that direct verbalization of the concepts represented in a domain specific ontology is not a promising endeavour. According to their approach, XML transformations are performed on text plan trees in order to produce text specification trees using Extensible Stylesheet Language Transformations (XSLT) which implies that text planning is embedded in the templates.

In the same vein, Bontcheva and Wilks (2004) presented the template based MIAKT system that supports a lexicon and uses an ontology-based aggregation strategy to reduce definite noun phrases. Simple aggregation is carried out at discourse level by joining RDF statements that have the same first argument and the same property name or if they are sub-properties of *attribute* or *part-whole* properties. The authors have demonstrated the usefulness of performing aggregation and applying some kind of discourse structures in the early stages of the microplanning process.

The ONTOSUM system (Bontcheva 2005) is an extended version of MIAKT; more oriented towards the user (in terms of length and format) and is less restricted to the ontology structure to increase portability. The system is implemented as a set of components in the GATE infrastructure and aims to generate summaries from a set of statements being given in the form of RDF/OWL. Statements are processed without any modifications, the only pre-processing task is to remove repetitive statements that have the same property and arguments. In addition, the system also removes statements containing inverse properties that share the same arguments. Summary structuring is done with the help of a set of pre-defined discourse schema.

Mellish and Pan (2008) experimented with knowledge represented in OWL. They focused on the problem of selecting the relevant material for inclusion into the final natural language output of an NLG system. Their work is different from previous approaches in that it verbalizes the ontology class axioms. Mellish and Pan (2008) argued that although

[2] As it turns out, generation results from independent surface realizers are usually not directly comparable because of the differences in the input representations.

most of the available ontologies contain some linguistic information, Web ontology representations are not adequate for generating texts in natural language. They distinguish between top-down and bottom-up methodologies for content determination and argue that text coherence plays an important role in this kind of formal logic knowledge-base. In their view, linguistic complexity does not necessarily mirror the complexity of the underlying logical formula; that complexity may very well depend on the mapping between logical formulas, the surface realization and the underlying linguistic resources that are available for the system. To obtain basic knowledge of how text coherence is manifested in a domain, it is necessary to study discourse structures that are commonly used in naturally occurring texts within this domain.

The work presented by Mellish and Pan (2008) is similar to work by other researchers who pioneered natural language generation from the perspective of Controlled Natural Language (CNL) (Fuchs and Schwitter 1996; Schwitter and Tilbrook 2004). These approaches focus on verbalization rather than on generation, with emphasis on the English language. When verbalizing a web ontology, sentences are formed on the basis of the logical patterns of this ontology. Recent work found that verbalization of this kind depends on the exploitation of a consensus model to allow adequate natural language generation (Power 2010).

Earlier work on NLG from Semantic Web ontologies applied verbalization methods to realize ontology statements in natural language (Wilcock 2003; Bontcheva and Wilks 2004; Mellish and Sun 2006a). These methods often take one statement – one sentence approach, and assume each ontology statement is realizable in one sentence.

While the majority of generation applications have been developed for English, comparatively small number of studies have been conducted to explore their applicability to other languages. The only multilingual systems we are familiar with in the context of the SW are (ILEX) (O'Donnell et al. 2001), M-PIRO (Androutsopoulos et al. 2001) and NaturalOWL (Androutsopoulos, Kallonis and Karkaletsis 2005; Galanis and Androutsopoulos 2007).

The Intelligent Labelling Explorer (ILEX) (O'Donnell et al. 2001) is an example of a system that has been developed to generate natural language descriptions about artifacts from a dynamic structured representation environment. The system is capable of generating domain-dependent descriptions in a hypermedia environment. Its components exploit the fact than an RDF graph can be made to correspond to the structure of a coherent text. Bellow follows an example of a description generated by the ILEX system, presented by O'Donnell et al. (2001).

> **This jewel** is a necklace and was made by a British designer called Edward Spencer. **It** is in the Arts and Crafts style and was made in 1905. **It** is set with jewels. **It** features rounded stones.

In ILEX, the user selects an object from the ontology, for example, by clicking on a thumbnail image in a web museum. The system then uses a content selection algorithm based on the *interest scores* stored in the ontology, and the user's previous browsing history to choose the content the text should convey. The *interest scores* have previously been assigned by experts in the domain, and these scores may differ between different user types, e.g. adult, expert or child. The microplanning process of the text structure is comprised of four steps and is organized via rhetorical relations.

We could not find any information about how exactly lexical units are chosen to improve the coherence of the text.[3] From the above example of the output produced by the system, it is understood that no decisions regarding the use of referential expression are made by the linguistic realizer, because of three consecutive pronouns (marked with bold). One of the drawbacks of the ILEX system is that it requires an extensive amount of hard coded linguistic knowledge for each defined concept and property. While such a manual process is often necessary and important from linguistic point of view, it should ultimately be automated, or at least draw upon general linguistic resources.

M-PIRO is a source authoring generation system that produces personalized descriptions in English and Greek in the domain of art. It employs templates in a similar fashion as the ILEX system (O'Donnell et al. 2001) but extends the ILEX's personalization mechanism. The order of the facts (triples) conveyed in the output text are specified by the user explicitly, these are constrained by a fixed fact order for each user type. There are two types of referring expressions associated with each triple: pronoun (personal/demonstrative) and noun (genitive form/-full noun phrase); similar to ILEX, it is the user who chooses the type of referring expression by indicating its form explicitly. If no referring expression has been indicated by the user, the system, which is guided by hand-crafted rules, will choose between a personal pronoun or a full noun phrase. To our knowledge, the system does not differentiate between the languages regarding the choice of referring expression. The same procedure is applied regardless the output language.

[3] Throughout this thesis we use the expression *lexical unit* to refer to a lexical form together with a single distinguished sense.

NaturalOWL is a multilingual natural language generation system that has adapted many of the ideas from ILEX and M-PIRO to generate multilingual descriptions from Semantic Web ontology languages. In NaturalOWL, linguistic information including referential expression units are encoded in the ontology. There is a set of candidate referring expressions (a noun phase, personal and demonstrative pronouns) assigned to each ontology statement. This set is similar in all languages. An appropriate candidate is rendered regardless the language by employing a simple algorithm that builds on the Centering Theory (Grosz, Weinstein and Joshi 1995). An example of a generated description:

> This is a vessel. **It** is sculpted by Nikolaou. **Nikolaou** was born in Athens. **He** was born in 1918 and he died in 1998. **This vessel** is not exhibited in the National Gallery. **It** is one of the best ..

According to this example, a demonstrative pronoun is chosen to represent the inanimate main entity, *vessel* in the beginning of the description and when the focus of the entity has shifted.

As the two examples above demonstrate, texts generated from structured formal representations often contain chains of different linguistic elements that refer to the main subject entity. Cross-linguistic investigations into how coreference is expressed have shown that these chains bear language specific characteristics (section 2.1.2.1), and that theories formulated on the basis of English, such as Centering Theory, must be further specified and adapted to the language in question. Yet, none of the reviewed systems differentiate between the generation of referential expression elements depending on the language considered.

In summary, most of the researchers who have dealt with Semantic Web languages aim at domain independent solutions and focus on the semantics of the ontology rather than on the syntactic form of the language in combination with semantic knowledge. The reviewed systems are based on templates; they employ direct verbalization that is close to the ontology structure; there is no indication of how adaptable these approaches are to languages other than English. Despite the growing need to develop text generation systems/components that are capable of producing texts from the same knowledge source in more than one language, most generation approaches remain monolingual. Our literature survey shows there has been very little work on extensible multilingual language generation that seeks an architecture within which the work involved in adding a new language may be minimized.

2.1.2 Coreference in text generation

Coreference (or reference) is a linguistic phenomenon which implies there are two or more occurrences of lexical units that follow each other in a sentence or discourse. The term *anaphoric expression* or *referential expression* is usually used to describe this phenomenon. The first entity mentioned in the sentence or the discourse is often a proper noun, a noun or a noun phrase that refers to some entity in the external world. In linguistics, the term is often called the *antecedent*. In works on NLG, the terms *Main Subject Entity (MSE)* and *Center* are sometimes used. In the following discourse the center is *Girl Before A Mirror* and the referential expressions are: *This painting*, *It*, and *The work*.[4]

> 'Girl Before A Mirror' by Pablo Picasso. **This painting** was painted in March 1932. **It** was produced in the style Picasso was using at the time and evoked an image of Vanity such as had been utilized in art in earlier eras, though Picasso shifts the emphasis and creates a very different view of the image. **The work** is considered in terms of the erotic in Picasso's art.

The semantic and syntactic realizations of the above referential expressions (in bold) are depicted in figure 2.

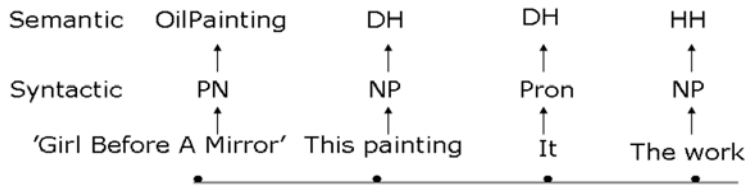


Figure 2: Coreference realization in a discourse.

As figure 2 exemplifies, discourses contain chains of referential expressions that bear both semantic and syntactic characteristics. Some of the lexical-semantic relations that may articulate the relation between a referential expression and an antecedent are: *hyponym*, i.e. the relation between a specific and a more general concept, for example oil painting is a hyponym of painting; *hyperonym*, also called superordinate, is the relation between a more general concept and a specific concept. It can be described in terms of *direct-hyperonym (DH)* and *higher-*

[4]The discourse example is taken from:
<http://www.pablopicasso.org/girl-before-mirror.jsp> (Last accessed: 2012-10-28)

hyperonym (HH), as seen in figure 2. For example: painting is a direct hyperonym of oil painting, artwork is a higher hyperonym of oil painting; *co-hyponyms*, i.e. lexical units which have the same hyperonym, for example, self-portrait and group-portrait; and *synonym*, i.e. the relation between two or more concepts which have the same meaning such as painting and picture. The linguistic elements that may express referential expressions include *gaps* (also called empty categories), *personal pronouns* (it, he, she), *demonstrative pronouns* (this), and *definite noun phrases* (the painting).

In NLG, most of activities involving generation of referential expressions have mainly focused on the syntactic realization of the referential expression (Gatt, Belz and Kow 2008, 2009; Belz and Kow 2010). Perhaps the most influential work for referential expression generation algorithms is the one by Dale and Reiter (1995) and Passonneau (1996). Work in the same lines has been carried out by many other researchers (McCoy and Strube 1999; van Deemter 2002; Krahmer and Mariet 2002; Krahmer, van Erk and Verleg 2003; Paraboni, van Deemter and Masthoff 2007; Croitoru and van Deemter 2007; Dale and Viethen 2009). A comprehensive survey of recent referring expression generation algorithms that have been proposed during the last two decades is found in Krahmer and van Deemter 2012.

In this work we do not try to re-implement any of the existing algorithms which we believe are computationally too expensive in the context of the Semantic Web. Instead, this work is concerned with establishing a modularized approach for generating chains of referential expressions by focusing on three lexical-semantic relations: direct-hyperonymy, higher-hyperonymy, and synonymy.

2.1.2.1 Discourse coherence theories

The notion of coherence. In linguistic literature there has been a lot of discussion about the notion of *coherence*. The term is typically understood as the phenomenon that contributes to the reader's understanding of a *discourse*. It is "a coherent sequence of utterances which together conveys a meassage to the addressee" (Halliday and Hasan 1976).

According to Halliday and Hasan (1976), coherence describes meaning relations between different parts of a text, such as paragraphs, sentences, clauses and is signaled by lexical choice and other linguistic cues. It can be divided into two types: *grammatical* and *lexical*. *Grammatical coherence* concerns the ways in which phrases and sentences are

related to each other and to other cohesive devices to create continuity in a text, such as ellipsis and conjunction. *Lexical coherence* concerns the ways in which lexical items are related to each other and to other cohesive devices in order to create continuity in a text. It may bear a semantic characteristic such as hyponymy and synonymy (Cruse 1986).

The notion of coherence presented by Halliday and Hasan (1976) was studied theoretically by many researchers and formulated on the basis of English (Prince 1981; Givón 1983; Grosz, Weinstein and Joshi 1995). These theories propose establishing a coherent description by replacing the lexical units pointing back to main subject entity with a pronoun or a full noun phrase. One of the theories that has been highly influential in NLG is the *Centering Theory* (CT) (Grosz, Weinstein and Joshi 1995).

The Centering Theory describes lexical cohesion by defining centering transition states which are intended to account for the local discourse coherence. It is formalized as a system of three transition relations: *continuation, retaining, and shifting* and constraints for describing how utterances (U) are linked to other utterances. According to the Centering Theory, each utterance is associated with a set of discourse entities called *Forward looking centers (C_f)* a subset of these entities have *Backward looking centers (C_b)*. The Backward looking center of U_{n+1} connects with one of the forward looking centers of U_n.

A *continuation transition* is established when an entity occurring in a subject position in an utterance provides the basis for continuing talking about that entity in the following discourse, for example by replacing it with a pronoun in the following utterance. As the example from Grosz, Weinstein and Joshi (1995) shows:

(1) **Susan** is a fine friend.
(2) **She** gives people the most wonderful presents.
(3) **She** just gave Betsy a wonderful bottle of wine.
(4) **She** told her it was quite rare. (Sudan told Betsy)

In this work the focus of attention is always on the entity mentioned in the beginning of a description and appears in subject positions in the following utterances of the discourse, therefore, the transition relation that is of interest is *continuation*. The constraint Grosz, Weinstein and Joshi (1995) put on the continuation transition is:

> If any element of $C_f(U_n)$ is realized by a pronoun in U_{n+1}, then the $C_b(U_{n+l})$ must be realized by a pronoun also.

By the first constraint of Centering Theory, if the center is realized by a pronoun, then its following referential expression must be realized by a pronoun as well. Following on the example from the previous section, according to the CT theory, the definite noun phrase *the work* will be replaced with a pronoun, as illustrated below.

> 'Girl Before A Mirror' by Pablo Picasso. **This painting** was painted in March 1932. **It** was produced in the style Picasso was using at the time and evoked an image of Vanity such as had been utilized in art in earlier eras, though Picasso shifts the emphasis and creates a very different view of the image. **It** is considered in terms of the erotic in Picasso's art.

The Centering Theory does not make any predictions about the type of linguistic expression different entities may bear in different languages. The theory has paved the way for further cross-linguistic research on the types of linguistic expressions different entities may bear in different languages or domains (Hobbs 1979; Givón 1983; Hein 1985; Ariel 1988; Hein 1989; Ariel 1990; Prince 1992; Vallduví and Engdahl 1996). These investigations into how coreference is expressed have showed that coherence depends on the target language and the domain.

Work on coreference in English and Swedish has showed referential expressions can vary depending on the language and the context and that typical semantic relations between an anaphoric noun expression and its antecedent are synonyms, hyperonyms or hyponyms (Hobbs 1979; Hein 1985, 1989). Ariel (1988, 1990) found that Hebrew, among other languages, has higher occurrences of zero anaphora and that other linguistic units, which are marked for gender, number and person, are equality informative as pronouns in cases of continuation.

The Centering Theory has also been a source for studying the usage of referring expressions in distinct languages. Yeh and Mellish (1997) established possible rules for generating referential expressions in Chinese and implemented these rules to generate descriptive texts, which were evaluated by native speakers of Chinese. Prasad (2003) studied the usage of referring expressions from the Centering theory point of view and developed an algorithm to generate referential chains in Hindi. Similar approaches for characterizing referential expressions have been proposed and implemented in Japanese (Walker, Cote and Iida 1996),

Italian (Di Eugenio 1998), Catalan and Spanish (Potau 2008), and Romanian (Harabagiu and Maiorano 2000). Other researchers have showed certain languages disallow the use of 'topic-continuing' pronoun *it* (Prince 1994).

The primary source for studying coherence by the majority of linguistics researchers mentioned above has been written and spoken outputs of native writers of the language in arbitrary domains.

2.2 Ontologies and the Semantic Web

The concept *ontology* is defined differently in different disciplines. In philosophy, the notion of ontology goes back to Aristotle who investigated questions like: "What characterizes being? What is being?"

The notion of ontology that is articulated in this thesis is the one adapted from the discipline of AI and computer science. According to these disciplines Ontology has the following definition:

> An ontology is an explicit specification of concepts and the relations among them. It provides a shared understanding of some domain of interest and a formal vocabulary for information exchange (Schalley and Zaefferer 2007).

Several types of ontologies are mentioned in the literature (Gruber 1995; Uschold and King 1995; Studer, Benjamins and Fensel. 1998; Guarino 1998; Lassila and McGuinness 2001; Gómez-Pérez, Fernández-López and Corcho 2004; Hovy 2005). Here, we only mention four types relevant for this thesis; all have been considered and discussed by the above authors.

(1) Top level ontologies express very basic knowledge. They are general ontologies to which other ontologies can be connected, directly or indirectly. Examples of ontologies that fall into this type are Suggested Upper Merged Ontology (SUMO), Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE).

(2) Domain ontologies describe the conceptualization of particular domains, and are useable within that domain. An example is the CIDOC-CRM.

(3) Linguistic ontologies describe syntactic and lexical representations whose objective is linguistic realizations. An example is the Suggested Generalized Upper Merged Ontology (GUM).

(4) Application ontologies describe concepts that are often a specialization of several domain ontologies. An example is the Wine ontology.

The ontologies introduced by different researchers can be classified into two types: lightweight and heavyweight (Studer, Benjamins and Fensel. 1998). Lightweight ontologies contain concepts, relations, taxonomies and properties. Heavyweight ontologies contain in addition axioms and constraints. Both types are suitable to be used by computer applications to serve various needs such as data storage, exchange, reasoning, etc. (Staab and Studer 2004).

For an ontology to be used within an application it must be represented in an appropriate knowledge representation language, that is, a representation suitable for computer processing. Sowa (2000) specifies several principles that a knowledge representation for natural language systems should follow: (1) It should be able to provide answers to questions within a domain. (2) It should be flexible and allow pragmatically efficient computation. (3) The specification language should be designed for the Semantic Web. (4) There should be support tools that are able to descibe it.

Ontology-based applications which combine software modules with possibilities for reuse and future extensibility requires common ontology languages and formats. The Semantic Web ontology languages specified by the World Wide Web Consortium (W3C) provide exactly the logical formalism needed to allow deployment of knowledge representation in the Semantic Web. In the next section we elaborate the notion of SW and motivate its importance for computational linguistics.

2.2.1 Semantic Web (SW)

> The Semantic Web is an extension of the current Web in which information is given well defined and explicit meaning. (Berners-Lee, Hendler and Lassila 2001).

In the context of SW, extension of the current Web is achieved through *semantic markup* associated with *resources*. Resource is the name of the basic reference mechanism of the Semantic Web that is denoted by Uniform Reasource Identifiers (URIs). Semantic markup (or metadata) allow to represent information in a form that can be processed by computer. Using this technology, anyone can contribute to our knowledge by sharing information that can be manged in a meaningful way. To illustrate the idea behind the SW, let me cite Allemang and Hendler (2008):

> The main idea of the SW is to support a distributed Web at the level of the data rather than at the level of the presentation. Instead of having one webpage to point to another, one data item can point to another, using global references called Uniform Reasource Identifiers (URIs). The SW infrastructure provides a data model whereby information about a single entity can be distributed over the Web. The single coherent data model of an application is not held inside the application but rather is part of the Semantic Web infrastructure.

In the following, we present the annotations and data models that have been introduced by the World Wide Web Consortium to allow reuse and sharing on an WWW scale.

2.2.1.1 *Modeling language standards*

The World Wide Web Consortium has developed a number of modeling language standards that use the idea of class hierarchy for representing commonality and support the representation and use of metadata. These modeling languages provide the basic machinery that we can use to represent the extra semantic information in the Semantic Web. These are distinguished by three properties: *annotations* to associate metadata with resources; *integration* for combining information sources; and *inference* for reasoning over the information. The modeling languages are built on top of one another and differ in their level of expressivity. Figure 3 depicts the collection of languages and their properties defined by the W3C.

Below follows a brief introduction of the development process of this collection of languages. Thereafter we introduce the formal language for representing knowledge that underlies the Semantic Web modeling languages and provide an extended description of the Web Ontology Language OWL and its subsets.

2.2.1.2 *Semantic Web markup languages*

The **Extensible Markup Language (XML)** is a user definable and domain specific markup language (Bray et al. 2008). The syntax structure of XML makes it possible to specify data and meta-information about data that describes it in the form of a labelled tree, where every node in the tree consists of a label, attribute/value, and content.

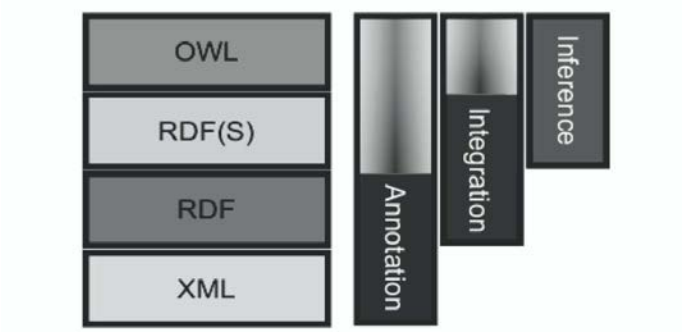


Figure 3: The languages defined by the W3C.

The **Resource Description Framework (RDF)** is a language for representing information about *resources* in the World Wide Web, by providing metadata about Web entities (Lassila and Swick 1999). It is the data model that the semantic web infrastructure uses to represent distributed web of data which is based on the idea of identifying things using Web URIs. RDF has an XML syntax consisting of constructs of the shape `<subject, predicate, object>`, in the SW this shape is referred to as a *triple* or *triple statement*.[5] A triple or set of triples can be seen as forming a graph, consisting of nodes and edges. The nodes correspond to either a resource or a value, and the edges correspond to the relationships between these nodes. The labels on the nodes and edges are URIs. An example of the RDF triple: `<Guernica,createdBy, PabloPicasso>` presented in this form is illustrated in figure 4.

RDF Schema (RDFS) is the name of the vocabulary, which extends RDF by defining the primitives for creating ontologies. It includes constructs of the type: `rdfs:Class`, `rdfs:subClassOf`, etc. with which relationships can be defined among classes. An example of such triple is: `<OilPainting, rdfs:subClassOf, Painting>`. RDFS also includes the constructs *domain* and *range* to describe the relationship between properties and classes. The domain of a relation is a set of objects for which it is defined, and the range is the set of values it can take.

Web Ontology Language (OWL) is an ontology language that is richer than XML, RDF, and RDFS. The language provides additional vocabulary along with a formal semantics (McGuinness and van Harme-

[5]The W3C terminology *subject*, *predicate*, and *object* should not be confused with linguistic terms.

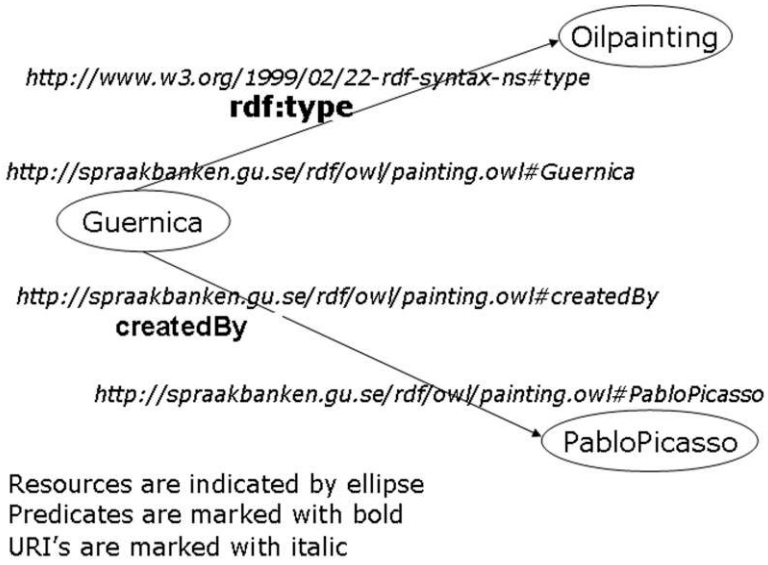


Figure 4: An RDF graph.

len 2004). It comes with three syntaxes: XML syntax,[6] the Abstract syntax,[7] and RDF.[8] It allows boolean combinations of class expressions: `owl:unionOf`, `owl:intersectionOf`, etc. The intersection constructor allows intersections of named classes and restrictions.

The basis for the Web Ontology Language, OWL design is description logics (Berners-Lee 1998). The logic on which OWL is based includes features such as role inclusion axioms ($\mathcal{H}$), nominals ($\mathcal{O}$), inverse role ($\mathcal{I}$) and number restrictions ($\mathcal{Q}$, if quantified, $\mathcal{N}$ otherwise). The language became a W3C recommendation in 2004 (Bechhofer et al. 2004).

OWL has two versions: OWL 1 and OWL 2. **OWL 1** has three increasingly expressive sublanguages: OWL Lite, OWL DL, and OWL Full (McGuinness and van Harmelen 2004). Of these three, OWL-DL is the most expressive which is still decidable language. **OWL 2** is a revision of OWL 1 to which new functionalities have been added (W3C 2009). All the OWL 2 profiles allow pruning that is a modeling approach that provides support for encoding domain independent and domain dependent information which allows some filtering that is in particularly meaningful for homogeneous knowledge bases. For ex-

[6] <http://www.w3.org/TR/xmlschema-2/> (Last accessed: 2012-12-12)
[7] <http://www.w3.org/TR/owl-semantics/> (Last accessed: 2012-12-12)
[8] <http://www.w3.org/TR/owl-ref/> (Last accessed: 2012-12-12)

ample, it is possible to declear *Painter* as a class and an individual without risking that this modeling will lead to undecidability. Other fuctionalities in OWL 2 and its sublanguages include: keys, property chains, richer datatypes and data ranges, qualified cardinality restrictions, asymmetric, reflexive and disjoint properties, enhanced annotation capabilities.

It is too early to determine which of these ontology languages is the most suitable for natural language processing, in particular for natural language generation. The ontology language developers themselves have not yet reached a consensus on which features each language must have, or which syntax is most appropriate for expressing knowledge on the Web. However, given that existing ontology Web languages have been successful and are widely used in practice by many applications in several fields, it is necessary to experiment with these emerging languages.

2.2.2 Description Logics (DL)

Description Logics (DL) is a family of knowledge representation formalisms that provide the formal logical base semantics underpinning the W3C ontology languages described in the previous section.[9] It provides a well-deifned semantics on different levels which supports the definition and use of subsumption hierarchies. It is a natural logic framework in which formal descriptions of classes, individuals and the relationships between them can be made, i.e. a set of individuals with common properties can be defined and ordered under subsumption, and assertions can be made about properties of and relations between individuals. Further details about the DL languages can be found in the appendix on DL terminology in Baader et al. (2003).

A knowledge representation system based on Description Logics is made up of two parts, a terminological part, called *TBox* and an assertional part, called *ABox*. The TBox defines concepts and ABox states facts belonging to these concepts. "The roles of TBox and ABox are motivated by the need to distinguish general knowledge about the domain of interest from specific knowledge about individuals characterizing a specific world or situation under consideration" (Nardi and Brachman 2003). In such system concepts can be checked for *(un)satisfiability* and *(in)consistency*. Satisfiability checks allow knowledge designers to check

[9] See the description logics community <http://dl.kr.org/> (Last accessed: 2012-12-12)

Table 2.1: Conventional DL notation (Baader et al. 2003).

DL syntax	**Abstract Syntax**
C	Class(C)
$C \sqsubseteq D$	subClassOf(C D)
$C \equiv D$	EquivalentClasses(C D)
$C \sqcap D$	IntersectionOf(C D)
$\exists$R.C	SomeValuesFrom(R C)
$R \sqsubseteq S$	SubPropertyOf(R S)
≥ 1 R $\sqsubseteq$ C	Property(R domain (C))
$\top \sqsubseteq \forall$R.C	Property(R range (C))
R:a	ObjectHasValue(R a)
C(a)	ClassAssertion(C a)
R(a b)	PropertyAssertion(R a b)

that a domain model is correct and that the expected subsumption relationships hold. Consistency checks can be applied to sets of assertions (ABox) to find out whether a particular individual is an instance of a given concept.

In chapter 4 we use a Description Logic formalism to describe the painting ontology following notational conventions according to the syntax given in table 2.1, where *C* and *D* denote concepts. *R* and *S* denote roles. Individuals are denoted by the lower-case letters *a* and *b*. $\forall$R.C and $\exists$R.C denote universal and existential quantification.

2.3 Computational lexical-semantic resources

This section provides background information about the computational lexical-semantic resources that are employed in some parts of the thesis (see chapters 12, 13 and 15). The resources have been employed to built the application ontology and to assign linguistic knowledge to the ontology content. They are also exploited by the generation grammar.

2.3.1 Princeton WordNet

The Princeton WordNet (PWN) is a free large-scale electronic dictionary developed at Princeton University by George Miller and his col-

leagues (Fellbaum 1998).[10] WordNet 3.1 contains approximately 100,000 synsets which are organised in a hierarchical fashion. A synset is a word sense associated with a group of one or more synonymous words.

PWN organises synsets in terms of the hierarchical is-a relation hyponymy, its inverse hyperonymy and part-whole relations meronymy and holonymy. Sense keys of lexical units have the following representation: *p*ainting%1:06:00:. This representation is composed of: the lexical unit (painting), the word class (1 for indicating a noun), the lexical file number (06), and a sense number (00).

2.3.2 SALDO

The Swedish Associative Thesaurus 2 (SALDO) is an extensive electronic semantic lexicon for the modern Swedish written language that is freely available for Swedish language technology (Borin, Forsberg and Lönngren 2008; Borin, Forsberg and Lönngren 2008).[11] The semantic lexicon has been developed from *The Swedish Associative Thesaurus* (SAL) (Lönngren 1989; Borin 2005) mainly based on corpora, teaching material, and scientific texts. SALDO contains over 100,000 lexical units (word meanings). It is available in the standard model LMF (Francopoulo et al. 2006) and has been recently linked to Core WN (Pedersen et al. 2012). A comparison of the lexicon with other large coverage lexicons is decribed by Borin and Forsberg (2009).

Each lexical unit in SALDO, covering: nouns, verbs, adjectives, adverbs, numerals, prepositions, pronouns, proper nouns, conjunctions, interjections and multiword expressions is associated with at least one other lexical unit. For example, if we look up the Swedish noun *oljemålning* 'oil painting', we find it is associated with the nouns *målning* and *olja*. *Målning* 'painting' is its primary descriptor, *olja* 'oil' is its secondary descriptor. Sense keys of lexical units have the following representation: *målning..1*, which specifies the lexical unit *målning* and its sense indicated by the number *1*.

[10] <http://wordnetweb.princeton.edu/> (Last accessed: 2012-12-12)

[11] <http://spraakbanken.gu.se/saldo/> (Last accessed: 2012-12-12)

2.3.3 MultiWordNet

MultiWordNet (MWN) is a multilingual lexical database which is aligned with Princeton WordNet 1.6 (Pianta, Bentivogli and Girardi 2002).[12] It provides access to lexicons in six languages: Italian, Spanish, Portuguese, Hebrew, Romanian and Latin, that share the same backbone as PWN. MultiWordNet has been developed on the basis of English and Italian synsets, it supports consistency with the PWN model by allowing lexical gaps in cases when lexical concepts are not available in the language. In the current version of MultiWordNet (version 1.4.2) there are around 5,261 synsets, with an average of 1.47 synonyms per synset (Ordan and Wintner 2005; Ordan et al. 2007) available through the multilingual lexical database.

Sense keys of lexical units have the same representation that is available in PWN, however the only information that is available through the MWN is the database location number of the different lexical units, for example 02957109.

2.3.4 FrameNets

A framenet is a resource that provides information about semantic relations which hold between lexical units whose meanings are partially constituted by their relations to other lexical units. It is a predicate argument resource that identifies meaning preserving transformations such as active/passive, verb alternations and nominalizations by providing detail descriptions of verbs, nouns and adjectives.

Fillmore, Johnson and Petruck (2003) were the first to develop an electronic framenet resource for English in the Berkeley FrameNet (BFN) project.[13] A *semantic frame* in the BFN consists of various *slots* called frame elements (FEs), and their *fillers*, called lexical units (LUs). A singel semantic frame carries information about the different syntactic realizations of the frame elements (syntactic valency), and about their semantic characteristics (semantic valency). Lexical units associated with a certain frame are based on shared semantics. Frame elements are the semantic properties (semantic roles) of the lexical units that participate in the frame. Details are given in chapters 12 and 13.

Many NLP researchers have started to acknowledge the value of a semantic-lexical resource such as BFN (Boas 2009) and today more so-

[12] <http://multiwordnet.fbk.eu/english/home.php> (Last accessed: 2012-12-10)
[13] <http://framenet.icsi.berkeley.edu/> (Last accessed: 2012-12-10)

phisticated and computationally oriented framenets are freely available in several languages (Erk et al. 2003; Subirats and Petruck 2003; Ohara et al. 2003; Borin et al. 2010). Recently there has also been some work on integrating FrameNet information in the the Grammatical Framework, GF (Gruzitis, Paikens and Barzdins 2012).

2.4 The Grammatical Framework (GF)

Three of the most important linguistic components of a grammar-based multilingual language generator are syntactic and morphological structures of languages, and multilingual lexicons. A framework that provides access to such valuable resources is the Grammatical Framework, GF (Ranta 2004, 2011).[14] GF is a logical framework that is based on a general treatment of syntax, rules, and proofs by means of a typed λ-calculus with dependent types (Nordström, Petersson and Smith 1990; Ranta 1991, 1994). The formalism is based on Martin-Löf's type theory (Martin-Löf 1984), it is a functional programmering language similar to Haskell. The formalism is originally designed for displaying formal propositions and proofs in natural language but is oriented towards multilingual grammar development and generation (see Ljunglöf, 2004 for a comparison of GF with other grammar formalisms such as combinatory categorial grammar, CCG).

GF has a bidirectional architecture designed explicitly for computational purposes. It is capable of both parsing and generation, which means we can go both from the semantic representations to syntactic structures and vice versa, as depicted in figure 5. As the figure shows, parsing and generation use the same grammar but only one part of the system is employed during generation.

GF treats grammars in a declarative way by separating between abstract and concrete levels, called abstract syntax and concrete syntax. The abstract syntax is guided by the structure of the semantic input while the concrete syntax specifies the linguistic knowledge needed to produce natural language utterances. More than one concrete syntax can be built on top of an abstract syntax to generate words, phrases, sentences, and texts in any natural language.

Abstract and concrete modules are top-level in the sense that they appear in grammars used at runtime for parsing and generation. They can be organized into inheritance hierarchies in the same way as object-oriented programs. GF also comes with a Resource Grammar Library

[14] <http://www.grammaticalframework.org/> (Last accessed: 2012-12-12)

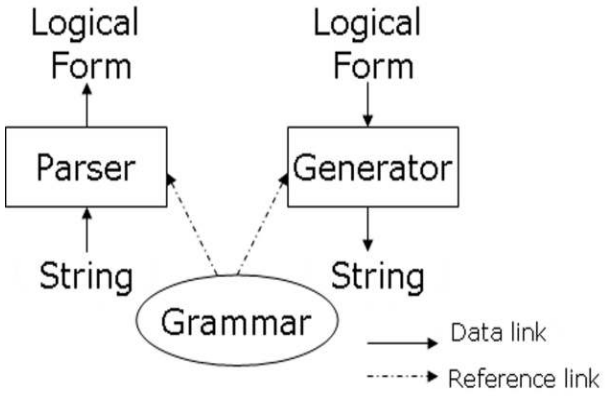


Figure 5: A sketch of bidirectional architecture from Strzalkowski (1994).

(RGL) (Ranta 2009). It is a set of parallel natural language grammars with a common Application Programming Interface (API),[15] covering almost 30 languages.

2.4.1 Multilingual language generation in GF

A computational perspective such as the one taken by GF has been argued to have a potential advantage for natural language generation grammars (Teich 1999). The obvious advantage is the close relation between functional and syntactic perspectives, i.e. the development cycle of the text and linguistic structures are interleaved. This requires some effort of the grammarian when designing a grammar application; the design of the abstract grammar must accommodate linguistic realizations defined in the concrete; writing the abstract while thinking about the concrete syntax of different languages simultaneously is a challenging task for any application developer.

The functional grammar specifications supported by GF permit a language grammar to be specified at a variety of levels of abstraction and preserve word order variations. These features are especially relevant for a multilingual natural language generation system.

GF has been used as a linguistic resource for various language generation tasks. Some examples of the research areas it has been explored in are: generation from formal and informal language specifications (Jo-

[15] <http://www.grammaticalframework.org/lib/doc/synopsis.html> (Last accessed: 2012-11-28)

hannisson 2005; Burden and Heldal 2011), spoken dialogue systems (Ljunglöf and Larsson 2008), controlled natural languages (Ranta and Angelov 2010; Khegai, Nordström and Ranta 2003), ontology translation (Enache 2009), multilingual ontology verbalization (Enache and Angelov 2010; Dannélls et al. 2011).

The two-level grammar specifications have proven advantageous for encoding ontologies. This was demonstrated by Enache (2009) where SUMO, the largest open-source ontology was translated to GF. The system is also capable to perform efficient parsing that is robust enough to be used in real time application (Angelov 2011).

What makes GF attractive for MLG, apart from the advantages mentioned above, is the fact it provides access to the morphology and the syntax of nearly 30 languages. Although the syntactic coverage is limited for some languages, we are not aware of any other grammar formalism with support for this amount of languages.

Moreover, with the GF mechanism it is possible to specify one high-level description of a family of similar languages that can be mapped to several instances of these languages. In GF this is accomplished with the help of a *functor* module. This implies, that if semantic representations defined in the abstract syntax are constructed in a way that allows compositional structure by more than one concrete syntax, it becomes possible to share one resource between several languages and hence, reducing the complexity of linguistic resources.

GF comes with a source authoring environment which deploys similar techniques to those introduced by Power and Scott (1998); Power, Scott and Evans (1998); Dymetman, Lux and Ranta (2000), and van Deemter and Power (2003). There is a grammar developing tool available through Eclipse (Camilleri 2012),[16] and a web-based Integrated Development Environment (IDE).[17]

Other advantages of employing GF in this work are: obtaining access to: (1) the large Swedish computational dictionary, SALDO; (2) a wide coverage grammar for Swedish (Ahlberg 2010); and (3) Hebrew morphosyntactic linguistic specifications (Dannélls and Camilleri 2010).

[16] <http://www.grammaticalframework.org/eclipse/> (Last accessed: 2012-11-28)

[17] <http://www.grammaticalframework.org/~hallgren/Talks/GF/gf-ide.html> (Last accessed: 2012-11-28)

2.4.2 Multilingual grammar example

In this subsection we provide some explanation of the grammatical formalism by showing how a domain grammar is constructed.

Suppose we have the following set of ontology statements:

```
<owl:Thing rdf:about="&painting;DoraMaarauChat">
 <rdf:type rdf:resource="&painting;Painting"/>
 <createdBy rdf:resource="&painting;PabloPicasso"/>
</owl:Thing>
```

From these statements we would like to generate the following sentences: *'Dora Maar au Chat' is a painting. It was painted by Pablo Picasso.* Or express the same information in one sentence: *'Dora Maar au Chat' is a painting by Pablo Picasso.*

As we already mentioned, the abstract syntax gives a structural description of a part of the domain. It has the ability to utilize the same semantic categories differently depending on the complexity of the context and syntax. There are different possibilities in how to encode syntactic variations in the abstract syntax. One of these is illustrated bellow.

Abstract syntax

```
abstract Example = {

 cat Person ;
     PEntity ;
     PType;
     Description ;

 fun describeEntity :
     PEntity -> PType -> Person -> Description ;
     DoraMaarauChat : PEntity ;
     Painting       : PType ;
     PabloPicasso   : Person ;
}
```

The abstract syntax is a context-free grammar without terminals, and where each rule has a unique name. An abstract rule in GF is written as a typed function and is specified with a *fun* declaration . Categories (types) are specified in GF with a *cat* declaration.

The above syntax introduces four categories `Person`, `PEntity`, `PType`, `Description` and four functions for building trees with these

categories. `describeEntity` is a function which takes three arguments `PEntity, PType, Person` and returns a `Description`.

`PabloPicasso` is a constant of type `Person`, `DoraMaarauChat` is a constant of type `PEntity` and `PPainting` is a constant of type `PType`. Each abstract category has a corresponding linearization type in the concrete syntax.

The concrete syntax that is built on top of this abstract syntax, is formulated as a set of linearization rules. These rules are declared differently for each target language. In addition, each concrete syntax also contains grammatical parameters and functions to ensure grammatical correctness for each language and deal with the domain specifications. Two concrete syntaxes for representing the categories and functions defined in the above syntax are specified below.

Concrete syntax for English

```
concrete ExampleEng of Example = {

lincat
  PEntity =  Str ;
  PType =  Str ;
  Person =  Str ;
  Description =  Str ;

lin
 describeEntity pent ptyp pers =
  let str1 : Str =
  ({s = pent.s ++ " is " ++ artIndef ++ ptyp.s}).s;
  str2 : Str =
  ({s = "It is " ++ " painted by " ++ pers.s}).s ;
    in (str1 ++ "." ++  str2 ++ ".")    |
  let str12 : Str =
    ({s = pent.s ++ " is " ++  artIndef ++
     ptyp.s ++ " by " ++ pers.s}).s  in str12 ;

  PabloPicasso = {s ="Pablo Picasso"} ;
  DoraMaarauChat = {s = "Dora Maar au Chat"} ;
  PPainting = {s = "painting"} ;

oper
  artIndef : Str  =
  pre {"a" ; "an" / strs {"a";"e";"i";"o";"u"}} ;
}
```

Each category defined in the abstract is given a linearization type declared with *lincat*. In the example above four categories are linearized as strings: `PEntity`,`PType`,`Person` and `Description`. Each function is linearized in the concrete with *lin*. `describeEntity` contains two strings `str1` and `str2` that are composed with the operator `++`. Each string contains a set of strings and records concatenated with the same operator. The operator `|` denotes variants. With this operator we avoid defining a new function for expressing the same semantic knowledge with two different syntactic structures. The `oper` function `artIndef` is using the built-in haskel macro `pre` to generate the right English article. If the first token of a string is a vowel it will generate the article *an*, otherwise *a*. `PEntity`,`PType`,`Person` are defined as records containing just one field, `s`. We define an operation `oper` to generate a correct article depending on whether the following word starts with a vowel or a consonant.

The Swedish concrete syntax `ExampleSwe` is also built on top of the abstract syntax `Example`.

Concrete syntax for Swedish

```
concrete ExampleSwe of Example = {

lincat
  PEntity = Str ;
  PType = {s : Str  ; g : Gender} ;
  Person = Str ;
  Description =  Str ;

 lin
  describeEntity ptyp pent pers =
   let str1 : Str =
     ({s = pent.s ++ " är " ++ artIndef ! ptyp.g
                                 ++ pent.s}).s ;
     str2 : Str =
     ({s =  case ptyp.g of
           {Neu => "Det" ; Utr => "Den"} ++
                " målades av " ++ per.s}).s  ;
       in (str1 ++ "." ++  str2)   |
   let str12 :  Str =
     ({s = pent.s ++ " är " ++ artIndef ! ptyp.g
      ++ ptyp.s ++ " av " ++ pers.s}).s in str12 ;
```

```
PabloPicasso = {s ="Pablo Picasso"} ;
DoraMaarauChat = {s = "Dora Maar au Chat"} ;
PPainting = {s = "Painting"  ; g =  Utr} ;

param
      Gender = Neu | Utr ;
oper
 artIndef : Gender  ->  Str =
       table { Neu => "ett" ; Utr => "en" } ;
}
```

Since Swedish nouns have a gender associated with them, we introduce the parameter type `param Gender` for noun gender, with two records `Neu` and `Utr` and add a gender parameter `g : Gender` to `PPainting`. The inflection rule `artIndef` is defined as an operator to deal with gender inherence, it contains an inflection table that takes a gender argument. Instead of inflection rules it is also possible to encode inflection tables directly in the grammar (as seen in `str2`). `str1` and `str2` yield *Dora Maarau Chat är en målning . Den målades av Pablo Picasso*. `str12` yields *Dora Maarau Chat är en målning av Pablo Picasso*.

As seen in the concrete syntax for Swedish, the concrete grammar contains explicit linguistic knowledge about the language. Such linguistic knowledge is encoded in the modules by using features including parameters, tables and records represented as types and values represented as strings.

The grammar developer does not have to define all these features if he/she chooses to utilize the resource grammar. By utilizing the GF constructors defined in the RGL, the developer does not have to care about gender agreement, words inflections, etc. All these relevant linguistic features are already defined in the RGL. The grammar developer can exploit a language resource grammar by importing the resource grammar module in the concrete grammar and use the language API to construct the grammar, as illustrated below.

Concrete syntax for Swedish using RGL

```
concrete AbsSweRGL of Abs =
     open SyntaxSwe, ParadigmsSwe  in {

 lincat
  PEntity = NP ;
  PType = CN ;
  Person = PN ;
  Utterance =  Text ;

 lin
 describeEntity pent ptyp pers  =
  let str1 : Phr = mkPhr (mkS presentTense
   (mkCl  pent  (mkVP (mkNP a_Art  ptyp))));

       str2 : Phr = mkPhr (mkS pastTense
   (mkCl  it8utr_Pron (mkVP
   (mkVP (mkVP (PassV2 paint_V2)
   (SyntaxSwe.mkAdv by8agent_Prep (mkNP pers))))));

   in
       mkText str1  (mkText str2);
 |
   let str12 : Phr = mkPhr (mkS presentTense
   (mkCl  pent  (mkVP (mkVP (mkNP a_Art ptyp))
   (mkAdv by8agent_Prep (mkNP  pers)))));
   in
       mkText str12 ;

 lin
   PabloPicasso =  mkPN "Pablo Picasso" ;
   DoraMaarauChat = mkNP (mkPN "Dora Maarau Chat") ;
   PPainting = mkCN (mkN "målning") ;
}
```

The Swedish resource library is imported with the `open` operator. Two modules are imported in `AbsSweRGL`: `SyntaxSwe` and `ParadigmsSwe`. In the above implementation we find the following categories and functions imported from the RGL for Swedish. Four category types: `NP` (Noun Phrase), `CN` (Common Noun), `PN` (Person Name) and `Text`. Seven functions: `mkPhr` (make Phrase), `mkS` (make Sentence), `mkCl` (make

Clause), `mkNP` (make Noun Phrase), `mkVP` (make Verb Phrase), `mkAdv` (make Adverbial), and `mkText` (make Text). An explanation of the resource categories and function are given in appendix E.

3 Data collection and analysis

This chapter describes the data collection and the data analysis. Three distinct analyses are presented: syntactic analysis using state-of-the-art NLP tools, semantic analysis using domain ontology concepts, and referential expression analysis using lexical-semantic relations.

3.1 The corpus data

One way of identifying the linguistic features that characterize a domain in different languages is through corpus analysis. Since there is a shortage of corpora of artwork descriptions in general and paintings in particular in any of the languages we are interested in, we gathered data from the cultural heritage domain to study the domain-specific conventions and the ways of signalling linguistic content in English, Swedish and Hebrew.

The nature of similarity of the collected texts is their content. All texts consist of comprehensible, well-formed work-of-art descriptions. The texts are gathered from museum websites, and are all written by native speakers of the language. This type of text collection is often referred to as multilingual *Comparable Corpora* in Applied Linguistics (Hunston 2006). Comparable corpora contain texts on the same topic collected from different sources. The purpose of gathering such corpora is to study linguistic innovators, differences in syntax and terminology acquisition.

The texts were extracted automatically from digital libraries that are available through online museum databases. The majority of the Swedish descriptions were extracted from the World Culture Museum,[18] the majority of the English descriptions were collected from the Met

[18] <http://collections.smvk.se/pls/vkm/rigby.welcome> (Last accessed: 2012-08-19)

Table 3.1: Statistics of the text collections.

	English	Swedish	Hebrew
Number of descriptions	394	386	110
Number of tokens	42792	27142	5690
Number of sentences	1877	2214	445
Avg. sentence length	24 (tokens)	13 (tokens)	13 (tokens)
Avg. description length	5 (sentences)	6 (sentences)	4 (sentences)

Museum.[19] The majority of the Hebrew descriptions were extracted from Artchive.[20]

A Web crawler was used to search and retrieve all html pages containing works-of-art descriptions. From the retrieved web pages, html tags and other metadata were removed. Some unstructured sentences that may have appeared in the beginning or at the end of a coherent description were also removed. Table 3.1 gives an overview of the three text collections.

As shown in table 3.1, the average length of an object description is very similar for English, Swedish and Hebrew. From the table we learn the Hebrew dataset is much smaller compared to the English and Swedish datasets. Note the big difference in the average amount of tokens in English sentences compares to Swedish and Hebrew which might depend on the commonly used writing style from the websites the texts were extracted from.[21] The rich production of compounds characterizing Swedish, among other languages (Rosell 2009), may explain the low amount of tokens found for Swedish as compared to English. When we analyze the data (section 3.2.4) we examine whether differences in sentence length (table 3.1, fourth row) affect the choice of the referential expression.

3.2 Data annotation and analysis

When generating text from formal semantic representations, it is important to understand how occurrences of instances of semantic concepts are realized and combined syntactically, together with the different types of grammatical functions they may fulfill when occurring with other instances of concepts.

[19] <http://www.metmuseum.org> (Last accessed: 2012-08-19)
[20] <http://www.artchive.com/> (Last accessed: 2012-08-19)
[21] The terms lexical unit and token are used interchangeably.

One way to discover lexical and grammatical functions for structuring a discourse is through a traditional analysis of text examples. The analysis carried out in this study aims to bring understanding about how artwork descriptions are realized linguistically by human writers. The approach combines quantitative and qualitative discourse analysis of the corpus data using the methodology of corpus linguistics to explore syntactic and lexical-semantic relations in special type of texts.

3.2.1 Syntactic processing

One part of content analysis is to identify the frequent grammatical categories co-occurring with particular verbs and to examine which lexical units are tagged with these categories in three languages. For the purpose of acquiring this information automatically, we chose to annotate the texts with a dependency parser (Nivre 2005).

A representation based on dependency structure provides two valuable pieces of information: (1) a binary relation between two words, i.e. the head and the dependent. (2) grammatical relations among two or more words, e.g. subject, direct/indirect object, modifier, etc. While the former is more purely syntactic, for example: in *to Philadelphia* the preposition *to* is the head of the proper noun *Philadelphia*, the latter have a complex mapping to semantic argument structure, for example in *Robertson painted this miniature*, *Robertson* is the (nominal) grammatical subject (nsubj) of *painted* and *miniature* is the grammatical direct object (dobj). To precisely interpret the argument structure of the verb *paint*, we would need world knowledge to understand that *miniature* is an artwork.

The syntactic analysis was carried out automatically. All sentences in each corpus were tokenised, part-of-speech (PoS) tagged, lemmatized and parsed using open source software. HunPoS (Halácsy, Kornai and Oravecz 2007; Megyesi 2009), an open source Hidden Markov Model (HMM) tagger based on the Trigrams'n'Tags (TnT) tagger (Brants 2000) was used for tagging the English and the Swedish data. Meni and Elhadad's (2006) PoS tagger was utilized for tagging the Hebrew texts. The English, Swedish and Hebrew tagging models were downloaded from their respectively web-pages.[22] [23] [24] Maltparser, version 1.4 (Nivre et al. 2007) was used for parsing the English, Swedish and Hebrew data.

[22] <http://code.google.com/p/hunpos/downloads/list> (Last accessed: 2012-08-17)
[23] <http://stp.lingfil.uu.se/~bea/resources/hunpos/> (Last accessed: 2012-08-17)
[24] <http://mila.cs.technion.ac.il/mila/eng/tools.html> (Last accessed: 2012-06-17)

Table 3.2: Tagger and parser models accuracy.

Language	Tagger	Malt-Parser
Eng	96.7 %	88.1 %
Swe	95.9 %	86.3 %
Heb	92.4 %	83.5 %

The models that have been used for parsing the English and Swedish data are available from the Maltparser web-page.[25] The model for parsing the Hebrew data was trained with MaltParser on the Hebrew Treebank and is described by Tsarfaty, Nivre and Andersson (2012).

Table 3.2 shows the accuracy of each of the tagging and parsing models given in percentages of the success of predicting the correct PoS tag and head-word. These results were reported by Goldberg, Adler and Elhadad (2008), Nivre et al. (2007) and Tsarfaty, Nivre and Andersson (2012). Each of the listed taggers (table 3.2) was trained on a dataset consisting of around one million tokens, each parser was trained on unlabelled data.

Table 3.3 shows three extracts from each of the three datasets. Each sentence is annotated and parsed with a dependency parser. Note different taggers and parsers use different PoS tags and dependency labels. The tag sets for each language are specified in appendix A and B. The transliteration table for Hebrew is specified in appendix D.

Given a set of text examples annotated with part-of-speech tags and with syntactic representations consisting of grammatical relations between lexical units, it is possible to acquire statistical information about how lexical units are realized syntactically, for instance how many nouns appear in subject positions, what are the lexical units they are realized with, how many prepositions appear in direct object positions, etc. Table 3.4 gives frequency counts of binary syntactic dependencies for: Proper nouns (PN), Numerical Expressions (NE), Nouns (NN), Verbs (VB), Prepositions (PP) and Pronouns (Pro). As this table illustrates, the most frequent syntactic dependencies are nouns dependent on verbs (first and fourth rows) and nouns dependent on prepositions (sixth row). Occurrences of proper nouns and pronouns in subject position are very low in the Hebrew dataset.

[25] <http://maltparser.org/mco/mco.html> (Last accessed: 2012-08-17)

Table 3.3: Data annotated with dependency parsers.

Eng	Swe	Heb
1 This DT 2 det	Målningen _ NN UTR | SIN | DEF | NOM 4 DT	AIWR NN 0 hd
2 work NN 3 nsubj	av av PP _ 1 ET	FL POS 1 posspmod
3 is VBZ 0 null	Erik Erik PM NOM 2 PA	EQDT NNT 2 gobj
4 among IN 3 prep	XIV 14:e RO NOM 5 SS	ICXQ NNP 3 gobj
5 the DT 7 dep	är vara VB PRS | AKT 0 ROOT	EL IN 1 mod
6 most RBS 7 advmod	utförd utförd PC PRF | UTR | SIN | IND | NOM 5 SP	H DEF 7 def
7 celebrated JJ 4 pobj	av av PP _ 5 OA	QJQWMBH BN 5 pobj
8 of IN 7 prep	den den DT UTR | SIN | DEF 10 DT	yyLRB PUNC 7 appos
9 those DT 8 pobj	holländske _ JJ POS | MAS | SIN | DEF | NOM 10 AT	AWMNWT NNT 8 hd
10 painted VBN 9 partmod	konstnären konstnär NN UTR | SIN | DEF | NOM 11 DT	QBWRH NN 9 gobj
11 between IN 10 prep	Steven Steven PM NOM 7 PA	NWCRIT JJ 9 mod
12 1870 CD 11 pobj	van van PM NOM 11 HD	yyRRB PUNC 8 punct
13 and CC 12 cc	der der PM NOM 11 HD	FL POS 8 posspmod
14 1874 CD 12 conj	Meulen _ PM NOM 11 HD	PRISILH NNP 13 gobj
15 . . . 3 punct	. _ RG NOM 11 ET	, PUNC 8 punct
16		MAH CD 1 hd
17		4 CD 16 hd
18		L PREPOSITION 16 mod
19		SPIRH NN 18 pobj
20		. PUNC 1 punct

Table 3.4: Statistics of binary syntactic dependencies.

(relation; head,dependent)	Eng	n/1000	Swe	n/1000	Heb	n/1000
(subj; VB, NN)	1235	29.4	725	26.8	250	44.6
(subj; VB, PN)	757	18	135	5	48	8.6
(subj; VB, PRO)	604	14.3	391	14.4	43	7.6
(dobj; VB, NN)	1793	42.6	1050	38.8	149	26.6
(dobj; VB, PP)	1932	46	709	26.2	78	13.9
(dobj; PP, NN)	5356	127	1363	50.4	175	31.2
(dobj; PP, NE)	400	9.5	208	7.7	25	4.4

3.2.2 Semantic processing

The strategy to perform the semantic content analysis described in this section follows Kilgarriff's approach (Kilgarriff 2001; 2010):

- identify the most **frequent semantic concepts** used throughout the documents;
- **classify words** according to these concepts, to give a content analysis dictionary;
- count the number of **occurrences of each semantic concept**.

3.2.2.1 Most frequent semantic concepts

The schema of concepts and relations that is used in this study is the one provided by the CIDOC-CRM, version 5.0.1 (see section 4.2).[26]

By studying the source texts in the three languages, we identified the most frequently occurring semantic concepts. These concepts, manually selected, are listed in alphabetical order in table 3.5.

3.2.2.2 Classification of words into semantic concepts

For each language, two annotators were given: (1) the original text collection annotated with only id numbers; and (2) the set of semantic concepts (table 3.5) with which they should manually annotate the texts. They were also given the CIDOC-CRM manual (Crofts et al. 2009) in order to explore further each of the semantic concepts using the ontol-

[26] <http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.1_Nov09.pdf> (Last accessed 2012-12-12)

Table 3.5: Semantic concepts most prominent in the source texts.

Actor (ACT)	Legal Body (LEB)
Actor Appellation (AAP)	Man-Made_Object (MMO)
Collection (COL)	Material (MAT)
Conceptual Object (CON)	Place (PLC)
Dimension (DIM)	Time-span (TMP)
Entity (ENT)	Title (TIT)
Event (EVT)	

ogy schema and the concept definitions. An extract from a work-of-art description from the English text collection is given below.

> <dd id=96> This is a sketch for a large canvas by Rubens with workshop assistance, which was formerly in the Kaiser-Friedrich-Museum, Berlin, and destroyed in 1945. The sketch and the Berlin picture are generally dated to about 1630, or to the early 1630s. </dd>

The data annotation process was exploratory in the sense that annotators were discussing their way through the annotation. For instance, we found during the annotation process that there are three elementary semantic concepts, viz *Conceptual Object* (CON), *Event* (EVT) and *Represented* (ENT), which frequently appear in the datasets. A token or a phrase annotated with one of these concepts describes what is depicted in the artwork, or provides historical information about it. These three concepts, in particular *Represented*, which corresponds to the top-level concept *CRM_Entity*, comprise all possible things in the universe. In most cases these concepts span longer text fragments, for example: *the English tradition of topographical painting* (*Conceptual Object*), *Rummet är drottningens matsal på Stockholms slott* (*Represented*), 'The room is the queens's dinning room in the Stockholm's castle'. Of this reason they are not considered by the language generator (see chapter 4). Moreover, because the concept *Represented* is only annotated in the Swedish data and partly in the English data, there is no statistical information about it in table 3.8.

The semantic annotation was carried out independently of the syntactic annotation. That is to say, the syntactic annotation played no role in the semantic annotation. This decision influenced the statistical measurements of inter-annotator agreement, e.g. we were only able to com-

pare the lexical units that were annotated by both annotators in each language (see next subsection). However, relying on the chunks provided by the parsers or other state-of-the-art chunkers could have been more problematic mainly because there is a lack of consensus regarding the assumptions made by different chunkers in different languages (Abney 1991; Goldberg, Adler and Elhadad 2006). We felt that it would have been detrimental to employ an automatic chunker for each language at the point of annotation.

The annotators were encouraged to carry out the semantic annotation at word level and to avoid recursion. In addition to the semantic annotations, the annotators were asked to annotate three semantic concepts: *Man-Made_Object*, *Actor Appellation* and *Actor* with coreference links indicated by free variables, i.e. *i, j, m*, etc. Examples of the semantically annotated sentences are given below.

```
This [[work]MMO]j is among the most celebrated of
those painted between [1870 and 1874]TMP.
[[It]MMO]j  depicts [Eakins']ACT]k boyhood friend
[Max Schmitt]ACT]q.

[[Målningen]MMO]i av [[Erik XIV]ENT]i är utförd av
[den holländske konstnären]AA
[[Steven van der Meulen]ACT]j.

[[AIWR]MMO]i FL [EQDT ICXQ]ENT EL [HQJQWMBH]PLC
yyLRB [[AWMNWT QBWRH NWCRIT]MMO]i yyRRB FL
[[PRISILH]ACT]j, [MAH 4 LSPIRH]TMP.
```

Table 3.6 shows some of the most frequent lexical units we extracted from the datastes using these annotations. From these occurrences we learn that multiword units are common for expressing the concepts *Actor appellation, Legal Body and Time* in English. In the Swedish dataset we find many compounds associated with the concepts *Material, Actor appellation, Legal Body* and *Collection*. In Hebrew, multiword units are commonly used for expressing *Material* .

Inter-annotator agreement

A Java program was written to generate two lists of lexical units annotated by each pair of annotators. The method for gathering each list

Table 3.6: Word lists based on semantic annotations.

Concept	Eng	Swe	Heb
ACT	*Robertson, Stuart, They, Stuart's, Willing, Degas, his, Sharples, he, Smith Stephen's, Washington's, Powers'*	*Rembrandt,Lucas,Jan,Steven, Pietro,Auguste,Anders, Eva,kejsar,Aleksis,Carl, Lauri,Alpo,Gunnar,Johan*	**אבנר, גרשוני, שירן, צבי**, **שרגא, צבי, לחמן**
AAP	*miniaturist, politicians, diplomat, student, member, portraitist, artist, painter, shipbuilder, maker, president, sculptor, Impressionists, New York collector, colorist*	*akvarellmålare,porträttörer, Realisterna,betraktaren,skulptör, konstnärsparet,konstnären,friluftsmålare*	**אמנים, צייר, אמן**
LEB	*Museum of Western Art committee of the American Art-Union Museum, Memorial, Arts, Library*	*museet,Mariner,kulturskola, Konstmuseets,Slott,stadsmuseum, Konstakademien,Förlag,kulturskola*	**במוזיאון, מוזיאון**, **מוזיאון מטרופולין**, **מוזיאון אארמיטאז**
MMO	*copies, portrait, examples, novel, miniature, picture, scenes, painting, photograph, drawings*	*Tavlan,Pianopall,fotografiet, Sebastian,bild,Dashavatara, keramikfragment,originalkortet*	**דיוקן, יצירה**, **פסל, שקף, תמונה**, **רישום, קומפוזיציה, איור**
MAT	*marble, ivory, gold, bronze, copper, zinc, paint, oils*	*olja,fibermaterial,lera,textil,koppar, ull,snäckskal,chi-väv,siden,skinn,bambu, papper,ullgarner,lack,keramik,sten,silkespapper*	**נייר צילום, מים, זכוכית**, **צבע תעשייתי, אקריליק, ברזל**, **עפרונות צבעוניים, נייר, שיש**
PLC	*Dublin, New York, America, river, Philadelphia, studio, England, Island, Courtyard, resort, D.C., room, House,*	*Stockholm,Salzburg,Haag, Nationalmuseum,Sverige,Gripsholm, Vishvanathatemplet,Varanasi,Mexico*	**אוסטריה,ווינה**, **פריז, תל אביב**, **בבית, צרפת**
TMP	*a year and a half, fall of 1834 the first half of the nineteenth century 1793, years, century, (1808–1861), century, period*	*AD, tiden,högtider,1920,1931, idag,nu,e.kr,2004,1800-talets, 1866,1865,tid,idag,1739 dåtidens,samtidigt,tidevarv*	*1990-1998* **שנות, משנת**, *1638* **לערך, מהמאה, בשנת**
COL	*private collection, personal collection*	*Samlingen,Adney,collection, porträttsamling,galleri,serie, konstsamling,Dockskåpsinvent*	**תערוכה**, **סדרה של** *7* **הדפסים**, **סדרת צילומים, סדרה, מיצג**

was as follows: for each lexical unit in a sentence, if the lexical unit starts and ends with a square bracket, add it to the list, else if the lexical unit starts with a square bracket add it to the list until a lexical unit ending with a square bracket is found. With this method it is possible to compare single lexical units with phrases, for example: both the lexical unit *painting* and the phrase *the famous painting* will result in a match for which a vector will be computed if both are semantically annotated.

A metric was computed by calculating the index of similarity between each item appearing in the gathered lists of lexical units. If a lexical unit appears in both lists and is tagged with the same tag, assign value 1, else if the tag differs, assign value 0. The computed metric based on this pairwise-agreement between two coders was used to calculate the inter-annotator agreement with the Fleiss' kappa statistical measure for assessing reliability (Fleiss 1971). It is an extension of the kappa statistic first proposed by Cohen (1960) given in equation (1):

$$\mathcal{K} = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

P(A) is the proportion of times that the annotators agree. P(E) is the proportion of times that the annotators would be expected to agree by chance. The results of the inter-annotator agreement calculations are reported in table 3.7.

Table 3.7: Results of the inter annotators agreement calculation.

	$\mathcal{K}$
English	0.93
Swedish	0.87
Hebrew	0.92

According to the interpretation of the kappa calculation values, the agreement is almost perfect for the three languages since $\mathcal{K} > 0.8$ (Cohen 1960). A closer look at the concepts where the annotators disagreed on shows that the concepts giving rise to the disagreement depended on the language. In the English data there were several disagreements between *Title* and *Man-made object*, in the Swedish data between *Represented* and *Title*, in Hebrew data between *Conceptual object* and *Event*.

Table 3.8: Frequencies of the concepts appearing in table 3.5. Absolute frequency (abs), relative frequency per 1000 tokens (n/1000).

Concept	English		Swedish		Hebrew	
	abs	n/1000	abs	n/1000	abs	n/1000
ACT	2833	66	984	**36**	429	**75.2**
AAP	750	17.5	193	7	193	33.8
COL	128	2.9	39	1.4	25	4.3
CON	1267	29.6	451	16.7	37	6.5
DIM	36	0.8	73	2.7	67	11.7
EVT	1380	32	620	22.9	99	17.3
LEB	241	5.6	47	1.7	13	2.2
MMO	2686	**62.7**	1225	45	303	53.1
MAT	40	0.9	439	16.2	128	22.4
PLC	723	16.8	350	12.9	89	15.6
TMP	711	16.6	286	10.5	107	18.7
TIT	88	2	541	20	53	9.2

3.2.2.3 *Occurrences of semantic concepts*

The number of occurrences of the semantic concepts in each dataset are displayed in table 3.8. According to this table, the most frequent concept for English is *Man-made object* while for Swedish and Hebrew it is *Actor*. *Actor* is the second most frequent concept for English and *Man-made object* is the second most frequent concept for Swedish and Hebrew. The concepts *Dimension* and *Material* appear more frequently in the Hebrew dataset than in English. *Actor Appellation* is much more frequent in the English and the Hebrew datasets than in Swedish. *Title* appears frequently in the Swedish dataset, but as noted in section 3.2.2.2, this is one of the concepts where the annotators disagreed. All other concepts appear to be quite marginal.

3.2.3 Referential Expressions (RE)

The set of semantic concepts provided by the domain ontology CIDOC-CRM are general and require a meaningful, more specific, semantic definition. This has been observed in the past (Binding 2010) and can also be seen in table 3.6, which clearly shows that many of the CIDOC-CRM concepts can be fine-grained to fully describe other entities such as countries, cities, streets, buildings, surnames, etc.

Table 3.9: Occurrences of lexical units classified into fine-grained semantic concept hierarchy of the CIDOC-CRM concept Man-Made Object.

Concept	Lexical units		
	English	Swedish	Hebrew
Man-Made Object	object	föremål	obyeqt **אוביקט**
Artwork	artwork, work,	verk, konstverk	ycyrat aomanot **יצירת אומנות**,
	composition		ycyra **יצירה**,
Painting	picture, painting	bild,	cyor **ציור**,
		målning, tavla	iyor **איור**
			tmona **תמונה**
Portrait Painting	portrait	porträttmålning	potret **פוטרט**,
			dyoqan **דיוקן**

Since the focus of this work is the semantic concept *Man-Made Object*, the concept was fine-grained to cover specific types of paintings. A detailed description of this extension is provided in section 4.2. In table 3.9, an example of the concept's hierarchy and the lexical units associated with each concept in each language are presented.

Each of the listed lexical units in table 3.9 can appear in the role of a valid unit to refer to a work of art. In this study, we look closer at two linguistic forms of referring expressions: definite nouns and pronouns, focusing on three lexical-semantic relations that hold between a painting object and a definite noun, namely: *direct-hyperonym (DH)* is, in a concept-hierarchy, the first superordinate concept of the lexical unit in question. For example, Painting is direct-hyperonym of Portrait Painting, Man-Made Object is a direct-hyperonym of Artwork; *higher-hyperonym (HH)* is the second, third or fourth superordinate concepts of the lexical unit in question. For example, both Artwork and Man-Made Object are higher-hyperonyms of Portrait Painting; *synonym (S)*, the relation between two lexical units which appear within the same semantic concept, for example both artwork and work belongs to the concept Artwork. Frequency data for these linguistic forms and lexical-semantic relations are specified below.

Definite nouns We counted the frequencies of nouns that are listed in table 3.9 and that proceed definite and demonstrative articles. The results are reported in table 3.10.

From table 3.10 we learn that definite noun references are equally frequent in English and Hebrew and somewhat less frequent in Swedish.

Table 3.10: Occurrences of a selected set of nouns in definite form.

English			Swedish			Hebrew		
Def. article	abs	n/ 1000	Def article	abs	n/ 1000	Def. article	abs	n/ 1000
the/this	856	20	den/det -et/-en	408	15	ha ה	116	20.7

Pronouns The frequency of personal pronouns is reported in table 3.11. Pronouns seem to be more frequent in the Swedish data. Possessive pronouns are very common in English compared to Swedish. The frequency of pronouns in the Hebrew data is very low.

Table 3.11: Frequencies of personal pronouns in the source texts.

English			Swedish			Hebrew		
Pron	abs	n/ 1000	Pron	abs	n/ 1000	Pron	abs	n/ 1000
he/she	285	6	han/hon	299	11	hw הוא/ hy היא	5	1
it	497	11.6	den/det	788	29	ze זה/zo זו	13	2.1
they	44	1	de	36	1.3	hem הם/hen הן	3	0.5
his/ hers	438	10.2	hans/ hennes	107	4	šelo שלו/ šela שלה	2	0.3

Direct-hyperonym The frequency of the lexical units of the concept *Painting* is reported in table 3.12. From this table we might conclude that both English and Swedish tend to refer to a painting entity with the lexical-semantic relation direct-hyperonym.

Table 3.12: Frequencies of lexical units of direct-hyperonym in the source texts.

English			Swedish			Hebrew		
Painting	abs	n/ 1000	Painting	abs	n/ 1000	Painting	abs	n/ 1000
painting	230	5.8	målning	86	3.1	tmona	5	1
picture	185	4.3	tavla/bild	141	5.2	cyor/ayor	34	5.9
total	315	10.1	total	227	8.3	total	39	6.9

Higher-hyperonym Table 3.13 displays the frequencies of lexical units classified according to the concepts *Artwork* and *Man-made Object*. The frequency of lexical units annotated with these concepts is very high in the Hebrew dataset.

Table 3.13: Frequencies of lexical units of higher-hyperonym in the source texts.

English			Swedish			Hebrew		
Artwork	abs	n/ 1000	Artwork	abs	n/ 1000	Artwork	abs	n/ 1000
artwork/ work	234	5.4	konstverk/ verk	123	4.5	ycyrat omanot/ ycyra	87	15.2

3.2.4 Combining semantic, syntactic and RE

We analyzed the texts for patterns of coreference in the discourse by studying the semantic, syntactic and referring expression annotations simultaneously. An example of how the annotations were combined is given in figure 6. Two linked types of coreference have been explored: (1) reference, i.e. definite noun and pronoun; and (2) repetition, focusing on three lexical-semantic relations direct-hyperonym, higher-hyperonym and synonym. The analysis consisted of two phases: (1) analyze the texts for discourse patterns (DP); and (2) analyze the texts for patterns of coreference (CP) of the ontology concept *Man-made Object*.

In the remainder of this section we present some examples from the analysis of anaphoric expressions in the corpus of multilingual written object descriptions. In the beginning of each example we specify DP sequences by listing the chains of the semantic concepts participating in the discourse, and CP sequences by listing the chains of coreference relation types participating in each sentence. Note, each semantic concept is said to correspond to an ontology statement, a triple, according to the definition in section 2.2.1.

3.2.4.1 Direct-hyperonym

Direct-hyperonym is the most common type of reference that exists in English and Swedish. We found many examples of the relation direct-hyperonym in both languages. In the Hebrew data only a few occurrences of this relation were found. Below follow some examples:

(2)
- **DP:** [MMO TMP EVT ACT MMO REP]
- **CP:** [DH PRO]

(a) **This canvas**, *first exhibited in 1799, was sold by the artist in 1808 to his biographer, John Knowles.*

(b) **It** *illustrates a passage from "Paradise Lost".*

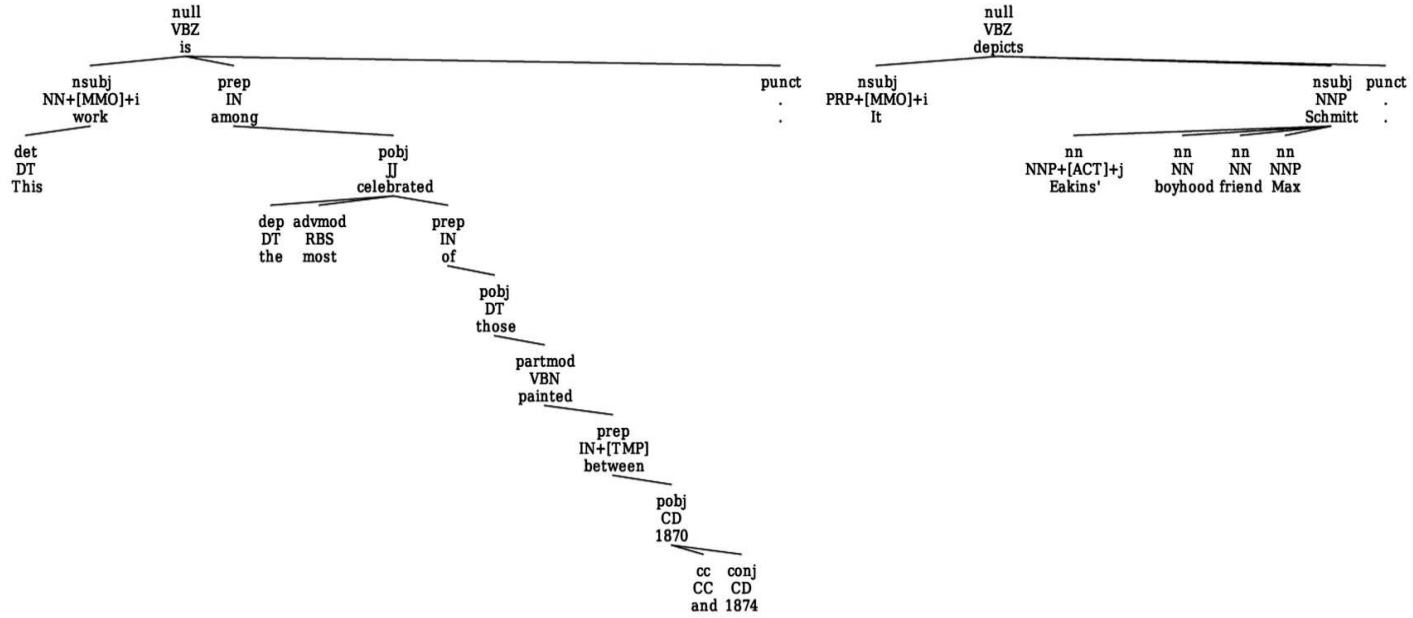


Figure 6: Dependency tree with semantic and coreference annotations for the sentence: *This work is among the most celebrated of those painted between 1870 and 1874. It depicts Eakins' boyhood friend Max Schmitt.*

(3)
- **DP:** [TIT MMO MAT REP MMO LEB TMP]
- **CP:** [DH DH]

(a) *Dockskåpsinventarie.* **Rund tavla** *av papper med*
Dollhouse-furniture. Round painting of paper with
målat motiv föreställandes en ängel med vingar
painted motive depict-PASS an angel with wings
och korslagda armar.
and crossed arms.
'Dollhouse-furniture. Round paper painting painted with a motive depicting an angel with wings and crossed arms.'

(b) **Tavlan** *hör till ett dockskåp som*
painting-DEF belongs to a dollhouse which
skänktes till Stockholms stadsmuseum 1950
donated-PASS to Stockholm-GEN city-museum 1950
av en privatperson i Stockholm.
of a private-individual in Stockholm.
'The painting belongs to a dollhouse that was donated to the Stockholm's city museum in 1950 by a private individual in Stockholm.'

The Hebrew data also contains definite nouns determined by the direct-hyperonym relation, as the noun *hacyor* in example (4b).

(4)
- **DP:** [MMO ACT TMP MMO EVT]
- **CP:** [DH DH]

(a) **CIWR QIR** *FL HAMN ABRHM AWPQ,*
painting wall of artist-DEF Abraham Aufek
1985-1988.
1985-1988.
' . 1985-1988 **ציור קיר של האמן אברהם אופק** '
'Wall painting by the artist Abraham Aufek 1985-1988. '

(b) **HCIWR** *QWFR BIN SPINT*
painting-DEF bind-MASC-SG between ship-GEN
HMEPILIM FJBEH BET NSIWN HELIH
immigrant sink-FEM in-at attempt immigration-DEF
LARC IFRAL WBIN EQDT ICXQ.
to-country israel and-between binding-GEN Itzhak,
הציור קושר בין ספינת המעפילים שטבעה בעת נסיון העליה '
לארץ ישראל ובין עקדת יצחק .'
'The painting connects the immigrant ship that sank while approaching Israel with the event of Itzhak's binding. '

3.2.4.2 Higher-hyperonym

Higher-hyperonym is a lexical-semantic relation of reference expressions that is very common in both English and Hebrew.

(5)
- **DP:** [MMO TIT MMO MMO ACT TMP]
- **CP:** [PRO PRO HH]

(a) **This** *is the only surviving fragment from a painting of the "Madonna and Child with Saint John the Baptist".*

(b) **It** *is* **an early work** *by Sarto, painted around 1506.*

(6)
- **DP:** [MMO ACT MMO MAT ACT PLC TMP]
- **CP:** [DH HH]

(a) *Along with* **other portrait busts** *of statesmen,*

(b) **this work** *was translated into marble after Powers settled in Florence permanently in 1837.*

Mixture of genders in a discourse is overwhelmingly common in Swedish, as illustrated in (7b) and (7c); although for some readers this might be considered incoherent. In these examples we also find a mixture of the lexical-semantic relations higher-hyperonym and direct-hyperonym.

(7)
- **DP:** [TIT DIM MMO MMO DIM MMO LEB COL TMP]
- **CP:** [HH PRO DH]

(a) *'Rembrandts* *självporträtt'* *är* *en* *till* *formatet* *liten*
Rembrandt-GEN self-portrait is a to format small
målning,
painting,
' Rembrandt's self-portrait is by size a small painting, '

(b) **den** *mäter* *endast* *12,2x15,5* *cm.*
it measure only 12,2x15,5 cm.
'it measures only 12,2 by 15,5 cm.'

(c) **Porträttet** *inköptes* *till* *Nationalmuseum* *i*
portrait-DEF purchase-PASS to national-museum in
samband *med* *Rembrandtutställningen* *år* *1956.*
connection with Rembrandt-exhibition-DEF year 1956.
'The portrait was purchased by the national museum in connection with Rembrandt's exhibition in the year 1956.'

Possessive constructions known as *free genitives*, where the head noun appears in the absolute state (Wintner 2000) are very common in the Hebrew dataset, as exemplified in (8a), i.e. **יצירה של האמן** 'artwork by the artist'. When this construction appears, we often find higher-hyperonymy.

(8)
- **DP:** [TIT MMO ACT MMO MAT ACT PLC TMP]
- **CP:** [HH DH]

(a) *"LICXQLH".* **ICIRH** *FL HAMN MFH GRFWNI*
Lisaacla. artwork of artist-DEF mose garshoni
HEWSQT BEIQDT ICXQ,
deal-DEF-FEM-SG in-binding-GEN Itzhak ,
' , **יצירה של האמן משה גרשוני העוסקת בעקדת יצחק** .**"ליצחקלה"** ' '
'Artwork by the artist Mose Grashoni that deals with Itzhak's binding,'

(b) *1982, 70x100 SUM.*
1982, 70 by 100 cm.
1982, 70 by 100 cm.
' .**ס"מ** , 1982, 70 x100 '

(c) **HCIWR** *HWA BBXINT XDIRH LMHWT*
painting-DEF is in-examining intrusion to-essence
HRGFIT FL TXWFT HIISWRIM FL HQWRBN.
emotional of feeling-GEN agony of victim-DEF
' **הציור הוא בבחינת חדירה למהות הרגשית של תחושת היסורים של הקורבן** . '
'The painting shows an incursion into the essence of emotional agony of the victim.'

In Modern Hebrew, the verbal agreement paradigm expresses definiteness, gender and number features that allow for the comprehension of the discourse content when a pronoun is dropped. The language is considered to be a 'partial pro-drop language' (Melnik 2007) where omissions of pronouns do occur in certain contexts. In our data, we found only a few examples where the subject pronoun is omitted, as exemplified in (9b).

(9)
- **DP:** [MMO ACT MMO MAT TMP DIM MMO REP]
- **CP:** [HH HH HH]

(a) **ICIRH** *FL HAMN IWRM RWZWB.*
artwork of artist-DEF yuram rozov.
'. יצירה של האמן יורם רוזוב '
'An artwork by the artist Yuram Rozov.'

(b) **BNWIH** *BCWRT DIPJIK , FMN EL BD,*
bulit-FEM-SG shape-GEN diptych , oil on cloth
1975, 70x110 SUM.
, 1975, 70 by 100 cm.
'. בנויה בצורת דיפטיך , שמן על בד , 1975 , 70x110 , ס"מ '
' is built in diptych form, oil on canvas, 1975, 70 by 100 cm.'

(c) **BICIRH ZW** *MWCGIM JIISIM BXLIPWT*
in-painting this depicted-MASC-PL pilots in-suit-PL
JISH.
flight
'. ביצירה זו מוצגים טייסים בחליפות טיסה '
'This work depicts pilots in flight suits.'

Pronoun is the most common linguistic unit of the reference expression in English. In the English data, we find many examples of this coreference tie. Typical types are demonstratives such as in (2a) and (10b).

(10)
- **DP:** [MMO MMO ACT MMO TIT PLC EVT TMP MMO]
- **CP:** [DH HH PRO]

(a) **This painting** *is* **a late work** *by Solario,*

(b) *and* **Ø** *is related to a large "Assumption of the Virgin" in the Certosa at Pavia, which was left unfinished on the artist's death in 1524.*

(c) **It** *is generally thought to reflect the knowledge of the work of Raphael gained by the artist.*

The coherence tie ellipsis was not analyzed extensively. However, we do mark its occurrence, for example in (10b), to show that it is commonly used when combining two semantic concepts, often in combination with the conjunction *and*. In Swedish, semantic concepts are

combined with relative clauses, in Hebrew with relative clauses or a comma. Also English tends to combine semantic concepts with commas, see example (2a). When two or more semantic concepts are combined with a conjunction, the subject is often omitted – this is common for all three languages.

3.2.4.3 Synonym

Several examples of synonym pairs exist in the Swedish data, as example (11) illustrates. It is less common in English and Hebrew where we found no examples of lexical-semantic relation synonymy.

(11)
- **DP:** [TIT MMO COL ACT MMO EVT]
- **CP:** [DH S]

(a) *Skolplansch "Smultron",* **tavla 11** *i*
school-wall-chart 'strawberry', painting-11 in
serien "Botaniska väggtavlor av J.
series-DEF "Botanical wall-painting-PL of J.
Eriksson", 25 st.
Eriksson", 25 pieces.
' Wild strawberry, painting 11 in the series "Botanical wall-painting-PL of J. Eriksson" 25 pieces.'

(b) **Bilden** *går även att se mer detaljerad med*
Picture-DEF exist also to see more detailed with
zoom-funktionen nere till vänster.
zoom-function-DEF down to left.
'The picture can also be seen in more details using the zoom function down to the left.'

Synonym words such as *bild* in (11b) are sometimes used to point to the content of the painting rather than to the physical object.

3.3 The results of the analysis

3.3.1 Syntactic structures

From the data analysis in section 3.2.1 we identified the syntactic structures that appear most frequently in the data and which all languages

have in common. Below we report on some examples of the most common syntactic structures observed in subject and direct object positions. These structures are covered in the application grammar for generating multilingual descriptions (see also chapter 9 in this thesis).

Subject position

- Proper nouns:
  Van Gogh
- Definite nouns:
  The portrait
  Målningen 'the painting'
  היצירה 'the artwork'
- Nouns preceding prepositions:
  a painting by Hals
  en målning av Gustaf Cederström 'a painting by Gustaf Cederström'
  יצירה של האמן 'painting by the artist'

Direct object position

- Numerical expressions preceding prepositions/prepositions preceding numerical expressions:
  dates from 1818
  utförd 1886 på Dalarö 'made 1886 in Dalarö'
  1886 צויר בשנת 'painted in 1886'
- Prepositions preceding nouns:
  displayed at the Paris Salon
  finns på Moderna museet 'is stored in the Moderna museet'
  מוצג במוזאון המודרני 'exhibited in the Moderna museet'

The syntactic analysis also showed that passive constructions are commonly used to describe artworks in the three languages. In Swedish, passive is usually formed with s-passive.

3.3.2 Discourse patterns

Empirical representations of the kind presented in sections 3.2.2 and 3.2.4 not only show how to lexicalize and combine semantic concepts according to the language specific patterns, but also guide discourse structures according to the discourse domain (chapter 9 in this thesis).

The analysis has shown how certain combinations of ontology statements are more appropriate for describing an artwork. Following these combinations, we were able to define a set of templates, each of which consists of different slot sequences. Each slot corresponds to a statement or a set of statements in the domain knowledge representation system. The templates and slots are specified in table 3.14. The template specification provides a set of ordering constraints over a pattern of statements in such a way as to yield a fluent and coherent output text. This approach is inspired by McKeown (1985) who described how to formalize principles of discourse for use in a computational process.

Following these pre-defined template specifications, we defined a discourse schema: *Description schema* (see below) consisting of two rhetorical predicates: *Identification–Predicate* and *Attributive–Predicate*.

Description schema:

Describe–Object $->$
  Identification–Predicate/
  Attributive–Predicate

Identification–Predicate $->$
  T1 , {T2 / T3}

Attributive–Predicate $->$
  T4 / T5

The description schema encodes the communicative goals and the structural relations that have been observed in the analyzed texts. Each predicate in the schema is associated with a set of templates (specified in table 3.14). The notation used to represent the schema is as follows: , indicates the mathematical relation *and*, *{}* indicates optionality, and / indicates alternatives.

Table 3.14: Template specification that governs text structures of a cultural object in a museum.

Name	Template slot
T1	(a) object's title | (b) object's creator
T2	(a) creator date of birth | (b) creator date of death
T3	(a) object id | (b) object material | (c) object size | (d) creation date | (e) creation place
T4	(a) current owner | (b) current location | (c) catalogue date | (d) collection
T5	(a) object's identifier | (b) identified place

3.3.3 Coreference strategies

The analysis of naturally occurring written examples described in section 3.2.3 has revealed a range of possibilities for constructing chains of coreference. Table 3.15 summarizes the hypothesized coreference strategies for generating reference forms. 1st, 2nd, and 3rd correspond to the first, second, and third reference expression pointing back to the main entity in the discourse. In summary, we found that:

- Pronoun is common in Swedish and English, and rare in Hebrew
- Direct-hyperonym is common in English, Swedish and Hebrew
- Higher-hyperonym is rare in English and Swedish, and common in Hebrew
- Synonym is common in Swedish, less frequent in English, and rare in Hebrew

Although the identified strategies are constrained by a relatively simple syntax and a domain ontology, they show clear differences between the languages. Direct-hyperonym in English appears after the antecedent or after a pronoun. In Swedish, direct-hyperonym follows or precedes empty categories. In some cases, we find occurrences of pronouns appearing after direct-hyperonym or empty categories. In Hebrew, higher-hyperonym usually appears after empty categories or higher-hyperonym. As table 3.15 shows, consecutive pronouns occur commonly in English, while consecutive higher-hyperonym noun phrases are common in Hebrew. The third row in the table is applied when there is a long distance between the referential expression and the main subject of reference, for instance when the sentence preceding the referential expression contains relative clauses.

Table 3.15: Coreference strategies.

DP	**CP**								
	English			Swedish			Hebrew		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
	DH	P		DH	P		DH	Ø	
	DH	HH	Ø	DH	Ø		DH	HH	
DP1	P	Ø	R	P	Ø	R	DH	Ø	R
	P	P	Ø	Ø	DH				
	P	P	DH	Ø	P	DH			
DP2				P	S	Ø			
							HH	HH	
				Ø	DH		HH	Ø	HH
DP3	P	P	DH	P	DH	DH			

The coreference strategies in table 3.15 are constrained by the following three discourse pattern principles:

- DP1: [MMO MMO ACT {[TMP], REP / REP, TMP }, [PLC]]
- DP2: [MMO, [TMP], [MMO], ACT, [DIM], PLC, REP]
- DP3: [MMO, [TMP], [MMO], {[ACT] [DIM] / [DIM] , [ACT]}, [REP], [PLC]]

To allow efficient computational processing relating to previous research on coherence in text generation (section 2.1.2), the coreference strategies in table 3.15 were simplified into table 3.16.

The generalized coreference strategies are motivated by semantic properties of the discourse. They follow the schema described in section 3.3.2 and are constrained by the distance between the referential expressions (table 3.16, first row). As it appears, sentence length influences the choice of the referential expression in all three languages.

As these simplified strategies illustrate, each strategy is instantiated differently depending on the language in question. There are two general principles that follow on from these simplified strategies:

(1) if the distance from the first occurrence of the entity in focus is greater than two sentences, repeat the entity in focus in definite form (R-DNP).

(2) if the distance from the first occurrence of the entity in focus is greater than two sentences and the RE in the previous sentence is a pronoun, or an empty category, use a pronoun in English and Swedish, and a direct-hyperonym in Hebrew.

Table 3.16: Generalization of the coreference strategies.

English			Swedish			Hebrew		
1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
PRO	PRO	R-DNP	DH	PRO	R-DNP	HH	HH	R-DNP
PRO	Ø	PRO	DH	Ø	PRO	DH	Ø	DH
DH	PRO	PRO	DH/S	DH/S	PRO	DH	HH	HH

3.3.4 Patterns of discourse and choice of RE

The analysis described in section 3.2.4 shows how semantic concepts are distributed over a discourse and the particular syntactic configurations of the instances of these concepts. The semantic concepts ACT and PLC often appear in direct object position, and are realized with a noun which requires a particular preposition in the three languages. For example, ACT often requires the preposition *by* in English, *av* 'by' in Swedish and with the genitive marker *šel* 'his' in Hebrew.

When studying the occurrences of semantic concepts we found semantic sequences that tend to co-occur in the same sentence. For example: ACT TMP, AAP ACT, TMP EVT. The maximum number of semantic concepts appearing in one sentence is five. In addition, we learned that the semantic concepts DIM and LOC rarely appear in the same sentence.

We also observed some linguistic differences regarding the choice of referential expressions. In English, demonstrative pronouns occur at the beginning of a description, regardless of the sentence's syntactic or semantic structure. Demonstrative and personal pronouns are commonly used to express the relations direct-hyperonymy and higher-hyperonymy. In Swedish, direct-hyperonyms usually appear when the antecedent is a compound. For example, the choice of the referential expression for *akvarellmålning* 'watercolour painting' is *målning* 'painting', for *oljeporträtt* 'oil portrait' is *porträtt* 'portrait'. In Hebrew, pronoun pro-drop occurs in sentences where the semantic concept MAT or TMP appears. Noun phares and demonstrative pronouns are commonly used to express the relations direct-hyperonymy and higher-hyperonymy.

3.4 Summary

3.4.1 Limitation of the study

Before discussing possible implications of the study, some limitations should be noted. Firstly, no comparison is made with other data is provided; we treat chains of referential forms as a phenomenon that is language-specific based merely on a rather narrow dataset, which may not necessarily be representative for the languages. Most examined texts were collected from specific websites so were presumably written by a small number of writers. Hence, the texts are based on a specific writing styles, which are not necessarily representative for most writings in the language. Secondly, the study is focused on a relatively small sample of object descriptions. To strengthen our findings, many more texts should be explored.

Other limitations of this study concern the semantic and syntactic annotation. According to the inter-annotator agreement calculations the manual semantic annotation was good. However, the annotation process would have been improved if inter-annotator agreement had been calculated during the annotation process. In that way, the annotators would have learnt where inconsistency occurred and improved their annotations for the remaining text.

As stated by previous researchers, it is difficult to compare two or more parsers since different parsers are designed to operate within frameworks based on different theoretical assumptions. Therefore, the figures provided in table 3.2 are not directly comparable. Closer inspection of the data revealed many tagging mistakes made on the collected data which also degraded the performance of the parsers. This is the cost of applying automatic methods. Some mistakes could have been avoided if part of speech tags were manually inspected and corrected.

3.4.2 Implications of the study

In spite of the above limitations, the results of this study are important from a computational linguistic point of view.

Our analysis showed which syntactic constructions are most frequent in the datasets, and how the examined semantic concepts are lexicalized. The semantic analysis revealed several semantic concepts that often precede each other. Following this observation, we were able to define a discourse schema that schematically describes how to struc-

ture descriptions about artworks. We also learned how these semantic concepts are combined syntactically.

We were able to identify a range of differences in how chains of coreference are constructed. These can be incorporated into any language generator which aims to produce coherent paragraph-sized descriptions. Table 3.15 summarizes the relations between an antecedent and an anaphoric entity revealed by the analysis. The entities that are usually chosen to express coreference in English are full noun phrases and demonstrative pronouns indicated by direct-hyperonymy relation. Coreference in Swedish is usually expressed by a synonymous noun, full noun phrase or a pronoun indicated by direct-hyperonymy. In Hebrew, higher-hyperonym is the most common lexical-semantic relation used to achieve coreference; pronouns are seldom used. Although, according to table 3.11, pronouns occur more frequently in the Swedish than in the English data, we found that English uses pronouns more often as a coreference tie, particularly demonstrative pronouns, to realize the first referential expression in the discourse.

Furthermore, it is often the case that a referring expression goes from a specific to more general description and not the other way round; this is a characterization that has been observed before. Ellipsis seems to be a rather frequent coreference tie in English, Swedish and Hebrew.

3.4.3 Conclusions

The study of naturally occurring texts in three languages has shown differences in how coreference chains are constructed in each language. We explored three lexical-semantic relations to investigate how their semantic properties can be accessed from an ontology. We narrowed the problem down to referential expressions pointing back to painting entities appearing in subject positions in a discourse.

From the corpus study presented in this section we gained domain and linguistic knowledge about: (1) the type of lexical information associated with semantic concepts; (2) the semantic information, which is relevant for the purpose of describing paintings; (3) syntactic knowledge about how to realize semantic concepts and referential expressions.

We identified some general principles that govern the distribution of referential forms and proposed simple procedures that are instantiated differently for each language. The kind of transition these procedures capture, contrary to Centering Theory, is Continuing.

We have also shown the CIDOC-CRM ontology schema provided by domain experts is understandable by layman users.

3.4.4 Future work

Further work could be carried out on the syntactically and semantically annotated data. This kind of work includes exploring syntactic differences between semantic categories. One could also explore how the ontology structure interacts with the syntactic structure of the language.

Much more could be done to extract information from the datasets and compare frequencies across the languages, for example, distribution of grammatical categories such as relative clauses, and units such as conjunctions that indicate coordination. There might be other sets of features, which play important roles for referential expression generation and could be used to annotate the data, such as the Centering discourse attributes, C_f and C_b. To enable further research, we intend to make the data freely available after some copy-write issues are resolved.

The discourse schema is formulated on the basis of a particular writing style with clear intentional goal of conveying a piece of predefined knowledge to a specific user group. It will be interesting to test how well it performs when generating descriptions about other artworks.

4 THE MLG DOMAIN APPLICATION

The multilingual domain application grammar presented in this section evolved from experiences with building an online grammar application within the MOLTO EU-project (Dannélls et al. 2012). Its design is based on previous knowledge we gained from developing a generation application that accommodates different user needs (see chapters 6 and 7 in this thesis).

4.1 Overview of the system

The purpose of our generation system is to produce coherent multilingual painting descriptions of Semantic Web ontology content by encoding linguistic knowledge about coreference principles. Figure 7 gives an overview of the natural language generator's components. The input to the generator is a set of predefined ontological statements, these are ordered according to the discourse patterns encoded in the abstract syntax. On top of the abstract syntax, three concrete syntaxes are built, one for each language. In every concrete syntax there is a specification of the coreference principles identified for each language. In section 4.3, these syntaxes are presented in more detail.

4.2 The application ontology

The application ontology employed in this work was designed to convey a fine-grained hierarchy of the concept *Painting* in an OWL compatible form. Its architecture, which represents a possible advance in the field of semantics in cultural heritage, is complementary to existing CH models, such as Europeana (Dekkers, Gradmann and Meghini 2009; Haslhofer and Isaac 2011),[27] the National Database Project of Norwe-

[27] <http://www.europeana.eu/portal/> (Last accessed: 2012-08-12)

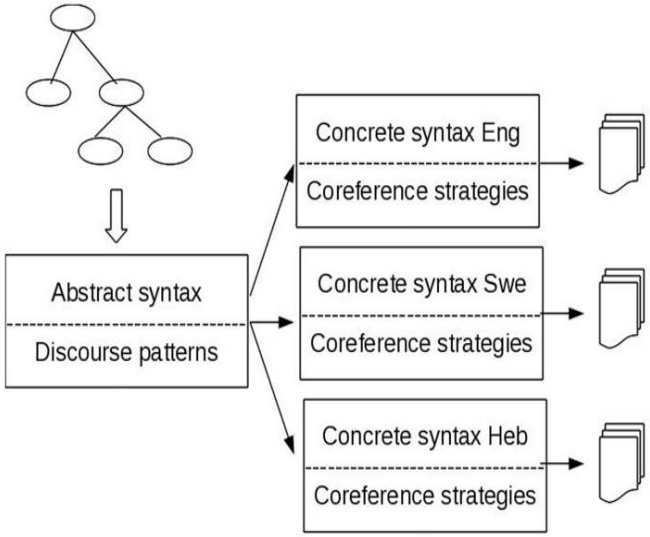


Figure 7: Overview of the natural language generator's components.

gian University Museums (Ore 2001).[28] and the Swedish Open Cultural Heritage (SOCH).[29] The idea behind the ontology development is to express semantic knowledge about paintings while supporting integration and interoperability with other ontology schemes (chapter 11 in this thesis). Our objective of developing the Painting ontology is to provide a feasible level of representation of the hierarchy of paintings in OWL.

The notion of painting

Collins English Dictionary (Sinclair 2001) gives three definitions for the term *painting*:

1. A picture which someone has painted.
2. The activity of painting a picture.
3. The activity of painting doors, walls and other parts of buildings.

[28] <http://www.muspro.uio.no/engelsk-omM.shtml> (Last accessed: 2012-08-12)
[29] <http://www.ksamsok.se/in-english/> (Last accessed: 2012-08-12)

The notion of *painting* that is adapted in this particular perspective corresponds to the first dictionary definition. The meaning conveyed in this definition refers to an intentionally created artifact entity. The term can be represented with the following class hierarchy:

$$\text{Painting} \subset \text{Artwork} \subset \text{Artifact}$$

We can further define the term *painting* as an artifact that is made up of *surface material*, and *paints*. Paints are identified by *mediums* that describe the nature of the painting and contain the information needed to identify and retrieve it. Mediums are associated with *painting techniques* that are applied by a human, or more specifically a *painter*. A Painting technique implies the *style* of the painting. Styles proclaim a particular *time period*, etc. Each of the entities given in italics are potential classes admitted in the painting ontology.

4.2.1 The construction of the ontology

The painting ontology was constructed manually using the Protégé editing tool, version 4.1.[30] It contains 94 classes and 98 properties. The main reference model of the painting ontology is the OWL 2 implementation of the domain specific CIDOC-CRM ontology.[31]

The Conceptual Reference Model (CRM) is an object oriented ontology developed by the International Council of Museum's Committee for Documentation (ICOM-CIDOC).[32] The model, consisting of about 90 classes and 148 properties (Doerr, Ore and Stead 2007; Doerr 2005), has been developed to help cultural and natural history organizations to store and share their data. It provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation (Crofts et al. 2009). Details are given in chapters 10 and 11 in this thesis.

4.2.2 Taxonomy and terminology specifications

As we mentioned in section 2.2.2, a common practice in ontology modeling is to divide the model into an intensional (TBox) and an extensional part (ABox). In this section we present the formal definitions and

[30] <http://protege.stanford.edu/> (Last accessed: 2012-12-12)
[31] <http://purl.org/NET/cidoc-crm/core> (Last accessed: 2012-12-12)
[32] <http://cidoc.ics.forth.gr/> (Last accessed: 2012-12-12)

axioms of terms in the TBox. DL expressions are used to present the terminology axioms in the ontology, CA denotes concept axioms and RA role axioms.

The taxonomy of paintings is considered under the CRM class `E22_Man-Made_Object` as specified in CA1 and CA2.[33]

(CA1) Artwork $\sqsubseteq$ E22_Man-Made_Object

(CA2) Painting $\sqsubseteq$ Artwork

Five classes for characterizing a painting were identified (these classes are put in italic): *Material*, what the painting is made of; *Painting Technique*, the technique that was applied to create the painting; *Dimension*, the size and form of the painting; *Device*, the instrument it was painted with; *RepresentedEntity*, what the painting depicts. Many of these classes are already defined in the CIDOC-CRM ontology (Crofts et al. 2009). The properties that link a painting object with each of these classes are: *hasMaterial*, *hasPaintingTechnique*, *hasDimension*, *isPaintedWith*, *depicts* (RA1–RA5).

(RA1) hasMaterial $\sqsubseteq$ P45_consists_of
≥ 1 hasMaterial $\sqsubseteq$ Painting $\top \sqsubseteq \forall$ hasMaterial.Material

(CA3) Material $\equiv$ itemMaterial $\sqcap$ E57_Material

To differentiate between the actual material that was applied to create a painting and other materials the painting is made of we have defined two subclasses of `Material`, namely `PaintMaterial` and `SurfaceMaterial` (CA4–CA5). This distinction supports generation of natural language phrases such as "oil on canvas", "a fabric made of cotton".

(CA4) PaintMaterial $\sqsubseteq$ E57_Material

(CA5) SurfaceMaterial $\sqsubseteq$ E57_Material

(RA2) hasPaintingTechnique
≥ 1 hasPaintingTechnique $\sqsubseteq$ Painting $\top \sqsubseteq \forall$ hasPaintingTechnique.PaintingTechnique

(CA6) PaintingTechnique $\equiv$ itemTechnique

[33] Classes that start with the upper-case letter 'E' followed by a number originate from CRM; classes that start with lower-case letters originate from SOCH; classes starting with upper-case letters are defined in the painting ontology.

(RA3) hasDimension $\sqsubseteq$ P43_has_dimension
≥ 1 hasDimension $\sqsubseteq$ Painting $\top \sqsubseteq \forall$ hasDimension.Dimension

(CA7) Dimension $\equiv$ itemMeasurement $\sqcap$ E54_Dimension

(RA4) isPaintedWith
≥ 1 isPaintedWith $\sqsubseteq$ Painting $\top \sqsubseteq \forall$ isPaintedWith.Device

(CA8) Device $\sqsubseteq$ itemClass

(RA5) depicts
≥ 1 depicts $\sqsubseteq$ Painting $\top \sqsubseteq \forall$ depicts.RepresentedThing

To support the large spectrum of objects and things a painting might depict, four classes have been defined (CA9–CA13).

(CA9) AnimateThing $\sqsubseteq$ RepresentedThing

(CA10) InanimateThing $\sqsubseteq$ RepresentedThing

(CA11) AbstractThing $\sqsubseteq$ InanimateThing

(CA12) Event $\sqsubseteq$ RepresentedThing

Event comprises all types of events, like the second world war, *State* contains descriptions of different physical non-dynamic things like state of a room in a certain time. A good practice in natural language processing is to differentiate between animate and inanimate entities, hence CA9 and CA10.

After having defined the classes and properties needed to describe paintings, we are able to define a range of paintings according to their classification types. The hierarchy of paintings is exemplified in CA13–CA26.

(CA13) AbstractPainting $\equiv$ Painting $\sqcap \exists$depicts.AbstractThing

(CA14) AcrylicPainting $\equiv$ Painting $\sqcap$ hasPaintType:AcrylicPaint

(CA15) BarkPainting $\equiv$ Painting $\sqcap$ hasMaterial:Bark

(CA16) FrescoPainting $\equiv$ Painting $\sqcap$ hasPaintingTechnique:Fresco

(CA17) LandscapePainting $\equiv$ Painting $\sqcap \exists$depicts.LandArea

(CA18) MiniaturePainting $\equiv$ Painting $\sqcap$
hasDimension:MiniaturePaintingHeight $\sqcap$
hasDimension:MiniaturePaintingWidth

(CA19) PencilPainting $\equiv$ Painting $\sqcap$ isPaintedWith:Pencil

(CA20) PortraitPainting $\equiv$ Painting $\sqcap$ $\exists$depicts.AnimateThing

(CA21) GroupPortrait $\sqsubseteq$ PortraitPainting

(CA22) GroupPortrait $\equiv$ PortraitPainting $\sqcap \geq 2$ $\exists$depicts.AnimateThing

(CA23) SelfPortrait $\sqsubseteq$ PortraitPainting

(CA24) SelfPortrait $\equiv$ PortraitPainting $\sqcap$ $\exists$depicts.Artist:a $\sqcap$ $\exists$createdBy.Artist:a

(CA25) StillLifePainting $\equiv$ Painting $\sqcap$ $\exists$depicts.InanimateThing

(CA26) ThangkaPainting $\equiv$ Painting $\sqcap$ $\exists$depicts.AnimateThing $\sqcap$ hasMaterial:TibetanSilk

Below follows a set of ontology assertion axioms which illustrate how the ontology classes are lexicalized in the painting ontology by using three computational lexicons: SALDO, PWN and HWN. The string *AnnotationAssertion* is abbriviated with the letters AA.

```
ClassAssertion(:PortraitPainting :BellelliFamilyObj)
AA(rdfs:label :PortraitPainting
                          "portrait%1:10:00::"@eng)
AA(rdfs:label :PortraitPainting "porträtt..1"@swe)
AA(rdfs:label :PortraitPainting
                             "dyokan:02956204"@heb)

ClassAssertion(:OilPainting :BellelliFamilyObj)
AA(rdfs:label :Painting
                  "oil_painting%1:06:00::"@eng)
AA(rdfs:label :Painting "oljemålning..1"@swe)
AA(rdfs:label :Painting "cyor semen:02956100"@heb)
```

A weakness of this lexicalization approach is that linguistic information is outside the logic at the present time. We cannot draw useful inferences about languages and thereby automatically acquire linguistic knowledge that is necessary to map to semantic structure and syntactic specifications. For example, with the right framework, we could infer from the ontology that Swedish lexical unit *oljemålning* of the the class *OilPainting* is a compound that incorporates its direct-hyperonym.

4.3 The abstract and concrete syntaxes

In this section we describe how the abstract and concrete syntaxes are constructed.

4.3.1 The abstract syntax

The top level abstract representation consists of two abstract syntaxes: `PaintingText` and `PaintingLex`. `PaintingText` contains semantic concepts encoded as categories and a function which takes these categories as arguments.[34]

PaintingText

```
cat GenDescription ; Painting ; Painter ;
    PaintingType ; Size ; Material ; Year ;
    Museum ; Collection ; OptSize ;
    OptMaterial; OptYear ;
    OptCollection ;  OptMuseum ;
```

In the above example we find the category: `GenDescription`, i.e. a top category for combining the remaining semantic concepts, of which eight are ontology classes: `Painting`, `Painter`, `PaintingType`, `Year`, `Size`, `Material`, `Collection`, `Museum`, and five concepts defined as optional (`Opt`) categories: `OptSize`, `OptMaterial`, `OptYear`, `OptCollection`, `OptMuseum`. Representing the ontology classes as optional allows the generator to cover discourse pattern variations by following the `Description schema` presented in section 3.3.2. The function `MkDescription` is implemented to generate some of the discourse variations covered by this schema.

```
fun MkDescription :
    Painting -> Painter -> PaintingType ->
    OptYear ->  OptSize -> OptMaterial ->
    OptCollection -> OptMuseum -> GenDescription ;
```

`MkDescription` takes eight arguments, the first one is the category of type `Painting` followed by the category `Painter` and `PaintingType`.

[34] I will refer to some of the semantic concepts with the term *ontology concepts/classes* although their names are not always equivalent to the classes defined in the ontology.

With this representation we are able to generate a description that always starts with presenting the painting object, its creator and type followed by optional categories. For each optional category there is a function which takes an ontology concept as argument. Thus, there are five functions of the kind defined below.

```
MkCollection : Museum -> OptCollection ;
```

The categories `Year` and `Size` are identified in the ontology with literals, such as number and date values. In GF, we represent these categories with `Int` values:

```
YInt : Int -> Year ;
SIntInt : Int -> Int -> Size ;
```

In `PaintingLex` one-place functions without arguments (called constants) are defined uniquely for each ontology class of `PaintingText`:

PaintingLex

```
OilPainting : PaintingType ;
PortraitPainting : PaintingType ;
SelfPortrait : PaintingType ;

ElisabethCzapek : Painter ;
AxelSparre : Painter ;

GIM: Collection ;
GSM: Collection ;

GoteborgsCityMuseum : Museum ;
MuseumOfWorldCulture : Museum ;

Paper : Material  ;
Wood : Material ;

GIM1026Obj : Painting ;
GIM1027Obj : Painting ;
```

For example, `PortraitPainting`, and `OilPainting` are constants of the category `PaintingType`. `ElisabethCzapek` and `AxelSparre` are constants of the category `Painter`, etc.

4.3.2 The concrete syntaxes

Each category and function introduced in `PaintingText` abstract syntax and in `PaintingLex` abstract syntax has a corresponding linearization type in the concrete syntax for the language. The three concrete syntaxes presented in this section are built on these two abstract syntaxes and are implemented independently for each target language.

4.3.2.1 English concrete grammar

The concrete syntax for English, `PaintingTextEng`, uses the resource grammar library categories and functions. `GenDescription` is linearized with `Text`, it is a top-level unit in the discourse. `Painting`, `Collection`, and `Museum`, are linearized with `NP`. `Painter` is linearized with `PN`. `PaintingType` is linearized with a record containing four CN fields: one field represents the actual lexical unit of the painting type, the remaining three fields represent the lexical-semantic relations: synonym, direct-hyperonym, and higher-hyperonym. This modular approach keeps part of the hierarchical knowledge separate from the whole ontology, which permits for efficient computation. `Material` is linearized with `CN`, and `Year` and `Size` are linearized with `Adv`. The names of these categories are specified in the API (Appendix E).

PaintingTextEng

```
lincat
   GenDescription = Text ;
   Painting, Collection, Museum = NP ;
   Painter = PN ;
   PaintingType =
     {st : CN ; sy : CN ; dh : CN ; hh : CN} ;
   Material = CN ;
   Year, Size = Adv ;
```

`Opt` categories are linearized with `OptAdv` (Optional Adverbial).

```
lincat
   OptCollection, OptSize, OptMaterial = OptAdv ;
   OptYear, OptMuseum = OptAdv ;
```

The category `OptAdv` is implemented as a record containing an adverbial string field and a Boolean field to express optionality.

```
oper
 OptAdv = {s : Adv ; isGiven : Bool} ;
 mkOptAdv : Adv -> OptAdv = \a ->
            {s = a ; isGiven = True} ;
 noAdv = {s = mkAdv [] ; isGiven = False} ;
```

The operation `mkOptAdv` specifies how to construct an `OptAdv` with the category `Adv` as argument. By default optional categories are linearized with `noAdv`. Below follows some examples of how optional categories are linearized:

```
lin
 MkCollection collection =
        mkOptAdv (mkAdv to_Prep collection) ;

 MkYear year = mkOptAdv year ;
 YInt i = mkAdv from_Prep (symb i) ;
```

Every category is constructed with a different preposition. A time string is constructed with the preposition *from*, a string describing a collection is linearized with the preposition *to*, etc. These instantiations are specific for the text to be generated.

In the same module we define how the function `MkDescription` is linearized.

```
 lin MkDescription painting painter paintingtype
      year dimension material collection museum =
 let
   s1 : Text = mkText ... ;
   s2 : Text = mkText ... ;
   ....
   s5 : Text = mkText ... ;
 in
  mkText s1 (mkText s2 (mkText s3 (mkText s4 s5)));
```

The functor `mkText` is used to linearize each sentence and finally concatenate all sentences, here, five distinct sentences s_1,..s_5. Each sentence combines different categories. For example, `s1` and `s2` are two sentences that together captures four categories.

```
s1 : Text = mkText (mkS presentTense
     (mkCl painting (mkVP
     (mkNP (mkNP a_Art paintingtype.st)
     (mkAdv by8agent_Prep (mkNP painter))))));

s2 : Text = mkText (mkS (mkCl it_NP
                   (mkVP (mkVP date_V) year.s)));
```

With this representation, the linearization of s_1 and s_2 yields: *'Hisingen' is a portrait by Brynolf Wennerberg. It dates from 1889.* The third sentence s_3 yields: *It measures 57 by 84 cm*:

```
s3 : Text = mkText (mkS (mkCl it_NP (mkVP
 (mkVP (mkVPSlash measure_V2)
 (mkNP (mkN ""))) size.s)));
```

In this implementation, the choice of form for a referential expression is approached from the perspective of distance from the antecedent, following the procedure described in chapter 3. In the remaining sentences we therefore consider the number of sentences generated so far and select the appropriate referential expression accordingly. In s_4 the choice of the referential expression is either direct-hyperonym or a pronoun, i.e. `paintingtype.dh` or `it_NP`.

```
s4 : Text =
 case <collection.isGiven, dimension.isGiven> of {
 <True,True> => mkText (mkS (mkCl
 (mkNP the_Art  paintingtype.dh) (mkVP
 (mkVP (mkVPSlash belong_V2)
 (mkNP (mkN ""))) collection.s))) ;
 <True, False> => mkText (mkS (mkCl it_NP (mkVP
 (mkVP (mkVPSlash belong_V2)
 (mkNP (mkN ""))) collection.s))) ;
 _ => emptyText
};
```

If two sentences have been generated which is indicated with the sequence `<True,True>` for the categories `Collection` and `Dimension`

then the next referential expression to be generated is a direct-hyperonym. Otherwise, if only one sentence has been generated, indicated with the sequence, `<True,False>` a pronoun is generated. The same procedure as in s_4 is applied in the fifth and last sentence s_5.

```
s5 : Text =
  case <museum.isGiven, dimension.isGiven,
                    collection.isGiven> of {
 <True,True,True> => mkText (mkS (mkCl it_NP
   (mkVP (passiveVP display_V2) museum.s))) ;
 <True,True,False> => mkText (mkS (mkCl
      (mkNP the_Art paintingtype.dh)
  (mkVP (passiveVP display_V2) museum.s))) ;
 <True,False,_> => mkText (mkS (mkCl
      (mkNP the_Art paintingtype.dh)
  (mkVP (passiveVP hold_V2) museum.s))) ;
  _ => emptyText
 };
```

Here, also, a pronoun or a direct-hyperonym is generated depending on the number of preceding sentences generated.

In the second concrete syntax for English, entitled `PaintingLexEng`, follows a specification of how to linearize each category defined in `PaintingLex`.

PaintingLexEng

```
lin

-- painting types

PortraitPainting =
      mkPaintingTypes (mkCN (mkN "portrait"));
OilPainting =
      mkPaintingTypes (mkCN (mkN "oil painting"));
SelfPortrait =
      mkPortraitTypes (mkCN (mkN "self portrait"));

-- painters

ElisabethCzapek = mkPN "Elisabeth Czapek";
AxelSparre = mkPN "Carl Axel Ambjörn Sparre";
```

```
-- collections

GIM = mkNP (mkPN "the Gothenburg Industry Museum");
GSM = mkNP (mkPN "the Gothenburg City Museum");

-- museums

GoteborgsCityMuseum =
        mkNP (mkPN "the Gothenburg Art Museum") ;
MuseumOfWorldCulture =
        mkNP (mkPN "the Museum of World Culture") ;

-- material

Paper = mkCN (mkN "paper") ;
Wood = mkCN (mkN "wood") ;

-- paintings

GIM1026Obj = mkPainting "'From a bird view'";
GIM1027Obj = mkPainting "'Constructions'";
```

Paintings are linearized with their titles, and are constructed with a noun phrase. This is a design choice made by the grammar developer to improve readability of the generated texts.

```
oper

 mkPainting : Str -> NP = \s -> symb s ;
```

`Paintingtype` carries information about its direct-hyperonym, higher-hyperonym and synonym strings. These are used in the grammar to generate the appropriate referential expression.

We have defined several operations for dealing with classes that appear in different levels in the ontology hierarchy. For example, the class `Self portrait` appears two levels below `Painting`, while `Portrait` and `Oil painting` appear only one level below. `mkPaintingTypes` and `mkPortraitTypes` specify how to linearize the appropriate painting types for each class.

```
oper

 mkPaintingTypes : CN ->
 {st: CN ; syn : CN ; dh : CN ; hh :CN} =
 \pst -> { st = pst;
           syn  = mkCN (mkN "picture") ;
           dh  = mkCN (mkN "painting") ;
           hh = mkCN (mkN "artwork")
} ;

 mkPortraitTypes : CN ->
  {st: CN ; syn : CN ; dh : CN ; hh :CN} =
  \pst -> { st = pst;
           syn  = mkCN (mkN "portaiture") ;
           dh  = mkCN (mkN "portrait") ;
           hh = mkCN (mkN "painting")
} ;
```

4.3.2.2 Swedish concrete grammar

The Swedish concrete syntax uses the RGL for Swedish and is very similar to the English implementation just described. The same linguistic categories are used to linearize categories defined in `PaintingText` and `PaintingLex`. We will therefore not go into the same details but rather highlight where the two syntaxes differ.

PaintingTextSwe

Coreference chains in Swedish must cater for pronoun agreement at the discourse level. For this purpose, `genPron` is defined. The function considers the gender of the `PaintingType` and guarantees the correct form of the pronoun. `genPron` is defined as follows:

```
 genPron : Gender => Pron = table {
     Neutr => it_Pron ;
     Utr   => it8utr_Pron } ;
```

`it_Pron` will generate the pronoun *det*, `it_8utr_Pron` will generate the pronoun *den*. Note that agreement need only be considered when generation extends across sentence boundaries. Using the resource grammar, gender and number agreements within a sentence boundary are generated correctly.

```
s4 : Text = case <collection.isGiven,size.isGiven>
 of {
 <True,True> => mkText (mkS (mkCl
  (mkNP the_Art  paintingtype.dh)
  (mkVP (mkVP tillhoera_vb_1_1_V) collection.s)));
 <True,False> => mkText (mkS (mkCl
  (mkNP (genPron ! paintingtype.g))
  (mkVP (mkVP tillhoera_vb_1_1_V) collection.s)));
  _ => emptyText
  };
```

The implementation strategy for encoding chains of coreference is implemented in a similar way as for English and follows the procedure defined for Swedish (section 3.3.3).

Lexical units in `PaintingTextSwe` are imported from a subset of the Saldo dictionary (section 2.3.2). A dictionary entry is indicated by a unique name and identity numbers corresponding to the different declinations. It has the following representation in GF:

```
fun tillhoera_vb_1_1_V : V ;
lin tillhoera_vb_1_1_V =
   mkV "tillhöra" "tillhörde" "tillhört" ;
```

There are some differences in the choice of preposition in the linearization of adverbial phrase. The Swedish verb *tillhöra* 'belong' does not take a preposition. These variations are sensitive to the choice of lexical units for expressing the ontology content.

```
MkCollection collection =
mkOptAdv (mkAdv no_Prep collection) ;
```

Functions in `PaintingLexSwe` are linearized with lexical units from the Saldo dictionary. They have the following representation:

PaintingLexSwe

```
PortraitPainting = mkCN portraettmaalning_nn_1_1_N ;
OilPainting = mkCN oljemaalning_nn_1_1_N ;
```

4.3.2.3 Hebrew concrete grammar

While it is possible to employ the resource grammar library to write the grammar for English and Swedish, the grammar rules for Hebrew

had to be defined with variables and inherent features explicitly. This is because the complete grammar resource of Hebrew is not yet available. However, a large part of the Hebrew morphology that is already available (Dannélls and Camilleri 2010), was reused in this implementation.

PaintingTextHeb

The categories defined in `PaintingText` are linearized as strings where the record *Str* is used as the simplest type. `PaintingType`, `Material` and `Museum` are lineraized with gender parameters.

```
lincat
 GenDescription = Str ;
 PaintingType =  {st : Str ; sy : Str ;
         dh : Str ; hh : Str ; g : Gender} ;
 Material, Museum = {s : Str ; g : Gender} ;
 Painting, Painter, Collection, Size, Year =  Str ;
 OptCollection, OptSize  = OptAdv ;
```

The only parameter type and attributes for nouns defined in this grammar is Gender (Masculine, Feminine). Other linguistic features that could be indicated as parameter types and attributes are definiteness (definite and indefinite) and number (singular, plural and dual).

The operator `OptAdv` is implemented as a record consisting of a string and a Boolean field.

```
oper
 OptAdv = {s : Str ; isGiven : Bool} ;
 mkOptAdv : Str -> OptAdv = \a ->
            {s = a ; isGiven = True} ;
```

`genPronIndef` and `genPronDef` have been defined to generate a pronoun in either definite or indefinite from.

```
 genPronIndef : Gender => Str =
            table {Masc => "zh"; Fem => "zw"} ;
 genPronDef : Gender => Str =
             table {Masc => "hzh"; Fem => "hzw"} ;
```

Definite noun phrases are constructed with `the_Art` as follows:

```
the_Art : Str -> Str = \st -> "h" + st ;
```

The function `MkDescription` is a concatenation of strings as opposed to `Text` as in the English and Swedish implementations:

```
lin MkDescription painting painter paintingtype year
               size material collection museum =
 let
  s1 : Str = ... ;
  s2 : Str = ... ;
   ...
  s5 : Str = ... ;
 in
  ss (s1 ++ "." ++ s2 ++ .. ++ "."++ s5 ++ ".") ;
```

The strategy for encoding chains of coreference is implemented in a similar way to English and Swedish, and follows the procedure defined for Hebrew (section 3.3.3).

```
s1 : Str = ({s =  painting.s ++
 (paint_V2.s ! Part ! Vp3Sg paintingtype.g) ++
  by_Prep.s ++ painter.s }).s ;

s2 : Str = ({s = the_Art paintingtype.hh ++
 (complete_V.s ! Perf ! Vp3Sg paintingtype.g) ++
                                      year.s}).s ;

s4 : Str = case <collection.isGiven,size.isGiven>
 of {
  <True,True> => ({s = paintingtype.hh ++
         genPronIndef ! paintingtype.g ++
   (belong_V.s ! Part ! Vp3Sg paintingtype.g) ++
                              collection.s}).s ;
  <True,False> => ({s = paintingtype.hh ++
  genPronIndef ! paintingtype.g ++
  (belong_V.s ! Part ! Vp3Sg paintingtype.g) ++
                             collection.s}).s ;
   _ => emptyText
} ;
```

Verb lexemes are inflected for tense, person, number, and gender. Attributes for verb tense are: Perfect `Perf` (past tense), Participle `Part` (present tense), Imperfect `Imp` (future tense). These features are specified explicitly to generate the correct verb form.

Hebrew has seven verb pattern groups (called binyanim) which are associated with a fixed morphological form, e.g. pa'al: `C1aC2aC3`, pi'el: `C1iC2eC3`. The lexical representations of the Hebrew verbs are based on these groups. They are encoded in the lexicon as follows:

```
 paint_V2 = dirV2 (mkVPiel "Zyr") ;
 belong_V = mkVHifhil "syk" ;
 complete_V = mkVHifhil "slm" ;
```

Linearization of functions in `PaintingLexHeb` have the following representations:

PaintingLexHeb

```
PortraitPainting = mkPaintingTypes "dywqN" Masc ;
OilPainting =  mkPaintingTypes "Sywr smN" Masc ;
SelfPortrait = mkPortraitTypes "dywqN OZmy" Masc;

AxelSparre = {s ="kArl ASl Sprh"} ;
GIM = {s = "kwlqZyh sl mwzAwN htOsyh sl gwTnbrg"} ;
GoteborgsCityMuseum =
               {s = "mwzyAwN gwTnbrg" ; g = Masc} ;
Wood = {s = "OZ." ; g = Masc} ;
GIM1026Obj = mkPainting "mrAh sl Zypwr" ;
GIM1027Obj = mkPainting "bnywt" ;
```

`mkPainting` and `mkPaintingTypes` include a gender category and are defined as follows:

```
oper
 mkPainting : Str ->
  {s : Str; g : Gender} =
             \st -> { s = st ; g = Masc} ;

 mkPaintingTypes : Str -> Gender ->
  {st: Str ; syn : Str ; dh : Str ;
           hh : Str ; g : Gender} =
  \pst,gst -> {  st = pst ; g = gst ;
           syn  =  "tmona" ; g = Fem ;
           dh  = "Sywr" ; g = Masc ;
           hh = "ySywrh" ; g = Fem } ;

 mkPortraitTypes : Str -> Gender ->
  {st: Str ; syn : Str ; dh : Str ;
           hh : Str ; g : Gender} =
  \pst,gst -> {  st = pst ; g = gst ;
           syn  =  "AwTwpwrTrT" ; g = Masc ;
           dh  = "dywqN" ; g = Masc ;
           hh = "Sywr" ; g = Masc   } ;
```

`mkPaintingTypes` and `mkPortraitTypes` linearize each painting type with a string and a gender. Note, a different gender is linearized depending on the lexical unit, i.e. *tmona* and *ySywrh* have gender Feminine, *Fem*.

4.4 A generation example

Below follows a set of ontology statements from which to generate coherent multilingual descriptions.

```
<owl:Thing rdf:about="&painting;GSM940051Obj">
<rdf:type
   rdf:resource="&painting;WatercolourPainting"/>
<createdBy
   rdf:resource="&painting;BrynolfWennerberg"/>
<Dimension
   rdf:resource="&painting;GSM940051Dimension"/>
<hasCreationDate
   rdf:resource="&painting;GSM940051CreationDate"/>
<belongsTo rdf:resource="&painting;GSM"/>
<hasCurrentLocation
   rdf:resource="&painting;GoteborgsCityMuseum"/>
<hasUnitOfMeasure
      rdf:resource="&painting;Centimeter"/>
</owl:Thing>

<owl:Thing rdf:about="&painting;GSM940051Dimension">
<rdf:type rdf:resource="&painting;Dimension"/>
<heightValue
    rdf:datatype="&rdfs;Literal">57</heightValue>
<widthValue
    rdf:datatype="&rdfs;Literal">84</widthValue>
<hasUnitOfMeasure
    rdf:resource="&painting;Centimeter"/>
</owl:Thing>

<owl:Thing
      rdf:about="&painting;GSM940051CreationDate">
<rdf:type rdf:resource="&painting;TimePeriod"/>
<fromTimePeriodValue rdf:datatype="&rdfs;Literal">
```

```
 1889</fromTimePeriodValue>
<toTimePeriodValue rdf:datatype="&rdfs;Literal">
 1889</toTimePeriodValue>
</owl:Thing>

<owl:Thing rdf:about="&painting;GSM">
<rdf:type rdf:resource="&painting;Collection"/>
</owl:Thing>

<owl:Thing
     rdf:about="&painting;GoteborgsCityMuseum">
<rdf:type rdf:resource="&painting;MuseumBuilding"/>
<hasPermanentLocation
           rdf:resource="&painting;Gothenburg"/>
</owl:Thing>
```

The above ontology specifications are converted into verified trees which are sent as input to the language generator (see chapter 10 in this thesis). A verified tree has the following syntax:

```
MkDescription GSM940051Obj BrynolfWennerberg
  WatercolourPainting (MkYear (YInt 1889))
  (MkSize (SIntInt 57 84)) NoMaterial
  (MkCollection GSM) (MkMuseum GoteborgsCityMuseum)
```

Lexical information is encoded manually and separately for each language in the corresponding `PaintingLex` module. Texts are then generated by following the concrete syntax for each language. Some examples are given below.

1. 'Hisingen' is a watercolour painting by Brynolf Wennerberg from 1889. It measures 57 by 84 cm. It belongs to the Gothenburg Art Museum collection. The painting is displayed in the Gothenburg Art Museum.
2. 'Hisingen' är en akvarell av Brynolf Wennerberg från år 1889. Målningen är 57 cm lång och 84 cm bred. Den tillhör Göteborgs stadsmuseum's samling. Målningen återfinns på Göteborgs stadsmuseum.
   ' 'Hisingen' is a watercolour painting by Brynolf Wennerberg from the year 1889. The painting is 57 cm long and 84 cm wide. It belongs to the Gothenburg Art Museum collection. The painting is displayed in the Gothenburg Art Museum.'

3.

היסינגאן הוא ציור שמן על ידי ברינוף ונרברג משנת 1889.
התמונה בגודל 84 על 57 ס״מ .
היצירה שיכת לקולקציה של מוזיאון האומנות של גוטנבורג .
היצירה מוצגת במוזאון גוטנבורג .

'Hisingen is a watercolour painting by Brynolf Wennerberg from the year 1889. The size of the picture is 57 by 84 cm. The artwork belongs to the Gothenburg Art Museum collection. The artwork is displayed in the Gothenburg Art Museum.'

4.5 Experiments and evaluation

This section presents the pilot experiments that were carried out to test the output results of applying language-specific coreference strategies.

4.5.1 Experiment 1

The data available in the painting ontology (see chapter 14 in this thesis) was exploited to generate descriptions of paintings automatically in the three languages. The descriptions were generated by following the domain specific coreference strategies, which are constrained by the discourse pattern principles (table 3.15).

The purpose of the experiment was to test whether readers prefer pronouns over other linguistic elements in two, three and four sentence-long discourses, and to investigate whether the semantic content has an impact on the preference of the referential expression.

4.5.1.1 Generated descriptions

Forty description pairs were generated in each language. One description containing only pronouns as the type of referring expression; it was presented on the left column throughout the evaluation. One description containing different referring expressions that were automatically generated by applying the language-specific coreference strategies; it was presented on the right column throughout the evaluation. Examples of the generated description pairs are given in chapter 15 in this thesis. The English example of the material each subject received is given in figure 8.

Please look at each description-pair in each row in the table below and put an x on the left hand-side of the description you find most coherent and natural.

You are encouraged to choose one description per row.
Thank you for your participation.

	"The Large Square" is painted by Anna Lindskog. It is from 1885.		"The Large Square" is painted by Anna Lindskog. The painting is from 1885.
	"The Large Square" is painted by Anna Lindskog in 1885. It is displayed in the Gothenburg City Museum.		"The Large Square" is painted by Anna Lindskog in 1885. The painting is displayed in the Gothenburg City Museum.
	"The Large Square" is painted by Anna Lindskog. It dates from 1885. It is displayed in the Gothenburg City Museum.		"The Large Square" is painted by Anna Lindskog. The painting dates from the 1885. It is displayed in the Gothenburg City Museum.

Figure 8: A screenshot of the survey of experiment 1.

4.5.1.2 Evaluators

Nine human subjects participated in the evaluation, three native speakers of each language. The subjects were asked to make a forced choice between two versions of each description and mark the one they found most coherent and idiomatic based on their intuitive judgements.

4.5.1.3 Evaluation results

The results of the evaluation are reported in table 4.1. The left-hand column shows the number of times descriptions containing only pronouns as the form of referential expression were preferred by the evaluators. The right-hand column shows the number of times descriptions containing different forms of language-specific referential expressions were preferred the evaluators. From this table we learn that the evaluators approved at least half of the automatically generated descriptions.

4.5.1.4 Summary of the results

A closer look at the examples where chains of pronouns were preferred revealed that these occurred in English when a description consisted of two or three sentences and the second and third sentences specified the painting dimensions or a date. In Swedish, these were preferred whenever a description consisted of two sentences. In Hebrew, the evaluators preferred a description containing a pronoun over a description con-

Table 4.1: Results of pairwise evaluation.

	Pronouns	Pronouns/DH/HH/S
English	17	18
Swedish	9	29
Hebrew	6	28

taining higher-hyperonym NP when the description consisted of two sentences, the second of which concerned the painting dimensions.

4.5.2 Experiment 2

The same data as in experiment 1 was exploited in this experiment. The simplified procedures for generating chains of referential expressions depicted in table 3.16 were applied to generate descriptions of paintings automatically in the three languages. The purpose of the experiment was to test whether coreference strategies are superior when language and strategy match.

4.5.2.1 Generated descriptions

Thirty description triples were generated by mixing the three language specific strategies (three for each language). One description was generated by applying the language dependent coreference strategies, two additional descriptions were generated by applying the language dependent coreference strategies of the other two languages. The nine following combinations have been applied:

1. English descriptions produced by applying English (EE), Swedish (ES), and Hebrew (EH) coreference strategies;
2. Swedish descriptions produced by applying Swedish (SS), English (SE) and Hebrew (SH) coreference strategies;
3. Hebrew descriptions produced by applying Hebrew (HH), English (HE), and Swedish (HS) coreference strategies.

Below follow examples of the generated texts used in the evaluations.

English descriptions:

(EE) 'The sugar factory' is a painting by Carl Axel Ambjörn Sparre. It dates from 1885. It is displayed in the Museum of World Culture.

(EH) 'The sugar factory' is a painting by Carl Axel Ambjörn Sparre. The work dates from 1885. The work is displayed in the Museum of World Culture.

(ES) 'The sugar factory' is a painting by Carl Axel Ambjörn Sparre. The painting dates from 1885. It is displayed in the Museum of World Culture.

Swedish descriptions:

(SS) 'Sockerfabriken' är en målning av Carl Axel Ambjörn Sparre. Målningen är från år 1885. Den är utställd på Världskulturmuseet.
'The sugar factory is a painting by Carl Axel Ambjörn Sparre. The painting is from the year 1885. It is displayed in the Museum of World Culture.'

(SE) 'Sockerfabriken' är en målning av Carl Axel Ambjörn Sparre. Den är från 1885. Den är utställd på Världskulturmuseet.
'The sugar factory is a painting by Carl Axel Ambjörn Sparre. It is from the year 1885. It is displayed in the Museum of World Culture.'

(SH) 'Sockerfabriken' är en målning av Carl Axel Ambjörn Sparre. Konstverket är från 1885. Konstverket är utställt på Världskulturmuseet.
'The sugar factory is a painting by Carl Axel Ambjörn Sparre. The artwork is from the year 1885. The artwork is displayed in the Museum of World Culture.'

Hebrew descriptions:

(HH) בית חרושת של סוכר הוא ציור של קארל אקסל אמביורן שפארה.
היצירה הושלמה בשנת 1885. היצירה מוצגת במוזיאון של עולם התרבות.
'The sugar factory is a painting by Carl Axel Ambjörn Sparre. The artwork was completed in the year 1885. The artwork is displayed in the Museum of World Culture.'

(HS) בית חרושת של סוכר הוא ציור של קארל אקסל אמביורן שפארה.
התמונה הושלמה בשנת 1885. היא מוצגת במוזיאון של עולם התרבות.
'The sugar factory is a painting by Carl Axel Ambjörn Sparre. The picture was completed in the year 1885. It is displayed in the Museum of World Culture.'

(HE) בית חרושת של סוכר הוא ציור של קארל אקסל אמביורן שפארה.
הוא הושלם בשנת 1885. הוא מוצג במוזיאון של עולם התרבות.
'The sugar factory is a painting by Carl Axel Ambjörn Sparre. It was completed in the year 1885. It is displayed in the Museum of World Culture.'

4.5.2.2 Evaluators

Eighteen evaluators participated in web survey evaluations,[35] six native speakers in each language. The age of the evaluators ranged from 27 to 70. Participants were asked to rank the descriptions by following their intuitive judgments. They were encouraged to rank each description according to the following three scales: 1. coherent and formal; 2. coherent and less formal; 3. less coherent and informal. Figure 9 shows an example of the English web-based survey. In the following section we present the evaluation results for each language.

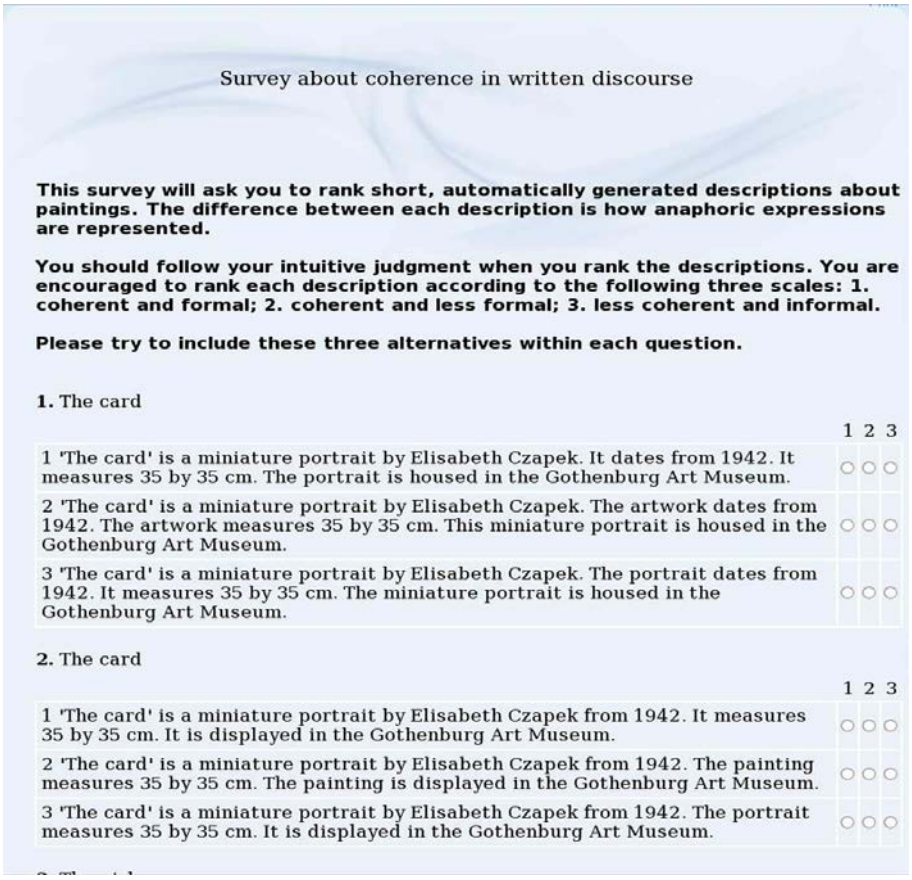


Figure 9: A screenshot of the English web-based survey.

[35]The three surveys were published via Webropolsurveys <https://www.webropolsurveys.com>

4.5.2.3 Evaluation results

According to the English evaluation results (figure 10 and table 4.2) there is a strong preference for descriptions produced with Swedish strategies. A closer look at the results shows, some readers found descriptions produced with Swedish strategies as good as the ones produced by applying English strategies. This is illustrated below (ranking values for each sentence are presented in parentheses: first, second, third).

(EE) 'Peter Ulrik Ekström' is a miniature portrait by J.K.F. Viertel. It dates from 1814. It measures 15 by 15 cm and is housed in the Gothenburg Art Museum. (3,3,0)

(EH) 'Peter Ulrik Ekström' is a miniature portrait by J.K.F. Viertel. The work dates from 1814. The painting measures 15 by 15 cm and is housed in the Gothenburg Art Museum. (0,1,5)

(ES) 'Peter Ulrik Ekström' is a miniature portrait by J.K.F .Viertel. The portraiture dates from 1814. It measures 15 by 15 cm and is housed in the Gothenburg Art Museum. (3,1,1)

Swedish strategies are preferred when three or four semantic concepts are realized in the preceding sentence. For example:

(EE) 'Sigrid Heurlin' was painted on wood by Eva Bonnier in 1886. It measures 65 by 51 cm. It is displayed in the Gothenburg Art Museum. (1,4,1)

(EH) 'Sigrid Heurlin' was painted on wood by Eva Bonnier in 1886. The painting measures 65 by 51 cm. The work is displayed in the Gothenburg Art Museum. (0,2,4)

(ES) 'Sigrid Heurlin' was painted on wood by Eva Bonnier in 1886. The portrait measures 65 by 51 cm. It is displayed in the Gothenburg Art Museum. (4,0,2)

There is also a strong preference for Swedish strategies when there are only two referring expressions in the discourse. As it appears, English readers do not prefer consecutive pronouns when a description is only two three sentence long. This is illustrated below.

Table 4.2: Evaluation results for English.

	1	2	3
EE	62 (35.2 %)	**100 (56.8 %)**	14 (7.9 %)
ES	**97 (54 %)**	29 (16.2 %)	53 (30 %)
EH	19 (10.7%)	50 (28.4 %)	**105 (59.6 %)**

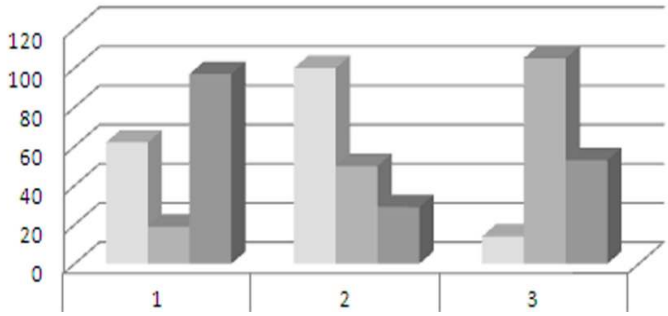


Figure 10: Illustration of the evaluation results for English.

(EE) 'The girl' is an oil portrait by Elisabeth Czapek from 1942. It measures 435 by 365 cm. It is displayed in the Gothenburg Art Museum. (1,3,2)

(EH) 'The girl' is an oil portrait by Elisabeth Czapek from 1942. The painting measures 435 by 365 cm. The painting is displayed in the Gothenburg Art Museum. (1,3,2)

(ES) 'The girl' is an oil portrait by Elisabeth Czapek from 1942. The portrait measures 435 by 365 cm. It is displayed in the Gothenburg Art Museum. (5,0,1)

According to the Swedish evaluation results (figure 11 and table 4.3), there is a strong preference for descriptions produced with the language strategies. Descriptions produced with Hebrew strategies have been highly ranked in several examples where the second sentence specified the painting dimension or the material, for example:

(SS) 'Edit Hedin' är ett oljeporträtt av Eva Bonnier från år 1889. Porträttet är 70 cm långt och 45 cm brett. Det återfinns på Världskulturmuseet. (4,2,0)
' 'Edit Hedin' is an oil portrait by Eva Bonnier from the year 1889. The portrait is 70 cm long and 45 cm wide. It is hosted in the Museum of World Culture. '

(SE) 'Edit Hedin' är ett oljeporträtt av Eva Bonnier från år 1889. Det är 70 cm långt och 45 cm brett. Det återfinns på Världskulturmuseet. (0,0,6)
' 'Edit Hedin' is an oil portrait by Eva Bonnier from the year 1889. It is 70 cm long and 45 cm wide. It is hosted in the Museum of World Culture. '

(SH) 'Edit Hedin' är ett oljeporträtt av Eva Bonnier från år 1889. Målningen är 70 cm lång och 45 cm bred. Målningen återfinns på Världskulturmuseet. (3,2,1)
' 'Edit Hedin' is an oil portrait by Eva Bonnier from the year 1889. The painting is 70 cm long and 45 cm wide. The painting is hosted in the Museum of World Culture. '

Many of the evaluated examples show the evaluators find descriptions produced with Hebrew strategies to be just as good as the ones produced with Swedish strategies.

(SS) 'Peter Ulrik Ekström' är ett miniaturporträtt av J.K.F. Viertel från år 1814. Porträttet är utställt på Göteborgs stadsmuseum.(3,3,0)
' 'Peter Ulrik Ekström' is a miniature painting by J.K.F. Viertel from the year 1814. The portrait is displayed in the Gothenburg Art Museum. '

(SE) 'Peter Ulrik Ekström' är ett miniaturporträtt av J.K.F. Viertel från år 1814. Det är utställt på Göteborgs stadsmuseum.(0,3,3)
' 'Peter Ulrik Ekström' is a miniature painting by J.K.F. Viertel from the year 1814. It is displayed in the Gothenburg Art Museum. '

(SH) 'Peter Ulrik Ekström' är ett miniaturporträtt av J.K.F. Viertel från år 1814. Konstverket är utställt på Göteborgs stadsmuseum.(3,0,3)
' 'Peter Ulrik Ekström' is a miniature painting by J.K.F. Viertel from the year 1814. The artwork is displayed in the Gothenburg Art Museum. '

Hebrew strategies were also preferred in cases where there was a mixture of gender, for example *konstverket* 'the artwork' has a neuter gender while *akvarellmålning* 'watercolour painting' is non-neuter.

(SS) 'Fem delar' är en akvarellmålning av Karl Larsson. Målningen är utförd på duk. Den är från år 1895.(3,3,0)
'Five pieces is a watercolour painting by Karl Larsson. The painting is made on canvas. It is from the year 1895.'

Table 4.3: Evaluation results for Swedish.

	1	2	3
SS	**107 (60 %)**	55 (31 %)	15 (8.4 %)
SE	10 (5.6 %)	51 (28 %)	**117 (65.7 %)**
SH	54 (30.6 %)	59 (33.5 %)	**63 (35.7 %)**

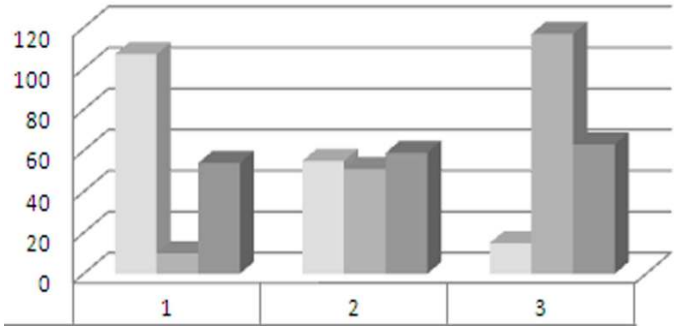


Figure 11: Illustration of the evaluation results for Swedish.

(SE) 'Fem delar' är en akvarellmålning av Karl Larsson. Den är utförd på duk. Den är från år 1895.(0,2,4)
'Five pieces is a watercolour painting by Karl Larsson. It is made on canvas. It is from the year 1895.'

(SH) 'Fem delar' är en akvarellmålning av Karl Larsson. Konstverket är utfört på duk. Konstverket är från år 1895.(3,1,2)
'Five pieces is a watercolour painting by Karl Larsson. The artwork is made on canvas. The artwork is from the year 1895.'

Other examples where descriptions produced with Swedish and Hebrew strategies were almost equally ranked when a description consisted only of two sentences.

(SS) 'Fem delar' är en målning på duk av Karl Larsson. Målningen finns på Världskulturmuseet.(3,3,0)
'Five pieces is a painting on canvas by Karl Larsson. The painting is hosted in the Museum of World Culture.'

(SE) 'Fem delar' är en målning på duk av Karl Larsson. Den finns på Världskulturmuseet. (0,1,5)
'Five pieces is a painting on canvas by Karl Larsson. It is hosted in the Museum of World Culture.'

(SH) 'Fem delar' är en målning på duk av Karl Larsson. Konstverket finns på Världskulturmuseet.(3,2,1)
'Five pieces is a painting on canvas by Karl Larsson. The artwork is hosted in the Museum of World Culture.'

According to the Hebrew evaluation results (figure 12 and table 4.4), Hebrew descriptions produced by applying Swedish strategies received the highest scores by the evaluators. These preferences occurred when there was a mixture of gender, i.e. when the referential expression has had a different gender than the antecedent. Some examples are:

(HH) חמש חתיכות הוא ציור מים על ידי קארל לרסון.
התמונה בגודל 58 על 48 ס״מ. היצירה מוצגת
במוזיאון של עולם התרבות. (0,3,3)

'Five pieces is a watercolour painting by Karl Larsson. The size of the picture is 48 by 58 cm. The artwork is displayed in the Museum of World Culture.'

(HS) חמש חתיכות הוא ציור מים על ידי קארל לרסון.
הציור בגודל 58 על 48 ס״מ. הוא מוצג
במוזיאון של עולם התרבות. (5,0,1)

'Five pieces is a watercolour painting by Karl Larsson. The size of the painting is 48 by 58 cm. It is displayed in the Museum of World Culture.'

(HE) חמש חתיכות הוא ציור מים על ידי קארל לרסון.
הוא בגודל 58 על 48 ס״מ. הוא מוצג
במוזיאון של עולם התרבות. (1,3,2)

'Five pieces is a watercolour painting by Karl Larsson. It's size is 48 by 58 cm. It is displayed in the Museum of World Culture.'

In many of the examples, the evaluators found the Hebrew descriptions produced with Swedish strategies just as good as the ones produced with Hebrew strategies.

(HH) חמש חתיכות הוא ציור מים שצויר על ידי קארל לרסון.
היצירה הושלמה בשנת 1895. (3,2,1)

'Five pieces is a watercolour painting painted by Karl Larsson. The artwork was completed in the year 1895.'

(HS) חמש חתיכות הוא ציור מים שצויר על ידי קארל לרסון.
התמונה הושלמה בשנת 1895. (3,2,1)

'Five pieces is a watercolour painting painted by Karl Larsson. The picture was completed in the year 1895.'

Table 4.4: Hebrew evaluation results.

	1	2	3
HH	**88 (48 %)**	60 (32.7 %)	35 (19 %)
HS	**93 (51 %)**	69 (37.9 %)	20 (10.9 %)
HE	29 (16.3 %)	54 (30.5 %)	**94 (53 %)**

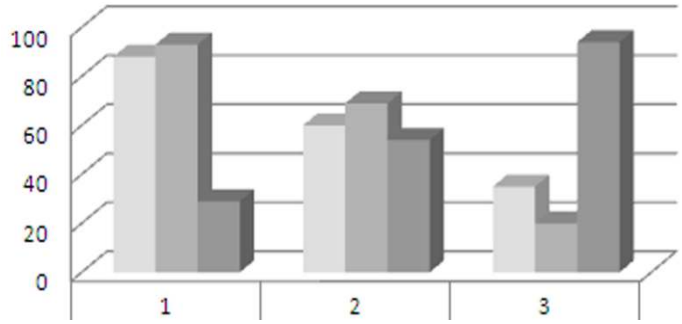


Figure 12: Illustration of the evaluation results for Hebrew.

(HE) **חמש חתיכות הוא ציור מים שצויר על ידי קארל לרסון.**
הוא הושלם בשנת 1895. (1,2,3)

'Five pieces is a watercolour painting painted by Karl Larsson. It was completed in the year 1895.'

There was a strong preference for descriptions produced with Hebrew strategies when a description was three or four sentences long.

(HH) **החוף צויר על ידי אוה בונייר בשנת** 1905.
היצירה שייכת למוזיאון התעשיה של גוטנברג.
היצירה מוצגת במוזיאון האומנות של גוטנברג. (5,1,2)

'The coast was painted by Eva Bonnier in the year 1905. The artwork belongs to the Gothenburg Industry Museum. The artwork is displayed in the Gothenburg Art. '

(HS) **החוף צויר על ידי אוה בונייר בשנת** 1905.
התמונה שייכת למוזיאון התעשיה של גוטנברג.
היא מוצגת במוזיאון האומנות של גוטנברג. (1,4,1)

'The coast was painted by Eva Bonnier in the year 1905. The picture belongs to the Gothenburg Industry Museum. It is displayed in the Gothenburg Art. '

(HE) החוף צויר על ידי אוה בונייר בשנת 1905.
הוא שייך למוזיאון התעשיה של גוטנברג.
הוא מוצג במוזיאון האומנות של גוטנברג. (1,1,4)

'The coast was painted by Eva Bonnier in the year 1905. It belongs to the Gothenburg Industry Museum. It is displayed in the Gothenburg Art. '

4.5.2.4 Summary of the results

The evaluation results of experiment 2 are summarized in table 4.5. Given the results of the discourse analysis provided in chapter 3, our assumption was that language strategies would be superior when language and strategy matched, therefore the final evaluation results are not exactly as predicted. As table 4.5 shows, the descriptions produced with Swedish strategies were favored by most evaluators in the three languages.

From these results we learn that readers prefer to realize a referential expression with a direct-hyperonym. Apparently, this relation that introduces a more general term increases the degree of coherence. The results also suggest that consecutive referential expressions of the same type do not contribute to the coherence of a text. The results given for Hebrew indicate that readers are likely to find discourses produced with the language strategies coherent, but this does not preclude other strategies resulting in a coherent text.

4.6 Discussion

The purpose of the painting ontology presented in this section is to store and present detailed information about paintings in the framework of the Semantic Web. The ontology contains only 94 classes and 98 properties and hence it is a relatively small knowledge base. We showed how its classes are integrated with classes in the CIDOC-CRM ontology and described the additional classes and properties needed to provide a meaningful semantic definition of the concept *Painting*. The ontology provides support for linking multilingual lexicons.

Although the painting ontology is not linguistically motivated, it can be used to encode linguistic knowledge because it rests on theories supporting standard frameworks, such as the ones found in Declerck et al. 2010, and Cimiano et al. 2011 for modeling and representing computational lexicons.

Table 4.5: Confusion metric for mixing coreference strategies.

	E	S	H
E	35.2 %	**54 %**	10.7 %
S	5.6 %	**60 %**	30.6 %
H	16.3 %	**51 %**	48 %

The generation application has several advantages. One is the ability to generate different patterns from one function, i.e. `MkDescription`. The ontology classes, such as the hierarchical structure of the semantic concept Painting, are efficiently encoded in the system as categories. Another is the generator's ability to produce texts with different syntactic constructions by combining different semantic categories (see chapters 12 and 13 in this thesis). What enhances this endeavour is the RGL, which facilitates the grammar writing. Using the RGL, it is easy to extend the syntactic coverage of the grammar and build new concrete syntaxes for other languages.

Concerning the experiments and the evaluation results, there are many factors that encourage different people to use different referential expressions when talking about things. The fact some evaluators chose to assign low rankings to certain discourses, does not necessarily mean these discourses were incoherent. Perhaps the guidelines were not clear enough for some of the untrained evaluators, who misunderstood the language properties they should consider. Another possible explanation is the difference between the patterns observed in the data analysis and user preferences. Something we discuss in section 3.4.1.

A better evaluation standard would have been to ask the evaluators to rate separately how coherent, formal and idiomatic the generated texts are, similar to the experiments as described by Reiter and Belz (2009). Perhaps, even make a further distinction between the different coherence types, such as grammatical and lexical. Another possibility is to evaluate the generated texts against the original texts. However, with an evaluation against a gold standard one may be confronted with the question of whether we should draw the conclusion that a text is good and coherent, because it is close to the gold standard? What counts as informative referential expression has cultural, sociolinguistic and other pragmatic aspects which have to be taken into considerations. The results of the evaluation must therefore be checked for viability among larger user groups.

There are other discourse constraints that limit a discourse from being coherent, such as choice of syntax and lexical preferences. These aspects of natural language have not been taken into consideration. It is possible that with varied syntactic complexities, the principles of building chains of referential expressions would have lead to different evaluation results.

The evaluation results, although performed with a small number of descriptions and evaluators, indicate that there are some general principles governing the distribution of referential forms in different languages and that language-specific coreference strategies lead to better generation results. The descriptions produced with the Swedish language-specific strategies improved the output results. Also, the difference in percentage between the preference of Hebrew descriptions produced with Swedish strategies as opposed to Hebrew descriptions produced with Hebrew strategies was minor. This indicates that Hebrew language-specific strategies improved the output results as well. However, the data used to draw these results is restricted in size and to strengthen the findings presented here the evaluation experiment should be carried out with more text and more readers.

5 SUMMARY AND CONCLUSIONS

This chapter summarizes the findings and the thesis contributions. It also highlights some further studies that could be carried out and the open questions that are yet to be answered.

5.1 Summary

One way to avoid producing natural language that reflects the knowledge representation system is by adapting linguistic knowledge. For this purpose, a quantitative and qualitative analysis was carried out. The results of the analysis contributed to a better understanding about: (1) the typical discourse patterns for the domain and how they can be devised in order to generate multilingual natural language; (2) how to differentiate between the languages regarding the generation of chains of referential expressions.

To test how well these results conform to the language, we propose a multilingual language generation architecture in which discourse patterns and referential expressions strategies are in focus.

When building a generation system that maps from some knowledge representation system that exists in the Semantic Web to natural language it is important to find the solutions that reduce the generation time. To cope with this requirement and avoid computationally expensive generation, we propose a modular system approach. One of the strengths of the presented approach is the separation of the part of the ontological knowledge which is not utilized by the generation system. This is an efficient way to generate from Semantic Web ontologies because the system does not need to process the whole ontology, a process which is highly time consuming.

5.2 MLG using coreference strategies

This is the first research on English, Swedish and Hebrew in the domain of CH that aims at identifying coreference similarities and differences between these languages. The corpus analysis has demonstrated that there are different criteria for how chains of coreference are realized in each language. The analysis showed that the most distinguishing criteria between the languages is manifested in the lexical-semantic relations between an anaphoric expression and an antecedent. In English, chains of coreference are built by combining pronouns and direct-hyperonym noun phrases. In Swedish, direct-hyperonym and synonym relations are more prominent when building coreference chains. A common lexical relation in Hebrew is higher-hyperonym. These differences are closely related to the semantic content.

Our experiments indicate that the notion of cohesion is unique for each language. However, we found that language-specific strategies might work for other languages. Although the corpus analysis showed there are several preferences made by native writers in each language, these do not necessarily hold among other native speakers of that language. This is not surprising considering the different aspects of a discourse that need to be taken into consideration. Coreference can be viewed from different perspectives of the reader's point of view. How different linguistic elements are reduced is an individual choice that depends on many aspects, such as mental accessibly, knowledge of the world, etc. Therefore it is very difficult to draw objective conclusions from the experiment results.

We can conclude that coreference strategies are important for signaling linguistic content as an aspect of a language generation process and that building chains of referential expressions is closely tied to the language in question. Hence, a language generator might benefit from language-specific definitions on how to realize referential expressions in different semantic and syntactic contexts. This kind of knowledge about how coreference is expressed in a discourse may improve on existing multilingual generation methods.

5.3 MLG from structured knowledge representations

This work has shown that for a natural language generator to benefit from Semantic Web ontologies, it must have access to a fine-grained hierarchical representation of the entity it is intended to generate a

description for. There has been some discussion in the NLG community about where to place the process of referential expression generation. This work suggests that the task of referring expression generation should be considered in the late stages of the microplanning phase to allow efficient computation.

In addition, the modular approach presented in this thesis brings new insights into how to approach multilingual generation from Semantic Web ontologies, something which in most cases is an expensive process to compute. This kind of modular approach is domain-independent and can be adapted to new domains.

5.4 Future directions

The syntactic analysis demonstrates that there exist concrete linguistic differences between the languages, which are related to grammatical cohesion and are not covered in this thesis. It will be interesting to test the generation approach on a wider range of linguistic constructions.

We have begun to explore lexicalizing the ontology content using computational lexicons. The same line can be followed to find interesting systematic variations and similarities between languages.

Another interesting future work that could be carried out is to test how descriptive the discourse patterns are by, for example, testing them with other domains. The coreference strategies can be tested with other languages, for example, one can use the Hebrew strategies to generate Arabic text, or use the Swedish strategies to generate texts in Danish and Norwegian. Comprehensive computational evaluations should be designed carefully using both quantitative and qualitative experiments.

Additional experiments could be carried out to test the performance of the modularized approach on larger amounts of data.

One of the benefits the Semantic Web brings with it is the ability to draw inferences about the knowledge it contains. Implicit knowledge might have a particular, quite specific interpretation depending on the context. This presents a difficulty with regard to the choice of the lexical and grammatical content of the generated text. There are still many open questions related to how language generators can benefit from this knowledge. An interesting next step is to extend our approach to incorporate implicit knowledge.

Part I

Generating tailored texts in the context of the Semantic Web

6 A SYSTEM ARCHITECTURE FOR CONVEYING HISTORICAL KNOWLEDGE

Dannélls, Dana 2008a. A system architecture for conveying historical knowledge to museum visitors. *Workshop on Information Access to Cultural Heritage (IACH)*, Lecture Notes in Computer Science. Berlin: Springer.

6.1 Introduction

Given the growing number of visitors looking for on-line information about museum collections and activities, it has become fundamental for various cultural institutions to enhance their visitors' ability to navigate on-line and to access information in the most effective way (Bowen and Filippini-Fantoni 2004). One of the questions posed by cultural organizations (e.g, libraries, museums and galleries) who utilize web-based applications that allow computer users to access their ontology-based museum data is how to accommodate different user needs while maximizing the user interaction and minimizing the production effort (Amato et al. 2008). In the context of the Semantic Web this task is closely related to the characteristics of the application-domain that the system is built upon (Mellish and Pan 2008) and the ability to assess and organize semantic information (Bontcheva and Wilks 2004).

In this paper we present a framework that is being developed to examine how to best express cultural heritage (CH) information encoded in a domain ontology to convey this historical knowledge to museum visitors. We are especially aiming to address the problem of how to establish linguistically motivated document plans about museum objects from ontological information formulated in an RDF language. The framework is innovative in that it provides a reusable solution to facilitate the development and enhancement of a Natural Language Generation (NLG) system given that the task is to research or engineer the production of text or speech from the domain-specific ontol-

ogy CIDOC-CRM.[36] Furthermore, the system reuses a dialogue framework and incorporates dialogue components with elaborated generation techniques.

6.2 The system architecture

To allow system development with reusable modules, flexible application building and easy combination of different techniques, the framework itself must be designed to support user-specific needs relating to the task of producing and realizing conceptual graphs. We argue in the favour of a system architecture using a highly specialized dialogue system framework called the Phase Graph Processor (PGP) (Degerstedt and Johansson 2003). The phase-based system contains several features which seem compatible with our requirements. First, information about the system state is kept in a "shared knowledge base" and can be accessed by system components allowing each component to utilize all the information that the system contains. Second, the system is organized into two layers: a phase-layer and a module-layer. The module-layer can consist of an unlimited number of modules, where each module is considered a free resource that can be used by any phase unit. These features make the system highly modular and allow us to experiment with reusable components using different approaches.

The system architecture is shown in Figure 13. The initial input data to the process are an ontology file and a user profile file. The system takes as input an ontology object, i.e., a concept which is going to be the subject of the retrieved set of statements (kind of conceptual graphs). The current output is a subset of statements describing the input object for the given case. The system is realized in Java and interoperates with other open source and Semantic Web technologies, including Pellet,[37] Jena,[38] and SPARQL. [39] The flow between the processor phases and the system components is described below.

6.2.1 Pragmatic and Memory Phase

The *Pragmatic Phase* maintains the model of the current situation on the basis of the user characteristics and the user knowledge, i.e., the

[36] The Conceptual Reference Model (CRM): http://cidoc.ics.forth.gr/
[37] http://pellet.owldl.com/
[38] http://jena.sourceforge.net/
[39] http://www.w3.org/TR/rdf-sparql-query/

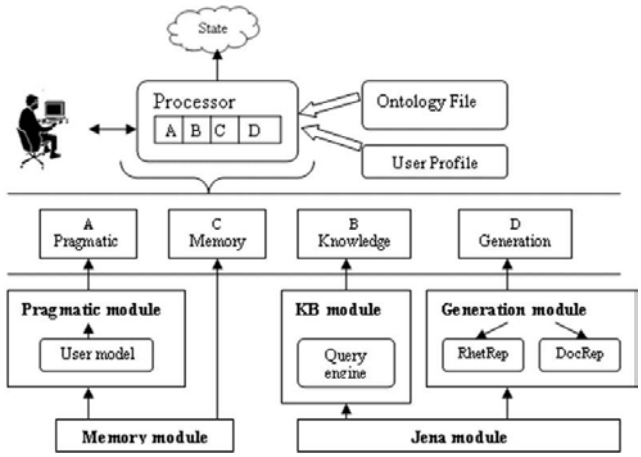


Figure 13: The overall museum guide system architecture.

information stored in memory. Its task is to decide on the appropriate content determination approach and explanation strategy to set the suitable user state accordingly. The task of the *Memory Phase* is to record the current dialogue turn and context tracking modules. It stores a list of facts generated so far in order to avoid generating redundant information. This list is also used to filter out repetitive RDF statements (Bontcheva and Wilks 2004).

The **user model** component stores the information from the user profile. A user profile contains information, such as the user's: (1) age range $a \in$ { 7-16, $\geq$ 17}; (2) expertise $e \in$ { expert, non-expert }; (3) generated facts per sentence $g \in \{1, 3, \geq 4\}$ (this factor determines the amount of facts that should be generated at each dialogue turn); (4) preferred textual complexity $l \in \{simple, complex\}$ (to help determining how knowledge will be selected, organized and realized).

6.2.2 Knowledge Phase

In the world of the Semantic Web, RDF syntax is used to describe a resource in terms of named properties and their values. The RDF's abstract syntax can be represented as a directed graph where each arc in the graph represents a *statement* [40] that the resource at the starting end of the arc, called the *Subject* of the statement has a property called the

[40] We use the term *fact* as a synonym for the term *statement*.

Predicate of the statement with a value called the *Object* of the statement.

In our system, the user input string corresponds to the subject of the statement, and the task of the *Knowledge Phase* is to select the appropriate subset of the available statements whose resource subject is equivalent to this string. The Jena components are utilized to store and parse the input ontology from which knowledge is retrieved. The **query engine** is responsible for formulating queries and retrieve statements from the stored ontology model using the SPARQL query language.

6.2.3 Generation Phase

As advocated by Bouayad-Agha, Power and Scott (2000), document planning can be divided into two stages: rhetorical representation and document representation. Following this approach, the planning process taken during the *Generation Phase* is divided into two steps: (1) **rhetorical representation**, deciding on how to select and organize the data. To allow a flexible interchange between various algorithms we implemented a method which defines a family of algorithms, and encapsulates each one, depending on the required output. We have so far added two algorithms that have been addressed through the framework of content planning (Mellish and Pan 2008): *top-down*, following the ontology structure and *bottom-up*, organizing facts in reverse to the original ontology structure; (2) **document representation**, responsible for distributing the available facts among sentences, paragraphs and perhaps vertical lists in the hope that it will permit a coherent realization as text. In later developments' this process will be limited by grammatical constraints derived from a realization component to cover all aspects of linguistic processing.

6.3 Initial results

We have been running our system on a range of user inputs. Here we report two of the results of an input string about the painting object "P_0922":

```
(1) UM {a18,eN,g3,lS}:

<P_0922, type, E38_1.Paintings>,
<P_0922, has_type, Portrait$>,
<P_0922, has_title, PortraitWithLockView>.


(2) UM {a44,eE,g3,lC}:

<P_0922, has_type, Portrait>,
<P_0922, has_title, PortraitWithLockView>,
<P_0922, was_created_by, LockViewCreation>.
```

In these results the problems connected with the personalized generation from the CRM ontology are made clearly visible: a possible lexical choice for verbalising "has_type" differ in the different contexts; the context provided in example (2) requires sophisticated realization techniques to avoid the generation of poor results.

6.4 Conclusion

In this paper we presented a personal museum guide system architecture that is being developed to exploit the linguistic aspects of the realization of a domain-specific ontology in relation to the user's interaction with this ontology. We argue in the favour of the reuse of a dialogue system where a user model is provided explicitly. The document planning process is applied incrementally and thereby integrates the rhetorical ontology structure and the generated text structure more tightly. These are the novel aspects of this approach. Future work aims to address verbalizations of the CH domain related aspects with respect to different personalized characterizations that answer a range of communicative goals.

7 GENERATING TAILORED TEXTS FOR MUSEUM EXHIBITS

Dannélls, Dana 2008b. Generating tailored texts for museum exhibits. *The 2nd workshop on language technology for cultural heritage (LaTeCH 2008)*, 17–20. Marrakech: ELRA.

7.1 Introduction

During the last decade, the awareness of the need for personalization has become fundamental for cultural institutions such as museums and libraries while aim to produce textual descriptions of museum exhibits tailored to the visitor's knowledge, interests, and personal preferences, such as preferred vocabulary, syntax, sentence length etc. One of the first examples of personalization in a museum context was developed in the Intelligent Labelling Explorer (ILEX) project,[41] by using Natural Language Generation (NLG) techniques. More recently, applications within the cultural heritage (CH) domain have seen an explosion of interest in these techniques (Novello and Callaway 2003; O'Donnell et al. 2001; Androutsopoulos, Oberlander and Karkaletsis 2007).

The process of NLG starts from an ontology that describes a certain domain. Recently, natural language generators that are targeted towards the Semantic Web ontologies have started to emerge. A strong motivation for generating texts from ontologies is that the information represented in an ontology has a true potential to provide a large amount of text if this text is realized correctly. Gradually, the cultural heritage knowledge domain which is often characterized by complex semantic structures and large amounts of information from several different sources will benefit from the complete generation of the information delivered in the ontology.

[41] http://www.hcrc.ed.ac.uk/Site/ILEXINTE.html

Web ontology languages pose many opportunities and challenges for language generators. Although standards for specifying ontologies provide common representations to generate from, existing generation components are not compatible with the requirements posed by these newcoming standards. This issue has been previously addressed by developing domain-dependent authoring interfaces that are built upon an ontology and that allows it to be deployed through knowledge editing (Brun, Dymetman and Lux 2000; Hartley et al. 2001; van Deemter and Theune 2005). These interfaces are links between the ontology and the user who can manipulate the content of the document indirectly in his/her own language. An example of a template-based authoring tool that makes use of this technique within the CH domain was presented by Androutsopoulos, Oberlander and Karkaletsis (2007). An alternative approach to template-based NLG that is particularly relevant in cases where texts are generated from logical forms in several languages simultaneously is a grammar-based approach (Bateman 1997).

In this paper we present a multilingual source authoring tool, which is built upon the grammatical framework (GF) formalism to generate texts from the underlying semantic representation that is based on the Conceptual Reference Model (CRM) domain ontology. The authoring environment is similar to those described by Scott (1999); Dymetman, Lux and Ranta (2000) and van Deemter and Power (2003). [42] The focus is on the process starting from a fixed semantic representation to a surface realization, with emphasis on the syntactical sentence structure, and the content variation.

The structure of the paper is as follows. In section 7.2 we elaborate the notion of ontology and describe both the reference ontology model and the grammar formalism that our application is built upon. Section 7.3 presents the grammar implementation and explains how it is utilized to generate tailored descriptions from a formal representation language. We finish with conclusions and a discussion of future work in section 7.4.

7.2 Background

In the context of the work presented here, an *ontology* is understood as a formal model that allows reasoning about concepts, objects and about the complex relation between them. An ontology holds meta-level in-

[42] The advantages of utilizing this family of domain authoring approaches that are coupled with multilingual text generation are elaborated by Scott (1999).

formation about different types of entities in a certain domain and provides a structure for representing contexts, it is not human readable as it is designed to be processed by computer systems.

Examples of Web ontology-languages that have been developed by the W3C Web-Ontology working group are OWL and DAML+OIL.[43] The basis for the design of these Web technology languages based on the RDF Schema is the expressive Description Logic (DL) *SHIQ* (Horrocks, Patel-Schneider and van Harmelen. 2003). These languages provide extensive reasoning capabilities about concepts, objects and relationships between them.

7.2.1 Generating from an ontology

In an ontology, an object may be described by semantic graphs whose nodes (concepts) represent parts of an object, and the arcs (relations) represent partial constrains between object parts. Each relation described in a logical language is binary, i.e. it connects between two nodes. In order to present a piece of information about an object represented in an ontology, multiple sentences must be formulated. It becomes valuable if these sentences that build the final text can be adapted to various contexts or users.

There has been some successful attempts to generate from ontologies (Wilcock 2003; Wilcock and Jokinen 2003; Bontcheva and Wilks 2004; Bontcheva 2005). Wilcock (2003) and Wilcock and Jokinen (2003) have shown how RDF/XML generation approach can be extended so that the information embedded in the ontology can be exploited to generate texts from Web ontology-languages such as DAML+OIL and OWL without the need for a lexicon. Bontcheva (2005) demonstrated how to minimize the effort when generating from Web ontologies while being more flexible than ontology verbalisers. Some of the difficulties reported by these authors concern lexicalization and in establishing context variations.

7.2.2 The CIDOC-CRM ontology

One initiative to enable an ontology in the context of the cultural heritage is the Conceptual Reference Model domain ontology. The International Committee for Documentation of the International Council of

[43] http://www.w3.org/TR/

Museums Conceptual Reference Model (CIDOC-CRM)[44] is a core ontology and ISO standard for the semantic integration of cultural information with library archive and other information (Doerr 2005). The primary role of the CRM is to enable information exchange and integration between heterogeneous sources of cultural heritage information.

The central idea of the CIDOC-CRM is that the notion of historical context can be abstracted as things and people. It concentrates on the definition of relationships rather than classes to capture the underlying semantics of multiple data and meta structures. It tends to provide an optimal analysis of the intellectual structure of cultural documentation in logical terms, which is available in several formats such as RDF and OWL that have hardly been explored yet. The work described in this paper is based on the OWL version of the ontology.[45]

7.2.3 The Grammatical Framework (GF)

The Grammatical Framework (Ranta 2004) is a functional grammar formalism based on Martin-Löf's type-theory (Martin-Löf 1975) implemented in Haskell.[46] GF focuses on language independent semantic representations. It differentiates between domain dependent and domain independent linguistic resources, as it is designed to be applicable both to natural and to formal languages. One abstract grammar can have several corresponding concrete grammars; a concrete grammar specifies how the abstract grammar rules should be linearized in a compositional manner.

Multilingual functional grammatical descriptions permit the grammar to be specified at a variety of levels of abstraction, which is especially relevant for constructing a detailed mapping from semantics to form. This aspect is crucial for natural language generation to work. What makes the grammar suitable for generating from ontologies and in particular from OWL, is that it allows multiple inheritance.

GF has three main module types: abstract, concrete, and resource. Abstract and concrete modules are top-level, in the sense that they appear in grammars that are used at runtime for parsing and generation. They can be organized into inheritance hierarchies in the same way as object-oriented programs. The main advantage with converting the on-

[44] http://cidoc.ics.forth.gr/

[45] http://cidoc.ics.forth.gr/OWL/cidoc_v4.2.owl

[46] Haskell is a standardized purely functional programming language with non-strict semantics. Similar to Lisp and Scheme.

tology to GF is that we can make use of the rich type system in the concrete syntax for capturing morphological variations.

Our approach is based on the idea suggested by Khegai, Nordström and Ranta (2003) who utilized GF to automatically generate multiple texts from semantic representations. The source authoring environment deploys similar techniques to those introduced by Scott (1999); Dymetman, Lux and Ranta (2000) and van Deemter and Power (2003).

7.3 Generating from the ontology

We chose for study a small amount of logical relations represented in the ontology and wrote a grammar that is capable to describe them in natural language through user editing. The following code is a fragment taken from the ontology we employed. The code states that the class *PaintingP9091* must have at least one value *TypeValue* on property *has_type*; the individual *TypeValue* is an instance of the class *cidoc:E55. Type*[47] and has two property values: "tool" and "painting".

```
<owl:Class  rdf:about="PaintingP9091">
 <owl:Restriction>
  <owl:onProperty rdf:resource="&cidoc;P2F.has_type"/>
   <owl:hasValue rdf:resource="#TypeValue"/>
 </owl:Restriction>
</owl:Class>
<owl:Thing rdf:about="#TypeValue">
  < rdf:type rdf:resource="&cidoc;E55.Type"/>
  <Tool  rdf:datatype="&xsd;string">tool
  </Tool>
  <Painting  rdf:datatype="&xsd;string">painting
  </Painting>
</owl:Thing>
```

The above fragment exemplifies the representation of the classes and relationships that are utilized by the grammar. In the grammar implementation, classes are represented as categories; properties are functions (rules) between two categories, where each property links between two classes; individuals are lexical categories (strings). Below is a representation of the *mkObject*, which corresponds to a function that links between the classes of an *Object*:
mkObject:ObjectNodeI→ObjectNodeII→ObjectNodeIII→Object;

[47] The notation &cidoc; is used instead of the whole namespace,i.e http://cidoc.ics.forth.gr/OWL/cidoc_v4.2.owl#

In the above example the *Object* category corresponds to *PaintingP9091*. Each *ObjectNode* is a class, according to the above ontology representation, *ObjectNodeI* corresponds to the cidoc class *cidoc:E55.Type*. It is followed by *ObjectNodeII*, i.e. *cidoc:E52.Time-Span* and *ObjectNodeIII*, i.e. *cidoc:E21. Person*, as shown below.

{Type} *instance_of* ObjectNodeI
{Time-Span} *instance_of* ObjectNodeII
{Person} *instance_of* ObjectNodeIII

Consequently, individuals such as "tool" and "painting" are terminals and are declared in the concrete syntax. In the next sections we describe the abstract and the concrete representations

7.3.1 The abstract representation

The abstract syntax is a context-free grammar where each rule has a unique name. An abstract rule in GF is written as a typed function. The categories and functions are specified in GF by *cat* and *fun* declarations. Below is a fragment of the grammar:

cat

Object ;ObjectNodeI ; Type ;
ObjectNodeII ; Time-Span ;
ObjectNodeIII ; Person ;

fun

HasType_This : Type → ObjectNodeI;
HasType_Here : Type → ObjectNodeI;
HasType_Template : Type → ObjectNodeI;
HasTimeSpan: Time-Span → ObjectNodeII;
CarriedOutBy_Painting: Person → ObjectNodeIII;
CarriedOutBy_Tool: Person → ObjectNodeIII;

The abstract syntax gives a structural description of a part of the domain. It has several advantages, one of which is the ability to utilize the same categories differently depending on the semantic complexity of the context. Here we declared three functions for the *ObjectNodeI* to achieve context variations, though very simple ones. Similarly, we

declared two functions for the *ObjectNodeIII*, however, the difference between *CarriedOutBy_Painting* and *CarriedOutBy_Tool* is the choice of the verb in the linearization rule. The verb *painted by* is applied when the subject is the noun *painting*, but the verb *created by* is applied when the subject is the noun *tool*, in cases when the object is an instance that belongs to the category *Person*.

7.3.2 The concrete representation

Each category and function introduced in the abstract syntax has a corresponding linearization type in the concrete syntax. Linearization rules are declared differently for each target language. In addition, each concrete syntax also contains grammatical parameters and grammar rules, which are used to ensure grammatical correctness for each language, in our case English and Swedish. An example of linearization rules taken from the English concrete syntax is the following:

lin

```
CarriedOutBy_Painting obj =
   {s = det ! obj.num ++ cop ! obj.num ++
         "painted by" ++ obj.s ; num=obj.num};

 Painting = {s = "painting" ; num = sg} ;
 Painting = {s = "paintings" ; num = pl} ;
```

Grammatical features are supported by GF and the agreement between the pronoun and the verb is enforced in the generated sentences. The variable *obj* represents a terminal string. The parameter *num* is an abbreviation for the parameter type "number", it contains the inherent number that can be either singular (sg) or plural (pl). The operation *det* is a determiner, and the operation *cop* is copula verb.

7.3.3 The authoring environment

Figure 3 illustrates the source authoring environment. The left-side window shows the abstract syntax tree, which represents the *Object* structure. The large window positioned to the right is the linearization area, the editing focus is presented as the highlighted metavariable ?3. The bottom area shows the context-dependent refinement for the *ObjecNodeIII*, there are two possible relations to choose from.

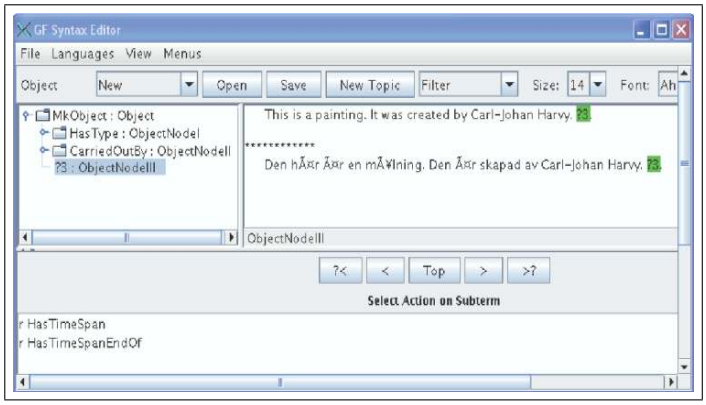


Figure 14: The GF source authoring environment.

The authoring tool that is built upon the GF grammar makes it possible to generate the following texts:

English
(1) Here we have a painting. It was painted by Carl-Johan Harvy. It was made in 1880.
(2) This is a tool. It was made in 1880. It was created by Carl-Johan Harvy.
(3) On the second floor of the history museum we have paintings. They were created by Siri Derkert. They were produced in Italy.

Swedish
(1) Här har vi en målning. Den är målad av Carl-Johan Harvy. Den är gjord på 1880 talet.
(2) Det här är ett redskap. Det är gjort på 1880 talet. Det är tillverkat av Carl-Johan Harvy.
(3) På andra våningen i historiska museet har vi målningar. De är tillverkade av Siri Derkert. De är producerade i Italien.

The difference between the first and second sentence is the order in which the *ObjectNodeII* and the *ObjectNodeIII* appears, this is done with the help of the *variants* function that allows for syntactic variations by reordering the linearized categories. The third sentence illustrates a typical example of a combined template and grammar based generation, e.g. the fixed sentence: "On the second floor of the history museum" that has been prewritten.

7.4 Conclusions and future work

In this paper we have presented a multilingual grammar-based approach, the aim of which is to generate exhibit descriptions following the CIDOC-CRM domain ontology. We chose for study a small amount of logical relations represented in the ontology and have started to examine the capabilities of utilizing a grammar to bridge between ontology representations and different users.

We suggest an approach to generate descriptions in English and Swedish. The suggested approach supports user preferences on receiving cultural heritage information from the Semantic Web. We show how the GF authoring tool, which allows users to choose the content and the form of the output text, can be utilized for this purpose.

Future work will focus on ontology studies and on particular problems of generating for cultural heritage. We are also planning to utilize the Resource Grammar Library that has been developed to provide the linguistic details for application grammars on different domains. This will be a step towards high quality summary generation. Our goal is to built a grammar that reflects the ontology structure and supports all the OWL features to allow the user to interact with the complete ontology.

Part II

Generating cultural content through discourse strategies

8 THE VALUE OF WEIGHTS IN GENERATED TEXT STRUCTURES

Dannélls, Dana 2009. The value of weights in automatically generated text structures. *Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Lecture Notes in Computer Science, LNCS 5449, 233–244. Berlin: Springer.

8.1 Introduction

The ability to generate natural language text from web ontology languages and more generally knowledge bases that are encoded in RDF (Resource Description Framework) imposes new demands on natural language generators that aim to produce written text either for textual presentation or for eventual use by text-to-speech system. One of these demands concerns the process of text planning. Text planning, also referred to *Document Planning* (Reiter and Dale 2000), is the process responsible for producing a specification of the text's content and structure. The fact that aspects such as the user characteristics, e.g., cognitive state, desires, the background domain knowledge, and linguistic properties must be taken into account and computed simultaneously during planning makes this process computationally hard and so far there has been little success in computing a general model with a suitable structure for generating from ontologies in general and from web ontologies in particular. This brings a need to find alternative strategies to generate knowledge from ontology languages, or alternatively to adapt previously presented ideas to the new emerging technology standards.

Recent attempts to develop natural language generators that support the Web Ontology Language (OWL) and similar Semantic Web languages,[48] treat the class hierarchy as kind of directed graph that is utilised to produce a coherent text (Bontcheva and Wilks 2004; Bontcheva

[48] http://www.w3.org/TR/

2005) with the most common algorithms including top-down approaches. To enhance personalisation and improve the clarity of the text content describing an object in a hierarchy, these approaches have been combined with comparison methods whose goal is to facilitate learning by relating new concepts to a user's existing knowledge (Milosavljevic 1997; Isard 2007). Yet, one of the main questions that arises in this context is how to capture and expose the relevant ontology content to the reader.

In this paper we present a text planning technique that has been developed to explore the value of assigning weights to ontology properties in addition to comparison methods. The generation technique is optimised to tailor descriptions about a concept from the rich logical structure of Web ontologies and was implemented as a part of a question-answering system. It combines top-down and bottom-up algorithms with enhanced comparison methods to produce a personalised text structure. To test the method performance we run the system on a range of user queries with different user preferences. The generation results indicate that the process of computing preferable property weights in addition to known generation techniques has a positive effect on the text structure and its content. An experiment was conducted to evaluate the generation results using human subjects. The evaluation results show the benefits of manipulating the ontology knowledge on the basis of pre-assigned property weights.

The remainder of this paper is structured as follows. In section 8.2 we describe the prior approaches in more detail. In section 8.3 we present the methodology of the generation machinery and the motivation behind the implementation. In section 8.4 we describe the implementation and the text planning approach. In section 8.5 we report on the experimental setup and present the evaluation results. In section 8.6 we discuss their implications and we conclude with section 8.7.

8.2 Background

8.2.1 Semantic web ontologies

An *Ontology* is defined as a representation of a shared conceptualisation of a specific domain and plays a vital role in the Semantic Web (Berners-Lee 1998) as it provides a shared and common understanding of a domain that can be communicated between people and heterogeneous, distributed application systems.

Web ontology languages are built upon the RDF and RDF Schema.[49] [50] The basic RDF data model contains the concepts of resource in terms of named properties and their values. It is an object-property-value mechanism, which can be seen as forming a graph where each edge represents a *statement* that the resource at the starting end of the edge, called the *Subject* of the statement has a property called the *Predicate* of the statement with a value called the *Object* of the statement. This is shown in Figure 15. Every elliptical node with a label corresponds to a resource and every edge in the graph represents the property of the resource. Formally:

Definition 1 An ontology O=(G,R) where G is a labeled graph and R is a set of rules. The graph G=(V,E) comprises a finite set of nodes V, and a finite set of edges E. An edge *e* belonging to the set of edges E is written as (n_1,α, n_2) where n_1 (the subject) and n_2 (the object) are labels of two nodes belonging to a set of nodes V and α is the label (the predicate) of the edge between them.

8.2.2 Planning the text structure from Web ontologies

The fact that the RDF's abstract syntax can be represented as a directed graph which corresponds to the structure of a coherent text was exploited by various authors who utilise top-down approaches to generate natural languages (O'Donnell et al. 2001; Wilcock and Jokinen 2003; Bontcheva and Wilks 2004). As pointed out by these authors, selection methods which follow the ontology graph structure pose several difficulties on the task of planning the text content. One of those is the fact that web ontologies are described as resources and are identified with URIs. This means that they can act as fields not just in the local store but anywhere they appear; when generating natural languages from ontologies it is not always clear where to begin to acquire knowledge about the concept that will be described. Recently, a new approach to content planning has been suggested by Mellish and Pan (2008) who impose a bottom-up method to identify appropriate text contents. They follow an approach that is associated with conversational maxims to select and plan consistent and informative contents (Mellish and Pan 2008; Young 1999). Our approach is most closely in line with Mellish and Pan (2008); Young (1999), however our goals are different. Mellish and Pan (2008) aim to find optimal axioms that are language motivated

[49] http://www.w3.org/RDF/

[50] http://www.w3.org/TR/rdf-schema/

by inducing new inferences, we aim to improve the text content and structure by combining different generation approaches with preferred property weights. Here we describe an attempt to enhance the input of an NLG system with some domain specific preferences in a way that is adaptive to the task at hand and test whether the generation results actually improve.

8.2.3 Tailoring the content and form of the text

Bontcheva (2005) extends the approach presented by Bontcheva and Wilks (2004) towards portability and personalisation. She presents an approach for producing tailored summaries by accounting for the user preferences that are imposed during the last generation phase, mostly to adapt the length of the generated text. No weights are computed to distinguish what should be included in the text content, and thus there is no adaptation in terms of the contextual information. In M-PIRO (Androutsopoulos et al. 2001), it is the user himself who chooses the information that should be included in the generated text and specifies his/her preferred language. This is accomplished through an authoring tool that makes the properties of the object visible to the user. The specified preferences are stored in a user model that is consulted during generation. Similarly to ILEX (O'Donnell et al. 2001) their user model contains scores indicating the educational value of the chosen information as well as how likely it is for him/her to find a particular type of information interesting. Our approach adds an addition feature to those as it allows to define a set of properties with higher weights which can be interleaved with the user model and computed during the comparison process.

8.3 Methodology

8.3.1 Conveying semantic information

To make certain predictions that will help us to convey an ontology content and will allow the system to generate certain continuities in the text structure, there are several questions that are asked, these are: what statements must occur; where can they occur; how often must they occur. Answers to these questions which guide our generation approach depend on the statement's property weight, the ontology content, the user preferences, the context, etc. Let us introduce the following text.

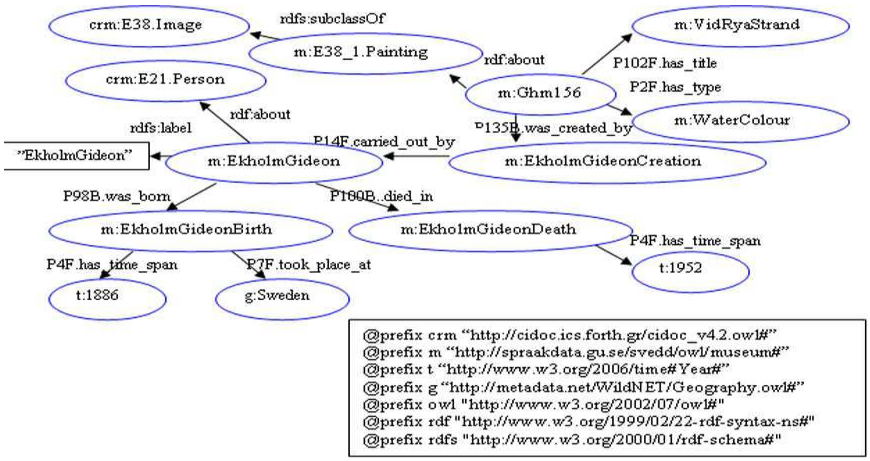


Figure 15: Classes and properties represented in RDF syntax.

Text T1
U: What is Ghm156?
S: Ghm156 is titled "Vid Rya Strand". Ghm156 was painted by Ekholm Gideon.
U: Who is Ekholm Gideon?
S: Ekholm Gideon is a painter. Ekholm Gideon was born in Sweden.

Text T1 is an example of a successful interaction sequence with the user, the user model in this context was: UM={a18,eN,g2,lS}, following the UM attributes described in section 8.3.2. The four statements that were generated by the system have received the highest property weights, given the ontology content. A fragment of the ontology from which the ontology statements were generated is shown is Figure 15, in this ontology four domain ontologies are emerged. We consider this interaction to be a successful one since in the text sequence produced by the system the generated ontology statements that are relevant to the topic of the conversation are presented. Thereby following the Grice's conversational maxims of quantity, i.e., the contribution to the conversation is informative. There is no abundance of information and the generated statements allow the user to ask back on one of the new concepts given in the generated description, e.g., the title, the painter place of birth, etc., from which the system can generate new descriptions relevant to natural language presentation.

To find an adequate sequence of statements about a concept described in the ontology and be able to present the related statements that are relevant in the context, are relevant to the user and eases the user understanding about it, we implemented a stepwise text planning (described in section 8.4). The planning procedure combines top-down and bottom-up algorithms with comparison techniques to generate relevant content about a concept described in an ontology. In addition it is possible to specify a set of properties with higher weights that can be computed during the comparison process.

8.3.2 Tailoring the ontology content

We aim to establish the rhetorical text content that supports reader and listener preferences. This is accomplished with the help of two modules: (1) the *User Module (UM)*, holds metadata information about the user's: age $a \in$ { 7-16, ≥ 17}; expertise $e \in$ { expert, non-expert }; generated facts per sentence $g \in \{1, 3, \geq 4\}$ preferred textual complexity $l \in \{simple, complex\}$. (2) the *Memory Module (MM)*, represents the user knowledge, filters out repetitive RDF statements and ranks the selected statements. As the discourse evolves the memory increases; depending on the user module, statements in the memory might receive higher selection priority (section 8.4.2.1). This information characterise the user specific part of the input to a single invocation of the generation system.

Similarly to Wilcock (2003) and Bontcheva (2005), we utilise the names of the ontology concepts and properties to generate the lexicon and produce the text content. Our point of departure is the English language in which the ontology information is given. However, we intend to map each concept and property to its appropriate lexical entry in languages other than English and implement a grammar that makes use of those entries to generate natural language contents.

8.4 Implementation

8.4.1 The generation machinery

Our approach was implemented within a question-answering system where users can access a knowledge base of information in natural language (Johansson, Degerstedt and Jönsson 2002). The system architecture is introduced by Dannélls (2008b). The initial input data to the process are an ontology file and a user profile file. The user profile holds

the user preferences that are stored in the UM. The ontology knowledge is held in a Jena store from which information is retrieved.[51] The system generates a description about the concept described in the ontology that was chosen by the user. The output is a set of content elements describing the input concept for the given case. It is a subset of verbalised statements describing the input concept.

8.4.2 Stepwise text planning

The text planning module is decomposed into two phases (Bouayad-Agha, Power and Scott 2000), it is a flexible approach that allows to exploit text possibilities (Mellish, Oberlander and Knott 1998):[52] (1) *rhetorical representation*, deciding on how to select and organise the data (see below); (2) *document representation* (also called surface realisation) distributing the available data among sentences, paragraphs and perhaps vertical lists in the hope that it will permit a coherent realization as text. Here we take a simple approach to complete the generation process, i.e., concepts are assumed to be lexicalised as nouns and properties as verbs.[53]

The rhetorical representation module acquisition problem is decomposed in two main steps: Content selection and Content organisation.

8.4.2.1 Content selection

Content selection operates over a relevant data that has been identified within the generator, see (1a), Table 8.1. Given the user query, the user model, the memory model, the ontology knowledge-base and a set of scored properties (edges) the task is to select the informative statements that meet the user request and that eases the user understanding about it.

First, all the edges in which the concept *n* appears in are selected. Second, every concept, i.e.,*n_new* other than the input one that has a path from *n* in G is selected. The selected edges are added to a sub-

[51] http://jena.sourceforge.net/

[52] This process of text planning is equivalent to the two processing modules: *Content Determination* and *Content Planning* that were proposed by Reiter (1994).

[53] Although there appears to be similarities between lexical entries and concepts, in linguistics and philosophy the term *concept* is defined as a nonlinguistic psychological representation of a class of entities in the ontology, where verbs distinguish what properties it has.

Table 8.1: Content selection algorithm, following the formal ontology Definition 1, section 8.2.1.

(1a) Statement selection:
function SELECT(n, G)
Input a node n, and an ontology graph G
For $n \in$ V
Add (V_n,E_n) to G′
For *n_new* $\in$ V
Add (V_{n_new},E_{n_new}) to G′
return G'

(1b) Score selected statement:
procedure SCORE(E,p)
Input a set of edges E, and a set of properties p
For $e \in$ E
Score(e)= W_α + $Hier_n$ + $Hist_n$

graph G', the prim sign ′ indicating a subset.

Scoring equation Scores are computed for every selected edge according to the equation presented in (1b) Table 8.1 that was partially inspired by Isard (2007).
W_α: the edge property weight;
$Hier_n$: hierarchical distance between the selected concept and the compared resource (i.e., the subject node of the edge in focus);
$Hist_n$: historical distance, i.e., the amount of generated edges after the edge in focus was presented to the user, 0 if it was never presented.

8.4.2.2 Content organisation

In this phase we assume there is no useful organisation to the taxonomic information of the selected subgraphs, or alternatively that such organisation as there is, follows the ontology structure. Given a set of scored edges that cover the input query, the task is to look for the relevant ones and organise them accordingly to generate the final output. This step is carried out by a stochastic search method (Mellish, Knott and Oberlander 1998; Mellish, Oberlander and Knott 1998).

The stochastic search method is a form of heuristic search that executes the following generic algorithm:

1. Randomly pick one or more edges from the set, in such a way as to prefer items with the highest scores.
2. Use these to generate one or more new random variations.
3. Add these to the set, possibly removing less preferred edges in order to adapt the size to the user requirements.

8.5 Evaluation

8.5.1 The domain ontology

Our domain ontology follows the CIDOC Conceptual Reference Model (CRM) thesaurus standard.[54] It is a conceptual model that subscribes an object-centred view of the CH domain. The model comprises 81 relations and 244 concepts and covers the semantic field of hundreds of schemata (Doerr, Ore and Stead 2007).

The domain ontology was created from the Carlotta database,[55] which is designed to be equally applicable to the CIDOC-CRM and covers objects from cultural history, photos, literature, archaeology, theatre, etc. It was enhanced with about 150 new concepts and properties, each of which was assigned with a *rdfs:label* that links its lexical string-name. Figure 15 illustrates a fragment of the data represented as RDF graph (in this graph only one label is made visible).

8.5.2 Adjusting the domain properties

The experiment presented here was carried out within the cultural heritage domain and the heuristics for selecting and ordering properties are based on the results from pilot studies that were conducted in order to gauge user needs and preferences with regards to browsing and exploring material related to artists and artwork (Capra et al. 2007; Clough, Marlow and Ireson 2008).

As a result of these studies four main information views that correspond to user needs were identified. We learned that the four views

[54] http://cidoc.ics.forth.gr/

[55] http://carlotta.gotlib.goteborg.se/pls/carlotta/welcome

Table 8.2: Property list for scoring edges (w1-less valuable, w2-valuable, w3-most valuable).

View	Property name	Weight
Artist name	carried_out_by	w3
Artwork titles	has_title	w3
Artist's birth date	was_born	w2
Artist's death date	died_in	w1
Artist birth place	took_place_at	w2

that were found most helpful and interesting for the variety of users correspond to five properties in our domain ontology. The identified views, their corresponding properties and their weights are specified in Table 8.2.

8.5.3 Experiment and result

We run the system on a range of user queries with different user preferences and generated two outputs for each query, one that applies the method described here (Our method), and another that does compute user preferences but does not utilise the scores and the ordering component, i.e., the results follow the ontology structure (Simple method). Below follows one example of the generation results.

Our method

Text T2
(1) U: What is Ghm156?
(2) S: Ghm156 is a painting. Ghm156 has title "Vid Rya Strand". Ghm156 was carried out by Ekholm Gideon.
(3) U: Who is Ekholm Gideon?
(4) S: Ekholm Gideon is a painter. Ekholm Gideon's birth took place in Sweden. Ekholm Gideon birth date has time span 1886.

Simple method

Text T3
(1) U: What is Ghm156?
(2) S: Ghm156 is a painting. Ghm156 has type water colour. Ghm156

Table 8.3: Evaluation results. Quality (Q), Coherency (C), Interest (I)

	Our method	Simple method
Q	3.6	1.8
C	2.3	2.8
I	3.8	1.9

has title "Vid Rya Strand".
(3) U: What is Ghm156?
(4) S: Ghm156 has time span 1924. Ghm156 was created by Ekholm Gideon's creation. Ghm156 was carried out by Ekholm Gideon.

In text T2 that was produced using our approach, three of the most important properties (according to our property set) are presented already after the first enumeration question, which enables the user to precede with the next question about the new concept, e.g., "Ekholm Gideon". When we employed the simple method approach, text T3, the user needs to repeat on the query about the same concept, e.g., "Ghm156" since the information provided after the first enumeration question does not contribute with informative knowledge. In this case the generated statements are not consist and violet Grice's maxim.

Fourteen interaction sequences, similar to the above examples, were generated and presented to non-experts human subjects, in total eleven subjects participated in the evaluation. Each participant was asked to evaluate the usefulness of each interaction sequence in terms of: (a) Quality (Q), whether the content of the generated statements were relevant and helpful in describing the required object; (b) Coherency (C), whether the generated text structures were coherent and made sense; (c) Interest (I), whether the presented statements (facts) invoked the user interest. For this evaluation a five-point scale (0-poor, 5-excellent) was used. We calculated the mean value of results, these are summarised in Table 8.3.

A closer look at the generated text structures that were presented in different points of the interaction sequences showed there were cases where the generated content contained a mixture of statements describing different concepts, yet that are all related to the required concept. This may explain why the simple method is superior in "coherency". On the other hand in "quality" and "interest", our method outperforms over the simple approach, which is encouraging.

8.6 Discussion

Though the idea of exploiting properties of the ontology concepts for generation tasks is not new, the approach here is new in regards to testing in practice how the choice of the property weights effects the text structure and its content with the aim to promote insights into generating knowledge from web ontologies. The fact that content determination is not bounded to the ontology structure makes it possible to gradually present information that accommodates to different contextual degrees and user needs.

The choice of employing a generation approach such as the one presented here that is compatible with employing a domain-specific ontology is based on the relative ease by which such knowledge might provide solutions for building domain-independent generators. Currently it is assumed that a task-specific approach such as the one presented here is tied to the domain ontology and operates at the object level, however, when merged with other ontologies it may operate on meta-level (Wilcock and Jokinen 2003).

The approach presented here was only tested on a small ontology with, where only a few subjects participated in the evaluation, a question that comes to mind is how well does it scale (Hardcastle and Scott 2008). From our observation we anticipate that operating on larger ontologies may give raise to several modifications, for example the selection strategy may result in a large content when retrieving all knowledge about an object, this might be limited by putting an exceeds threshold on the depth and length of the required graph.

The growing body of research that generates from non-linguistic structured databases has employed different comparison methods to enhance comprehension and improve the clarity of texts for the end-user (Dale and Reiter 1995; Milosavljevic 1997; O'Donnell et al. 2001). Comparison methods can reveal the patterns of contrast and similarity (Isard 2007) and have proven to be useful to remove redundant information, a problem that is exhibited in RDF's (Bontcheva and Wilks 2004). The specific selection strategy adopted here accommodates to these approaches and has proven these methods feasibility for generating from a web ontology.

8.7 Conclusion and future work

In this paper we presented a generation approach to text planning that has been developed to explore the value of assigning weights to domain specific properties. The generation method combines bottom-up and top-down approaches with enhanced comparison techniques to accommodate for the complex structure of Web ontologies. It was implemented within a question-answering framework where the primary goal was to tailor descriptions about a concept described in an ontology to a variety of users. The generated results show the benefits of assigning preferred property weights to enhance the quality and relevance of the generated content elements. A preliminary evaluation indicates that when several factors are enforced during planning, users' interest about the content describing an ontological concept seems to increase.

Although this study focused on a domain specific ontology, and conclusions were drawn based on a small amount of generation results, the findings and technical principles behind the presented methodology could likely to be generalised to other domains. Furthermore, an evaluation can potentially be repeated to confirm the generation results and to test how well does the method scales. Future work aims to assign grammar rules and lexical entries in order to produce coherent texts from the generated text structure elements. In this paper we emphasised mainly the text structure and rhetorical content, but it is necessary to cover linguistic aspects to motivate the chosen text structures for producing grammatically correct texts.

9 Discourse generation from formal specifications using GF

Dannélls, Dana 2010a. Discourse Generation from Formal Specifications Using the Grammatical Framework, GF. *Special issue of the journal Research in Computing Science* 46: 167–178.

9.1 Introduction

During the past few years there has been a tremendous increase in promoting metadata standards to help different organizations and groups such as libraries, museums, biologists, and scientists to store and make their material available to a wide audience through the use of the metadata model RDF (Resource Description Framework) or the Web Ontology Language (OWL) (Schreiber et al. 2006; Bryne 2008). Web ontology standards offer users direct access to ontology objects; they also provide a good ground for information extraction, retrieval and language generation that can be exploited for producing textual descriptions tailored to museum visitors. These advantages have brought with them new challenges to the Natural Language Generation (NLG) community that is concerned with the process of mapping from some underlying representation of information to a presentation of that information in linguistic form, whether textual or spoken. Because the logical structure of ontologies becomes richer, it becomes increasingly hard to devise appropriate textual presentation in several languages that humans comprehend (Hielkema, Mellish and Edwards 2008).

In this article we argue that discourse structures are necessary to generate natural language from semantically structured data. This argument is based on our investigations of text cohesive and syntactic phenomena across English, Swedish and Hebrew in comparable texts. The use of a discourse strategy implies that a text is generated by selecting and ordering information out of the underlying domain ontology, a

process which provides a resulting text with fluency and cohesion. It is an approach that relies on the principles drawn from both linguistic and computer science to enable automatic translation of ontology specifications to natural language. We demonstrate how discourse structures are mapped to GF's abstract grammar specifications from which multilingual descriptions of work of art objects are generated automatically. GF is a grammar formalism with several advantages which makes it suitable for this task – we motivate the benefits GF offers for multilingual language generation. In this work, we focus on the cultural heritage domain, employing the ontology codified in the CIDOC Conceptual Reference Model (CRM).

The organization of this paper is as follows. We present some of the principles of cohesive text structure (Section 9.2) and outline the difficulties of following these principles when generating from a domain ontology (Section 9.3). We show how discourse strategies can bridge the gap between formal specifications and natural language and suggest a discourse schema that is characteristic to the cultural heritage domain (Section 9.4). We demonstrate our grammar approach to generating multilingual object descriptions automatically (Section 9.5). We conclude with a summary and provide pointers to future work (Section 9.6).

9.2 Global and local text structure

Early work on text and context (Hasan 1985) has shown that cultural content is reflected in language in terms of text as linguistic category of genre, or text type. A text type is defined as the concept of Generic Structure Potential (GSP) (Halliday and Hasan 1989). According to this definition, any text, either written or spoken, comprises a series of optional and obligatory macro (global) structural elements sequenced in a specific order and that the obligatory elements define the type to which a text belongs. The text type that is expressed here is written for the purpose of describing work of art objects in a museum.

To find the generic structure potential of written object descriptions, we examined a variety of object descriptions, written by four different authors, in varying styles. Our empirical evidence suggest there is a typical generic structure potential for work of art descriptions that has the following semantic groupings:

1. object's title, date of execution, creation place
2. name of the artist (creator), year of birth/death
3. inventory number when entered to the museum, collection name
4. medium, support and dimensions (height, width)
5. subject origin, dating, function, history, condition.

To produce a coherent text structure of an object description the author must follow this semantic specification sequences that convey the macro structure of the text. Apart from the macro structural elements, there is a micro (local) integration among semantic units of the text type that gives the text a unity. These types are reflected in terms of reference types that may serve in making a text cohesive at the paragraph or embedded discourse level. Some examples of reference types are: conjunction, logical relationships between parts of an argument, consistency of grammatical subject, lexical repetition, consistency of temporal and spatial indicators. Thus local structure is expressed partly through the grammar and partially through the vocabulary.

9.3 The realities of a domain specific ontology

The ontology we utilize is the Erlangen CRM. It is an OWL-DL (Description Logic) implementation of The International Committee for Documentation Conceptual Reference Model (CIDOC-CRM) (Crofts et al. 2009).[56] The CIDOC-CRM is an event-centric core domain ontology that is intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information and museum documentation.[57] One of the basic principles in the development of the CIDOC CRM has been to have empirical confirmation for the concepts in the model. That is, for each concept there must be evidence from actual data structures widely used. Even though the model was initially based on data structures in museum applications, most of the classes and relationships are surprisingly generic. In the following we use this model to illustrate the limitation imposed by a domain specific ontology on

[56]The motivation behind the choice of DL is that it allows tractable reasoning and inference; it ensures decidability, i.e. a question about a concept in the ontology can always be answered; it supports the intuition that the model must be clear, unambiguous and machine-processable. These aspects are in particular important in computational setting, where we would like our logic to be processed automatically.

[57] The model was accepted by ISO in 2006 as ISO21127.

```
1.:RestOntheHunt_PE34604
2.        a     :CIDOC5.0.1.rdfsE22.Man-Made_Object;
3.  :CIDOC5.0.1.rdfsP30B.custody_transferred_through       :RestOntheHunt_PE34604;
4.  :CIDOC5.0.1.rdfsP51F.has_former_or_current_owner      :Museum_Hallwylska;
5.  :CIDOC5.0.1.rdfsP51F.has_former_or_current_owner      :Galleria_Spada;
6.  :CIDOC5.0.1.rdfsP43F.has_dimension      :width_1.2_m;
7.  :CIDOC5.0.1.rdfsP43F.has_dimension      :length_1.54_m;
8.  :CIDOC5.0.1.rdfsP50F.has_current_keeper      :Museum_Hallwylska;
9.  :CIDOC5.0.1.rdfsP52F.has_current_owner      :Museum_Hallwylska;
10. :CIDOC5.0.1.rdfsP48F.has_preferred_identifier      :PE_34604;
11. :CIDOC5.0.1.rdfsP2F.has_type      :oil_doth;
12. :CIDOC5.0.1.rdfsP103F.was_intended_for        :RestOntheHunt_function;
13. :CIDOC5.0.1.rdfsP101F.had_as_general_use      :RestOntheHunt_function;
14. :CIDOC5.0.1.rdfsP53F.has_former_or_current_location      :Stockholm;
15. :CIDOC5.0.1.rdfsP53F.has_former_or_current_location      :Rome;
16. :CIDOC5.0.1.rdfsP1F.is_identified_by      :PE_34604;
17. :CIDOC5.0.1.rdfsP1F.is_identified_by      :TA_959a;
18. :CIDOC5.0.1.rdfsP65F.shows_visual_item      :Inscription_of_RestOntheHunt_PE34604;
19. :CIDOC5.0.1.rdfsP45F.consists_of      :linseed_oil;
20. :CIDOC5.0.1.rdfsP45F.consists_of      :canvas_duck;
21. :CIDOC5.0.1.rdfsP62F.depicts          :Bambocciate;
22. :CIDOC5.0.1.rdfsP55F.has_current_location      :Stockholm;
23. :CIDOC5.0.1.rdfsP49F.has_former_or_current_keeper        :Museum_Hallwylska;
24. :CIDOC5.0.1.rdfsP54F.has_current_permanent_location      :Stockholm;
25. :CIDOC5.0.1.rdfsP70B.is_documented_in        :RestOntheHunt.jpg;
26. :CIDOC5.0.1.rdfsP108B.was_produced_by      :Creation_of_RestOntheHunt_PE34604 .
```

Figure 16: Formal specification of a museum object modeled in the CIDOC-CRM.

generation where concepts and relationships can not easily be mapped to natural language.

According to the CIDOC-CRM specifications, a museum object is represented as an instance of the concept *E22.Man_Made_Object*, which has several properties including:[58] *P55.has_current_location*, *P43.has_dimension*, *P45F.consists _of*, *P101F.had_general_use*, *P108B.was_produced_by*. A concrete example of a formal specification (presented in turtle annotation) of the *RestOntheHunt_PE34604* object that was modeled according to the CIDOC Documentation Standards Working Group is given in Figure 16.

Taking the domain ontology structure as point of departure, the information in hand is an unordered set of statements that convey a piece of information about an object. The information the *RestOntheHunt_PE34604* statements convey spans at least four of the semantic sequences that we outline in section 9.2. To generate a coherent text, some ordering constraints must be imposed upon them. This is in particular important because a statement may map to an addition set of statements about an object, for example the relationship *P108B.was_produced_by* maps to an

[58] *Property* is a synonym for *relationship* that maps between two instances. In this paper we use the term *statement* to refer to a relationship between instances.

instance of the concept *E12.Production* that has the following properties: *P14F.carried_out_by*, *P7F.took_place_at*, *P4F.has_time_span*.

9.4 From formal specifications to coherent representation

As we pointed out in the previous section, the structure of the ontology is not a good point of departure for producing coherent texts and therefore requires pre-processing. In broad terms this involves taking a set of information elements to be presented to a user and imposing upon this set of elements a structure which provides a resulting text with fluency and cohesion.

Some of the pre-processing steps that have been suggested by previous authors (O'Donnell et al. 2001; Bontcheva 2005) include removing repetitive statements that have the same property and arguments and grouping together similar statements to produce a coherent summary. Although there is a need to select statements that mirror linguistic complexity (Mellish and Pan 2008), most authors focus on the semantics of the ontology rather than on the syntactic form of the language. They assume that the ontology structure is appropriate for natural language generation, an assumption which in many cases only applies to English.

In this section we describe the approach we exploit to learn how the ontology statements are realized and combined in natural occurring texts. We perform a domain specific text analysis; texts are studied through text linguistics by which the critic seeks to understand the relationships between sections of the author's discourse.

9.4.1 Linking statements to lexical units

When text generation proceeds from a formal representation to natural language output, the elements of the representation need to be somehow linked to lexical items of the language. We examined around 100 object descriptions in English, Swedish and Hebrew and studied how statements are ordered, lexicalised and combined in the discourse. To capture the distribution of discourse entities across text sentences we perform a semantic and syntactic analysis, we assume that our unit of analysis is the traditional sentence, i.e. a main clause with accompanying subordinate and adjunct clauses. Below we exemplify how the

ontology statements are mapped to lexical items in the studied texts.[59] Statements:

1. *P55F. has_current_location* maps between instances of *E22.Man-Made-Object* and instances of *E53.Place* (see line 22, Figure 16)

2. *P52F. has_current_owner* maps between instances of *E22.Man-Made-Object* and instances of *E40. Legal Body* (see line 9, Figure 16)

3. *P82F.at_some_time_within* maps between instances of *E52. Time-Span* and *String* data values.

Text examples:
Eng> The subject made its first appearance [in 1880]$_{P82F}$. It is [now installed]$_{P52F}$ in the Wallace Collection[,]$_{P55F}$ London.

Swe> Först [på 1900 talet]$_{P82F}$ kom den till Sverige och [hänger nu på]$_{P55F}$ Gripsholms slott [i]$_{P52F}$ Statens porträttsamling.

Heb> hatmuwnah hegieh larisunah leAeretz yisraAel [besnat 1960]$_{P82F}$. hyA [sayeket le]$_{P52F}$-quwleqitzyah sel Amir bachar [senimtzet]$_{P55F}$ bemuwzeyAuwn haAretz betel Aabiyb

These text examples exhibit a few local linguistic differences between the languages. In English and Hebrew, the order of the statements is: 3,2,1 while in the Swedish text it is: 3,1,2. It is interesting to note how the domain entities and properties are lexicalized in the different languages. In all three languages the property *P82F.at_some_-time_within* is lexicalised with a preposition phrase. On the other hand, the lexicalisation of the property *P55F. has_current_location* differs significantly. Furthermore, in the Swedish text all statements are realized in one single sentence; the statements are combined with a simple syntactic aggregation using the conjunction *och* 'and'. Both in the English and the Hebrew examples, statements 3 and 2 are realized as two sentences which are combined with a referring pronoun, i.e. *it* and *hyA*. When generating natural occuring texts it is important to utilize a generation machinery that supports such syntactic variations. In section 9.5 we demonstrate how these variation are supported in the GF formalism.

Empirical representations of stereotypical clause structures such as presented above not only provide evidence on how to pair ontology

[59]The transliteration ISO-8859-8 ASCII characters of Hebrew are used to enhance readability.

Table 9.1: Template specification that governs text structures of a cultural object in a museum.

Name	Template slot
T1	(a) object's title | (b) object's creator | (c) creation date | (d) creation place
T2	(a) creator date of birth | (b) creator date of death
T3	(a) object id | (b) object material | (c) object size
T4	(a) current owner | (b) current location | (c) catalogue date | (d) collection
T5	(a) object's identifier | (b) identified place

statements with lexical units according to the language specific patterns, but also guide template constructions proceeding according to the organization of the domain semantics.

9.4.2 Template specifications

In section 9.2 we presented a five stage typical GSP for a work of art object description. To guarantee that the selected statements follow this structure, we defined a sequence of templates describing the discourse structure, this approach was first introduced by McKeown (1985). Each sequence in a template consists of slots that correspond to a set of statements in the domain knowledge.

The template specification as whole provides a set of ordering constraints over a pattern of statements in such a way that may yield a fluent and coherent output text. The templates and slots are specified in Table 9.1.

9.4.3 A discourse schema

A discourse schema is an approach to text structuring through which particular organizing principles for a text are defined. It straddles the border between a domain representation and well-defined structured specification of natural language that can be found through linguistic analysis. This idea is based on the observation that people follow certain standard patterns of discourse organization for different discourse goals in different domains.

Our text analysis has shown certain combinations of statements are more appropriate for the communicative goal of describing a museum object. Following our observations, we defined a discourse schema *Description schema* (see below) consisting of two rhetorical predicates (e.g. Identification–Property and Attributive–Property).[60] The schema encodes communicative goals and structural relations in the analyzed texts. Each rhetorical predicate in the schema is associated with a set of templates (specified in Table 9.1). The notation used to represent the schema: ',' indicates the mathematical relation *and*, '{}' indicates optionality, '/' indicates alternatives.

Description schema:

Describe–Object $->$
  Identification–Property/
  Attributive–Property

Identification–Property $->$
  T1 , {T2 / T3}

Attributive–Property $->$
  T4 / T5

An example taken from one of the studied texts:

> [T1b]Thomas Sully [T2](1783-1872) painted this half-length [T1a] Portrait of Queen Victoria [T1c] in 1838. The subject is now installed in the [T4d] Wallace Collection, [T4b] London.

The first sentence which corresponds to the rhetorical predicate *Identification –Property*, captures four statements (comprising the following relationships: *P82F.at_some_time_within*, *P14F.carried_out_by*, *P108B. was_produced_by* and *P102. has_title*) that are combined according to local and global text cohesion principles.

9.5 Domain dependent grammar-based generation

After the information from the ontology has been selected and organized according to the pre-defined schema, it is translated to abstract

[60]The notion of *rhetorical predicates* goes back to Aristotle, who presented predicates as assertions which a speaker can use for persuasive argument.

grammar specifications. The grammar formalism is the Grammatical Framework (GF) (Ranta 2004), a formalism suited for describing both the semantics and syntax of natural languages. The grammar is based on Martin-Löf's type theory (Martin-Löf 1984) and is particularly oriented towards multilingual grammar development and generation. GF allows the separation of language-specific grammar rules that govern both morphology and syntax while unifying as many lexicalisation rules as possible across languages. With GF it is possible to specify one high-level description of a family of similar languages that can be mapped to several instances of these languages. The grammar has been exploited in many natural language processing applications such as spoken dialogue systems (Ljunglöf and Larsson 2008), controlled languages (Khegai, Nordström and Ranta 2003) and generation (Johannisson 2005).

GF distinguishes between abstract syntax and concrete syntax. The abstract syntax is a set of functions (fun) and categories (cat) that can be defined as semantic specifications; the concrete syntax defines the linearization of functions (lin) and categories (lincat) into strings that can be expressed by calling functions in the resource grammar. [61] Each language in the resource grammar has its own module of inflection paradigms that defines the inflection tables of lexical units and a module for specifying the syntactic constructions of the language.

Below we present the abstract and concrete syntax of the rhetorical predicate *Identification–Property* presented in section 9.4.3.[62] Figure 17 illustrates the abstract syntax tree of our abstract grammar that reflects on the semantics of the domain and that is common for all languages.

abstract syntax
cat
IdentificationMessage; ObjTitle; CreationProperty; Artist; TimeSpan;
CreationStatement; ArtistClass; TimeSpanClass;
fun
Identification: ObjTitle → CreationStatement → IdentificationMessage;
CreationAct: CreationStatement → TimeSpanClass → CreationStatement;
HasCreator: CreationProperty → ArtistClass → CreationStatement;
CreatorName: Artist → ArtistClass;
CreationDate: TimeSpan → TimeSpanClass;
Year : Int → TimeSpan ;
RestOnTheHunt: ObjTitle;
JohnMiel: Artist;
Paint: CreationProperty;

[61] A resource grammar is a fairly complete linguistic description of a specific language. GF has a resource grammar library that supports 14 languages.

[62] The GF Resource Grammar API can be found at the following URL: <http://www.grammaticalframework.org/lib/doc/synopsis.html>.

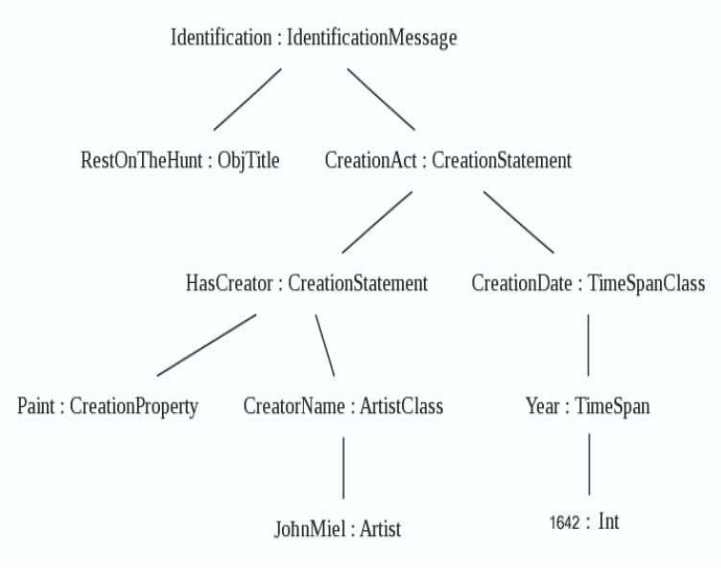


Figure 17: Abstract syntax tree for *Rest on the Hunt was painted by John Miel in 1642.*

The abstract specification expresses the semantics of the ontology and is language independent. What makes the abstract syntax in particular appealing in this context is the ability to expand the grammar by simply adding new constants that share both common semantics and syntactic alternations. For example, Beth Levin's (Levin 1993) English *Performance Verbs* class contains a number of verbs that can be added as constants of type *CreationProperty*, such as *draw* and *produce*, as follows: Paint, Draw, Produce : CreationProperty.

GF offers a way to share similar structures in different languages in one parametrized module called *functor* (Ranta 2009). In our implementation the common structure of the concrete syntax for English and Swedish is shared in a functor. Since the function *CreationDate* is linearized differently, it is defined separately for each language. This is illustrated below.

incomplete concrete syntax[63]
lincat
IdentificationMessage = S ;
TimeSpanClass, ArtistClass = Adv ;
TimeSpan = NP ;
CreationStatement = VP ;
CreationProperty = V2 ;
ObjTitle, Artist = PN ;

lin
Identification np vp = mkS pastTense (mkCl (mkNP np) vp);
CreationAct vp compl = mkVP vp compl;
HasCreator v np = (mkVP (passiveVP v) np) ;
CreatorName obj = (mkAdv by8agent_Prep (mkNP obj));
Year y = mkNP (SymbPN y) ;

concrete English syntax
lin CreationDate obj = (mkAdv in_Prep obj);

concrete Swedish syntax
lin CreationDate obj = mkAdv noPrep (mkCN year_N (mkNP obj));

The lexicon is implemented as an interface module which contains *oper* names that are the labels of the record types. It is used by the functor and by each of the language specific lexicons.

interface lexicon oper
year_N : N;
restOnTheHunt_PN : PN ;
johnMiel_PN : PN ;
paint_V2 : V2 ;

instance English lexicon
oper restOnTheHunt_PN = mkPN [“Rest on the Hunt”];
johnMiel_PN = mkPN “John Miel”;
year_N = regN “year”;
paint_V2 = mkV2 “paint” ;

[63]The word *incomplete* suggests that the functor is not a complete concrete syntax by itself.

instance Swedish lexicon
oper restOnTheHunt_PN = mkPN [“Rastande jägare”];
johnMiel_PN = mkPN “John Miel”;
year_N = regN “år”;
paint_V2 = mkV2 “måla” ;

In GF it is possible to built a regular grammar for new languages by using simple record types. In our case we implemented a small application grammar for Hebrew, i.e. *concrete Hebrew* that uses the same abstract syntax as for English and Swedish. In this module functions are linearized as strings where records *{s : Str}* are used as the simplest type.[64] We introduce the parameter type *Gender* with two values: Masc and Fem, these are used in table types to formalize inflection tables. In Hebrew, verb phrases are parameterized over the gender and are therefore stored as an inflection table {s : Gender => Str}; noun phrases have an inherent gender that is stored in a record together with the linearized string {s : Str ; g : Gender}.[65]

concrete Hebrew syntax
lincat
IdentificationMessage, TimeSpan, ArtistClass, TimeSpanClass = {s : Str};
Artist, ObjTitle = {s : Str ; g : Gender}; CreationProperty, CreationStatement = { s : Gender => Str};
lin
Identification np vp = {s = np.s ++ vp.s ! np.g };
CreationAct vp compl = { s = \g => vp.s ! g ++ compl.s };
HasCreator v obj = { s = \g => v ! g ++ obj.s};
CreatorName obj = { s = [“al yedey”] ++ obj.s };
CreationDate obj = { s = [“be”] ++ obj.s };
ObjTitle = {s = [“menuhat tzayydym”] ; g = Fem};
JohnMiel = {s = [“guwn miyAe”] ; g = Masc};
Paint = { s = table {Masc => “tzuwyr”; Fem => “tzuwyrah”}};
Param
Gender = Fem | Masc ;

[64] The resource grammar for Hebrew is currently under development.

[65] Hebrew has a more complex morphology as the one described here. However, in this implementation we changed the grammar so that it takes only care of gender agreement.

The complete grammar specifications yield the following text, in English, Swedish and Hebrew:

> Eng> Rest on the Hunt was painted by John Miel in 1642. The painting is located in the Hallwyska museum in Stockholm.
> Swe> Rastande jägare blev målad av John Miel år 1642. Tavlan hänger på Hallwyska museet i Stockholm.
> Heb> menuhat tzayydym tzuwyrah 'al yedey guwn miyAel be-1642. htmwnh memukemet be-muwzeyAuwn hallwiska be-stukholm.

This kind of multi-level grammar specification maps non-linguistic information to linguistic representation in a way that supports local and global text variations. For example, in the English and the Hebrew concrete syntax, the sentence complement is realized as a prepositional phrase (signalled by the prepositions *in* and *be*), but in the Swedish sentence, the complement is realized as a noun phrase (signalled by the noun *år*). In the above example this is illustrated in the linearization of *CreationDate*. In the Swedish concrete syntax no preposition is used (*noPrep*), and a different NP rule is applied to generate the noun phrase *år 1642*, i.e. CN→ NP → CN. Lexical variations are supported by the grammar as well, for instance, the verb *located* is not a direct translation of the Swedish verb *hänger* 'hang' but the interpretation of the verb in this context implies the same meaning, namely, the painting exists in the Hallwyska museum. The choice of the lexical unit are governed by the semantic structure of the ontology that is reflected in the abstract syntax.

While the functional orientation of isolated sentences of language is supported by GF concrete representations, there are cross-linguistic textual differences that we touched upon in section 9.4.1 and that are not yet covered in the grammar specifications, i.e. patterns with which cohesive and coherent texts are created. In English, cohesive means comprise conjunction, substitution and ellipsis that can frequently be used to realize a logical relation. In Swedish, cohesive means is often realized as elliptical item, preposition phrase, and/or punctuation. Whereas in Hebrew means of cohesion are realized through the verbal form, usage of ellipsis and conjunctive elements are not common.

9.6 Conclusion

In this paper we have presented a grammar driven approach for generating object descriptions from formal representations of a domain specific ontology. We illustrated how the lexicons of individual languages pair ontology statements with lexical units which form the backbone of the discourse structure. We demonstrated how schema based discourse structure is mapped to an abstract grammar specification using the domain specific ontology concepts and properties.

We are now in the process of the development of schemata that are being continually modified and evaluated; each rhetorical predicate should capture as many sentence structure variations as possible. A limitation of discourse schemata development is that it requires a lot of human efforts, however once a discourse schema is defined it can automatically be translated to abstract grammar specifications. This method of assembling coherent discourses from basic semantic building blocks will allow any generation system to assemble its texts dynamically, i.e. re-plan portion of its text and communicate successfully.

In the nearest future we intend to extend the grammar to support grouping of rhetorical predicates which requires a certain coverage of linguistic phenomena such as ellipsis, focus, discourse and lexical semantics. The long challenge of this work is in capturing linguistic properties of a language already during the schema development process to guide further development of language independent grammar specifications.

Part III

MLG generation from SW ontologies

10 THE PRODUCTION OF DOCUMENTS FROM ONTOLOGIES

Dannélls, Dana 2008c. The production of documents from ontologies. *Proceedings of the 18th european conference on artificial intelligence (ECAI)*, 36–38. Patras, Greece: IOS Press.

10.1 Introduction

A major challenge for a language generator developer who wishes to make use of Semantic Web ontologies is how to alter the input knowledge-base, so as to verbally express contents that describe a concept in an ontology. This task becomes even harder when the user preferences such as the preferred language, text length and syntax must be computed.

Our research project aims to adapt the presentation of a text content for a specific readership from Web ontologies. As a primary step towards accomplishing this aim we utilized a domain specific Web Ontology Language (OWL) and started to exploit how natural language texts may be produced from this expressive language. Below we outline a number of steps which we believe are significant for the quality of the produced text:

1. Selection of the axioms describing a concept;[66]
2. Presentation order of the selected axioms;
3. Verbalization and realization of the selected and ordered axioms.

In this paper we focus on the third step and show that given the selected ontology content, verbalization and realization of the relationships and classes describing a concept exhibit great variations, which depend on

[66] An axiom is an ontology statement which states the relationships among concepts.

the context in which they appear. We illustrate some of these variations and discuss their implications for text production.

The remainder of this paper is structured as follows. Section 10.2 provides an overview of previous work on generation from ontologies and discusses a number of the advantages and challenges that Web ontology languages pose to language generators. Section 10.3 provides a description of the domain ontology and the domain ontology language. Section 10.4 exemplifies the difficulties in verbalizing the knowledge contained in the ontology which we came across while attempting to produce coherent and cohesive texts. Section 10.5 ends up with conclusions and main directions for future research.

10.2 Background

There are many definitions for the term *ontology* (Staab and Studer 2004). In this context, an ontology is defined as a structured framework for modeling the concepts and relationships of some domain expertise, which provides the structural and semantic ground for computer based processing of domain knowledge. To allow better use of ontologies in applications, traditional ontology language standards such as DAML and OWL[67] have been specified by the World Wide Web Consortium (W3C). One of the purposes of these established standards is to enable better communication between humans and machines in which information is given a well defined meaning.

10.2.1 Generating from ontologies

Generation techniques deal with the process of converting semantic representation into surface form in a particular language. The features of the text produced are normally chosen with respect to a particular target reader group. There have been successful attempts to develop natural language generation tools that generate texts from Web ontology languages (Bontcheva 2005; Bontcheva and Wilks 2004; Wilcock 2003; Wilcock and Jokinen 2003).

Wilcock (2003) presents an approach in which the concepts defined in the ontology are employed for generating the lexicon. Bontcheva and Wilks (2004) concentrate on the semantic representations encoded in Semantic Web standards and discuss how these can be exploited to

[67] http://www.w3.org/TR/

generate text summaries. They point out the content of the ontology itself as a major factor for the quality of the output. Gawronska and Erlendsson Gawronska and Erlendsson (2005) show how biological ontologies as Kyoto Encyclopedia of Genes and Genomes, may be utilized for generating graphs representing the essential contents of biomedical scientific articles.

Mellish and Sun (2006b) describe the large extent of linguistic material in existing Web ontologies and its complexity. They exemplify how an extended text with multiple sentences can be generated from class axioms.

Similarly to Wilcock and Jokinen (2003); Bontcheva and Wilks (2004), this work is concerned with generating textual descriptions of concepts from a domain-specific ontology. As opposed to Mellish and Sun (2006b), this approach deals with individuals and requires manual input of the lexicon. In contrast to Wilcock and Jokinen (2003) who uses templates to produce texts, we intend to utilize a grammar-based surface realiser to enhance linguistic variations in the generated texts.

10.2.2 Opportunities and challenges

As pointed out by many authors, there are several advantages which make Web ontology languages such as OWL particularly suitable to generate from. For example, axioms can be seen as forming a graph in which routes between axioms correspond to different possible transitions in a coherent text (Mellish and Pan 2008); axioms can be used to accommodate a generation system to different contextual degrees and user needs; the use of multiple-inheritance converts the class hierarchy into a directed graph and not a tree structure.

Web ontologies provide implicit information about a domain. This is an advantage that has been exploited by a number of Natural Language Generation (NLG) systems (Paiva 1998) who utilize the domain background knowledge base to complete generation related tasks. In many domain ontologies the ontology concepts used to express classes and relationships are similar to their lexical entry, which in many aspects facilitate the generation tasks. However, natural languages are ambiguous and even ontologies which do not make a distinction between the ontology concepts and natural language words that describe them, contain ambiguities that need to be resolved.

To reveal implicit information about a concept, inferences have to be drawn. These inferences that are mostly based on DL (Reiter and

Mellish 1992), might render in different axiom sets, depending on the axiom selection constraints, such as constraints that are set due to the user preferences. Furthermore, it is necessary to fully understand what the knowledge in the selected axiom set actually states before natural language words can be expressed. The content and knowledge formalized in an ontology can lead to ambiguous content interpretations, and can also bring up problems during the process of verbalization. This has brought with it an awareness of the need to encode linguistic knowledge about concepts directly into ontologies (Judgem, Sogrin and Troussov 2007).

10.3 The domain ontology model

The work described in this paper is based on the CIDOC Conceptual Reference Model (CRM) ontology,[68] which is an initiative to construct an ontology within the Cultural Heritage (CH) domain. The CIDOC ontology consists of 81 relations and 244 concepts and is available in various formats, among which is OWL. It contains facts about concepts (sets of objects) and roles (binary relations) and provides a conceptual model that subscribes an object-centred view of the CH domain.

10.3.1 Population and maintenance

Since the CIDOC-CRM ontology does not contain information about individuals (single objects), populating the ontology was a necessary step. We enhanced the ontology with additional lexical entries, as well as new concepts and relationships.

On the task of ontology population, most of the work that has been carried out relates to information extraction from unstructured natural language text or semi-structured HTML pages (Karkaletsis, Valarakos and Spyropoulos 2005). In our work, the process of ontology population was conducted manually, it is based on a small corpus of CH texts that we have collected from internal museum repositories. Following the guidelines given by the reference document (Crofts et al. 2009) for filling in concept-values along with a thorough analysis of the information content, we have so far enriched the ontology with a total of 150 new concepts. Each concept was assigned with its lexical lemma that links to a lexical string-name.

[68] http://cidoc.ics.forth.gr/

10.3.2 The ontology terminology

An OWL ontology (lite or DL) has a description logic based semantics which consists of a set of axioms. Axioms assert facts about concepts (Tbox) and facts about individuals (Abox). Roles are usually asserted in the form of inclusion axioms.

As with any representation of an OWL ontology, the CIDOC CRM ontology contains classes (concepts) that define a group of individuals that belong together because they share some properties (roles). A subclass is a class that is a specialization of another class (its superclass). According to the CRM documentation, *specialization* means: (1) all instances of the subclass are also instances of its superclass; (2) the intension of the subclass extends the intension of its superclass; (3) the subclass inherits the definition of all of the properties declared for its superclass in addition to having one or more properties of its own.

Properties serve to define relationships of a specific kind between two classes. A property can have a subproperty which is a specialization of another property (its superproperty). A property must be defined with reference to both its domain and range. The term *specialization* in the context of properties has similar meaning as for classes with additional restrictions, i.e: (4) the domain of the subproperty is the same as the domain of its superproperty or a superclass of that domain; (5) the range of the subproperty is the same as the range of its superproperty or the subclass of that range.

10.4 Realization of a concept in the ontology

In the semantics of OWL, a given axiom may be expressed in several ways and may have more than one realization possibilities. In this section we exemplify some of the discussed challenges (see section 10.2.2) which are related to realization of concepts in the CIDOC-CRM ontology.

10.4.1 A concept representation

The following example, taken from our ontology, describes the class *EdelfeltProduction*. This particular class comprises a set of productions that has been carried out by Albert Edelfelt.[69]

[69] According to the CRM reference document: "a production can present activities, that are designed to, and succeed in, creating one or more new items".

The example presents an ontology content that describes the concept *EdelfeltPortraitProduction* formulated in an RDF syntax. The knowledge it conveys is that a production of a portrait took place in France and was made by Albert Edelfelt between 1880 and 1890.

```
<museum:EdelfeltProduction
   rdf:about="#EdelfeltPortraitProduction">
 <crm:P14F.carried_out_by>
  <crm:E21.Person rdf:about="#AlbertEdelfelt"/>
  </crm:P14F.carried_out_by>
  <crm:P12F.occurred_in_the_presence_of>
  <crm:E21.Person rdf:about="#AlbertEdelfelt"/>
  </crm:P12F.occurred_in_the_presence_of>
 <crm:P7F.took_place_at>
  <crm:E48.Place_Name rdf:about="#France"/>
  </crm:P7F.took_place_at>
 <crm:P4F.has_time_span>
<crm:E49.Time_Appellation rdf:about="#1880-1890"/>
  </crm:P4F.has_time_span>
</crm:E12.Production>
```

The class *EdelfeltProduction* is a subclass of *E12.Production*. *E12. Production* has multiple subclasses, i.e. *E11.Modification* and *E63.Beginning_of_Existence*, this is shown below.[70]

```
<owl:Class  rdf:about="&crm;E12.Production">
 <rdfs:subClassOf rdf:resource="&crm;E11.Modification />
 <rdfs:subClassOf rdf:resource="&crm;E63.Beginning_of_Existence />
</owl:Class>
```

E11.Modification is a subclass of *E7.Activity* and *E63.Beginning_of_Existence* is a subclass of *E5.Event*, hence the inferred relation *P12F.occurred_in_the_presence_of*.

[70] The notation &crm; is used as a shortcut for the complete URL to the CIDOC-CRM ontology.

10.4.2 Surface realization

Given an ontology, populated with individuals, and some user preferences, the task is to verbalize and realize the selected ontolgy content. A straightforward realization of the above content describing the concept *EdelfeltPortraitProduction* may result in the following:

> This Edelfelt portrait production was carried out by Albert Edelfelt. The Edelfelt portrait production occurred in the presence of Albert Edelfelt. The Edelfelt portrait production took place in France. The Edelfelt portrait production has time span 1880-1890.

Inferred knowledge Inferred relationships may have distinguished interpretations, therefore in order to resolve their meaning knowledge about the domain and the context in which a concept appears are required. For example, following the above ontology fragment, we interpretate that the inferred relationship *P12F.occurred_in_the_presence_of* carries out redundant information within the context of the concept *EdelfeltPortraitProduction*, and thus does not contribute with new information. As a result of this interpretation, the inferred relationship could be eliminated, or "selected" and verbalized instead of the relationship *carried_out_by*. On the other hand, when a production describes an activity which has resulted in a movie production, e.g. within the context of the concept *TheLordOfTheRingMovieProduction*, the inferred relationship *P12F.occurred_in_the_presence_of* will not provide redundant information but rather contribute with new knowledge.

Verbalization The choice of the lexical entry encoding a relationship is both domain and user dependent, for example, the relationship *carried_out_by* could be verbalized as either "painted by" or "created by" depending on the concepts it describes. Furthermore, the choice between synonyms for the relationship *created_by* are various: "produce by", "bring out by", "develop by", "acquire by", etc. Some differences in categorisations or internal makeup must be present if the difference in information content is to be consequential.

When verbalizing the description about the concept *EdelfeltPortraitProduction* we want to establish a text which is more similar to the following:

> This portrait production was carried out by Albert Edelfelt. The production took place in France. It covers the period 1880-1890.

Humans are able to recognize that semantic representations are intimately linked, this realization process could be also automated rather easily. However, the problem of how words and other linguistic phenomena might be integrated with the internal representations that support reasoning is yet to be explored.

10.5 Conclusion and future work

We presented an ongoing research and illustrated the problems we encountered while attempting to generate coherent and cohesive texts from a Web ontology language. This research work is based on the domain specific CIDOC-CRM ontology. Text planning follows the ontology axioms structure; the assertional part of the ontology is developed manually; both the terminological part and the assertional part are applied to present parts of the ontology.

This paper showed that although OWL provides powerful reasoning opportunities for natural language generators, it poses difficulties to language generators that need to be resolved. We highlighted the problem of distinguishing between the inferable relationships that contribute with new knowledge in a particular context. Relationships might have a particular, quite specific interpretation depending on the context in which they appear and the concept they describe. This invokes a difficulty on choice of a lexical entry encoding a relationship.

Our research work is only in its early stages. Exploiting OWL for realization purposes and finding general, domain-independent solutions requires a considerable amount of work. In the near future we are planning to address issues related to content selection and lexical determination of relationships between concepts, a task which depends on the chosen semantic content, the concept it describes, the class hierarchy that is utilized to represent the concept, and the target language.

11 A FRAMEWORK FOR IMPROVED ACCESS TO SW DATABASES

Dannélls Dana, Mariana Damova, Ramona Enache and Milen Chechev 2011. A Framework for Improved Access to Museum Databases in the Semantic Web. *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage. Workshop associated with the RANLP 2011 Conference*, Hissar, Bulgaria.

11.1 Introduction

During the past few years several projects have been undertaken to digitize cultural heritage materials (Clough, Marlow and Ireson 2008; Dekkers, Gradmann and Meghini 2009) through the use of Semantic Technologies such as RDF (Brickley and Guha 2004) and OWL (Bechhofer et al. 2004). Today there exist large number of digital collections and applications providing direct access to cultural heritage content.[71]

However, digitization is a labour intensive process and is long from being complete. Because of the heterogeneous data structures different museums have, digitally encoded cultural material stored in internal museum databases requires advanced mapping and vocabulary integration for it to be accessible for Semantic Web applications. In addition to establishing ways for managing various vocabularies, and for exploiting semantic alignments across them automatically (van der Meij, Isaac and Zinn 2010), computer engineers also need to investigate automatic methods to make this information available to computer users in different forms and languages that are available to them.

Our work is a step towards this direction. It is about an automatic workflow of sharing data infrastructures that is explicitly targeted towards the Semantic Web. We have developed a method to manage and

[71] http://www.europeana.eu/portal/

access museum data by integrating it within a series of interlinked ontological models. The method allows querying and generation of query results in natural language using the Grammatical Framework (GF). We have been experimenting with data collections from the Gothenburg City Museum that we made available for querying in the Museum Reason-able View loaded in the triple store OWLIM.

In the remainder of this paper we present the ontologies that were merged including CIDOC-CRM,[72] PROTON,[73], the Painting ontology and the data that we have been experimenting with (Section 11.2). We describe the creation of the Museum Reason-able View with structured query examples (Section 11.3). In Section 11.4, we introduce the Grammatical Framework and demonstrate the mechanisms of interfacing between the structured data and natural language. We provide an overview of related work (Section 11.5) and end with conclusions (Section 11.6).

11.2 The ontologies and museum data

11.2.1 The CIDOC-CRM

The International Committee for Documentation Conceptual Reference Model (CIDOC CRM) that was accepted by ISO in 2006 as ISO21127 (Crofts et al. 2009), is one of a widely used standards that has been developed to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information.

The CIDOC CRM, independent of any specific application, is primarily defined as an interchange model for integrating information in the cultural heritage sector. Although it declares rich common semantics of metadata elements, many of the concepts that are utilized for describing objects are not directly available in this model. To arrive at the point where information that is available in museum databases about paintings could be recorded using this model, we developed the painting ontology that integrates the CIDOC-CRM with more specific schemata.

[72] The Conceptual Reference Model (CRM): http://cidoc.ics.forth.gr/

[73] http://proton.semanticweb.org/

11.2.2 The Swedish Open Cultural Heritage (SOCH)

The Swedish Open Cultural Heritage (SOCH) is a web service used to search and fetch data from any organization that holds information related to the Swedish cultural heritage.[74]

The idea behind SOCH is to harvest any data format and structure that is used in the museum sector in Sweden and map it into SOCH's categorization structure. The data model used by SOCH is an uniform data representation which is available in an RDF compatible form.

The schema provided by SOCH helps to intermediate data between museums in Sweden and the Europeana portal. More than 20 museums in Sweden have already made their collections available through this service. By integrating the SOCH data schema in the ontological framework we gain automatic access to these collections in a semantically interoperable way.

11.2.3 The Painting ontology

The painting ontology is a domain specific ontology. It is designed to support integration and interoperability of the CIDOC-CRM ontology with other schemata. The main reference model of the painting ontology is the OWL 2 implementation of the CRM.[75] The additional models that are correctly integrated in the ontology are: SOCH, Time Ontology,[76] SUMO and Mid-Level-Ontology.[77] The painting ontology was constructed manually using the Protégé editing tool.[78] It contains 184 classes and 92 properties of which 24 classes are equivalent to classes from CIDOC-CRM and 17 properties are sub-properties of CIDOC-CRM properties.

Integration of the ontology concepts are accomplished by using the OWL construct: *intersectionOf* as specified in the following example. In this example, the class *Painting* is defined in the painting ontology as a subclass of *E22_Man-Made_Object* class from the CIDOC-CRM ontology and is an intersection of two classes, i.e. *item* from the SOCH schema and *PaintedPicture* from the Mid-Level Ontology.

[74] http://www.ksamsok.se/in-english/
[75] http://purl.org/NET/cidoc-crm/core
[76] http://www.w3.org/TR/owl-time/
[77] http://www.ontologyportal.org/
[78] http://protege.stanford.edu/

```
<owl:Class rdf:about="&painting;Painting">
 <owl:equivalentClass>
  <owl:Class>
    <owl:intersectionOf
           rdf:parseType="Collection">
      <rdf:Description
           rdf:about="&ksasok;item"/>
      <rdf:Description
           rdf:about="&milo;PaintedPicture"/>
     </owl:intersectionOf>
    </owl:Class>
   </owl:equivalentClass>
 <rdfs:subClassOf
     rdf:resource="&core;E22_Man-Made_Object"/>
</owl:Class>
```

The schemata that are stated in the above example are denoted with the following prefixes: painting ontology (&painting), SOCH (&ksamsok), Mid-Level-Ontology (&milo) and CIDOC-CRM ontology (&core).

11.2.4 Proton

PROTON (Terziev et al. 2005) is a light weight upper level ontology, which was originally built with a basic subsumption hierarchy comprising about 250 classes and 100 properties providing coverage of most of the upper-level concepts necessary for semantic annotation, indexing, and retrieval. Its modular architecture allows for great flexibility of usage, extension, integration and remodeling. It is domain independent and complies with the most popular metadata standards like DOLCE,[79] Cyc,[80] Dublin Core.[81]

PROTON is encoded in OWL Lite, and contains a minimal set of custom entailment rules (axioms). It is interlinked with CIDOC CRM, and is used in the data integration model to provide access to the Linked Open Data (LOD) for Cultural Heritage Damova and Dannélls (2011).

Table 11.1: A painting object representation in the GCM database.

Field name	**Value**
Field nr.	4063
Prefix	GIM
Object nr.	8364
Search word	painting
Class 1	353532
Class 2	Gothenburg portrait
Amount	1
Producer	E.Glud
Produced year	1984
Length cm	106
Width cm	78
Description	oilpainting represents a studio indoors
History	Up to 1986 belonged to Datema AB, Flöjelbergsg 8, Gbg
Material	oil colour
Current keeper	2
Location	Polstjärnegatan 4
Package nr.	299
Registration date	19930831
Signature	BI
Search field	BO:BU Bilder:TAVLOR PICT:GIM

11.2.5 The Gothenburg City Museum (GCM) database

The Gothenburg City Museum (GCM) preserves 8900 museum objects described in two of the museum database tables. These two tables correspond to two of the museum collections, i.e. GSM and GIM. Each of these tables contains 39 properties for describing museum objects. Table 11.1 shows 20 of these properties, including the object type, its material, measurements, location, etc. All properties and object values stored in the database are given in Swedish.

The Gothenburg City Museum's data that is used as our experimental data follows the structure of the CIDOC-CRM but it contains many

[79] http://www.loa-cnr.it/DOLCE.html
[80] http://www.ontotext.com/downloads/cycmdb
[81] http://www.cs.umd.edu/projects/plus/SHOE/onts/dublin.html

concepts that are not available in CIDOC-CRM. So, in order to be able to fully integrate the Gothenburg City Museum data into a semantic view it was necessary to make use of concepts and relationships from the remaining ontologies.

Figure 18 shows how elements from the Goethenburg city museum are represented with elements from different schemata, e.g. CIDOC-CRM, PROTON, SOCH and the Painting ontology.

11.2.6 DBpedia

DBpedia (Auer et al. 2007) is the RDF-ized version of Wikipedia, comprising the information from Wikipedia infoboxes, designed and developed to provide as full as possible coverage of the factual knowledge that can be extracted from Wikipedia with a high level of precision. DBpedia describes more than 3.5 million things and covers 97 languages. 1.67 million of DBpedia things are classified in a consistent ontology, including 364,000 persons, 462,000 places, and 99,000 music albums. The DBpedia knowledge base has over 672 million RDF triples out of which 286 million extracted from the English edition of Wikipedia and 386 million extracted from other language editions.

DBpedia is used as an additional source of data, which can enrich the information about the Gothenburg museum data. For example, their location identified with the DBpedia resource referring to the city of Gothenburg.

11.3 Integrating and accessing museum data

11.3.1 Integration for flexible computing

Integrating datasets into linked data in RDF usually takes place by indicating that two instances from two datasets are the same by using the built in OWL predicate: `owl:sameAs`.[82] However, recent research (Damova 2011; Damova et al. 2011; Jain et al. 2011) has shown that interlinking the models according to which the datasets are described is a more powerful mechanism of dealing with large amounts of data in RDF, as it exploits inference and class assignment.

We have adopted this approach when creating the infrastructure for the museum linked data, including several layers of upper-level on-

[82] http://www.w3.org/TR/owl-ref/

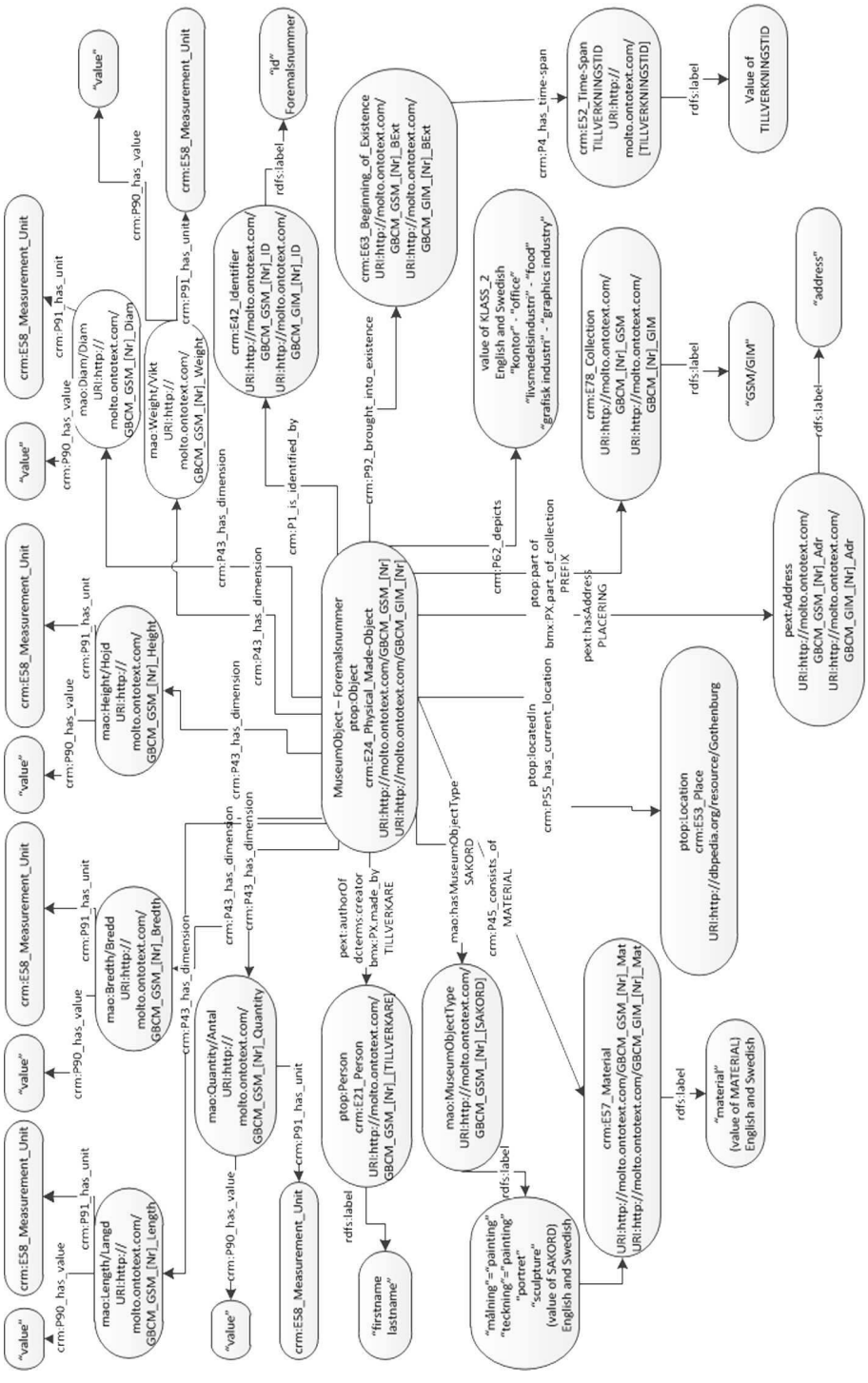


Figure 18: Dataset interconnectedness in the Museum Reason-able View.

tologies. They provide a connection to different sets of linked data, for example PROTON for the LOD cloud. They also provide an extended pool of concepts that can be referred to in museum linked data that do not directly pertain to the expert descriptions of the museum objects, and the strictly expert museum knowledge is left to CIDOC-CRM. This model of interlinked ontologies offers a flexible access to the data with different conceptual access points. This approach is implemented as a Reason-able View of the web of data (Kiryakov et al. 2009).

Using linked data techniques (Berners-Lee 2006) for data management is considered to have great potential in view of the transformation of the web of data into a giant global graph. Still there are challenges related to them that have to be handled to make this possible. Kiryakov et al. (2009) discuss these challenges and present an approach for reasoning with and management of linked data. In summary, a Reason-able View is an assembly of independent datasets, which can be used as a single body of knowledge with respect to reasoning and query evaluation. Each Reason-able View is aiming at lowering the cost and the risks of using specific linked datasets for specific purposes. We followed this approach when constructing the Museum Reason-able View with the data from the Gothenburg City Museum, DBpedia, Geonames and the ontologies listed in Section 11.2.[83]

The process of Gothenburg city museum data integration into the Museum Reason-able View consists in transforming the information from the museum database into RDF triples on the ontologies described in the previous section. Figure 19 shows the architecture of the Museum Reason-able View, which includes interconnected schemata and links to external to the Gothenburg museum data, such as DBpedia. The knowledge base contains close to 10K museum artifacts from the Gothenburg city museum, and the entire DBpedia.

11.3.2 Accessing Museum Linked Data

The Museum Reason-able View is loaded in OWLIM (Bishop et al. 2011) and its data are accessible via a SPARQL (Eric and Andy 2008) end point and keywords.[84] The queries can be formulated by combining predicates from different datasets and ontologies in a single SPARQL query, retrieving results from all different datasets that are part of the Reason-able View.

[83] Geonames website: http://www.geonames.org/

[84] The data is available at: http://museum.ontotext.com

11.3.3 The Museum Reason-able View

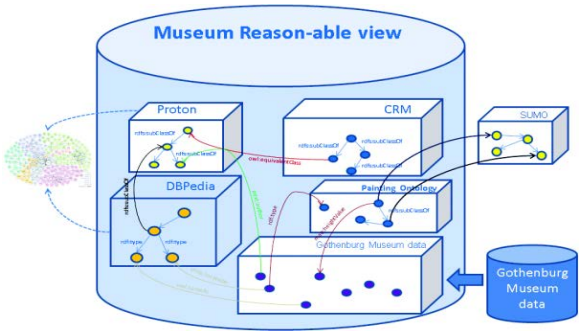


Figure 19: Integration of Gothenburg city museum data into the Museum Reason-able View.

A query example about the location, address, description and time of paintings by Carl Larsson is given below.

```
crm: <http://purl.org/NET/cidoc-crm/core#>
ptop: <http://proton.semanticweb.org/protontop#>
painting:
   <http://spraakbanken.gu.se/rdf/owl/painting#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
pext: <http://proton.semanticweb.org/protonext#>

select * where
{
?museumObject
        crm:P55_has_current_location ?location.
?museumObject painting:hasCategory
                    [rdfs:label "teckning"@sv].
?museumObject pext:authorOf
                 [rdfs:label "Carl Larsson"@sv].
?museumObject
        crm:P55_has_current_location ?location.
OPTIONAL {
 ?museumObject pext:hasAddress
                       [rdfs:label ?address].}
?museumObject  crm:P62_depicts  ?description .
?museumObject crm:P92_brought_into_existence
[ crm:P4_has_time-span [  rdfs:label  ?time ] ].}
```

SPARQL Query
Results for PREFIX crm: <http://purl....

Download in JSON | SPARQL Results in XML | SPARQL F

museumObject	location	collection	address	description	time
http://molto.ontotex...	http://dbpedia.org/r...	GIM@sv		glasindustri@sv	
http://molto.ontotex...	Göteborg@en	GIM@sv		glasindustri@sv	
http://molto.ontotex...	http://dbpedia.org/r...	GSM@sv	Huvudmagasinet Gårda@sv	1 göteborgsporträtt@sv	1889
http://molto.ontotex...	Göteborg@en	GSM@sv	Huvudmagasinet Gårda@sv	1 göteborgsporträtt@sv	1889
http://molto.ontotex...	http://dbpedia.org/r...	GSM@sv	Huvudmagasinet Gårda@sv	1 göteborgsporträtt@sv	1889
http://molto.ontotex...	Göteborg@en	GSM@sv	Huvudmagasinet Gårda@sv	1 göteborgsporträtt@sv	1889
http://molto.ontotex...	http://dbpedia.org/r...	GIM@sv		glasindustri@sv	
http://molto.ontotex...	Göteborg@en	GIM@sv		glasindustri@sv	
http://molto.ontotex...	http://dbpedia.org/r...	GSM@sv	Huvudmagasinet Gårda@sv	1 göteborgsporträtt@sv	1889

Figure 20: The results from a SPARQL query.

The above query returns the results that are depicted in Figure 20. Note that the returned location is the DBpedia resource about the city of Gothenburg. The results also show that museum items from the two collections – GIM and GSM – are harvested, which means that the data from the collections are integrated together and accessible from a single query point.

Other queries can be asked about the types of art work preserved in the museum, their material, or about artwork from a certain period of time, etc. Below follows another query example about the address, the time of paintings and the collection they are coming from.

```
select ?museumObject ?location ?collection
?address ?description ?time where
{?museumObject
          crm:P55_has_current_location ?location ;
ptop:partOf [ rdfs:label ?collection ] ;
painting:hasCategory [ rdfs:label "teckning"@sv ];
crm:P62_depicts ?description .
OPTIONAL {?museumObject pext:hasAddress
[ rdfs:label ?address ] .}
OPTIONAL {?museumObject
                  crm:P92_brought_into_existence
[ crm:P4_has_time-span [ rdfs:label ?time ] ] .}}
```

The Reason-able View is accessible with SPARQL queries, which require intimate knowledge of the schemata describing the data, and technical expertise in SPARQL. Moreover, the results from SPARQL are not always easy to understand, in particular if the retrieved information is given in a language other than English. This is why the results are send forward to the NLP component to verbalize the ontology links.

11.4 Ontologies verbalization

11.4.1 The Grammatical Framework (GF)

The Grammatical Framework GF (Ranta 2004) is a grammar formalism, based on Martin-Löf's type theory (Martin-Löf 1982). Its key feature is the division of a grammar in the abstract syntax-which acts as a semantic interlingua and the concrete syntaxes-representing verbalizations in various target languages (natural or formal).

GF comes with a resource library (Ranta 2009), where the abstract syntax describes the most common grammatical constructions allowing text generation, which are further mapped to concrete syntaxes corresponding to 18 languages.[85] The resource library aids the development of new grammars for specific domains by providing the operations for basic grammatical constructions, and thus making it possible for users without linguistic background to generate syntactically correct natural language.

To verbalize the data that is stored in the Museum Reason-able View, we utilize GF. The advantages of using GF for verbalization is three fold: it provides mechanisms for type checking, by validating coercions between the basic class of an instance and the class required by the definition of the relation that uses it; the framework offers support of direct verbalization which makes it easier to generate text from the ontology and so to create natural language applications using it without the aid of external tools; GF has a resource library that cover the syntax for 18 languages.

11.4.2 Translation of the Museum Reason-able View to GF

The capabilities of GF as a host-language for ontologies were already investigated by Enache and Angelov (2010), where SUMO, the largest open-source ontology was translated to GF. It was shown that the type system provides a robust framework for encoding classes, instances and relations. The same basic implementation design that was used for encoding SUMO in GF is applied in this work for representing the Museum Reason-able View.

The classes form a hierarchy modelled by an inheritance relation, which is the reflexive-transitive closure of the subclass relation `rdfs:subClassOf` from the ontology, are encoded as functions in the GF

[85] www.grammaticalframework.com

grammar. Other information stated in the ontology, is encoded in GF as axioms, external to the grammar. These are used for verbalization as in the following example from the OWL entry corresponding to the painting *Big Garden*:

```
<owl:NamedIndividual
  rdf:about="&painting; BigGardenObj">
 <rdf:type
  rdf:resource="&painting;Painting"/>
 <isPaintedOn
  rdf:resource="&painting;Canvas"/>
 <createdBy
  rdf:resource="&painting;CarlLarsson"/>
 <hasCreationDate rdf:resource=
                "&painting;Year1937"/>
</owl:NamedIndividual>
```

A representation of the instance *BigGardenObj* is defined as follows:

```
 fun BigGardenObj : Ind Painting ;
```

Where the *Painting* was defined previously as a class. The remaining information about *Big Garden* from the ontology is encoded as a set of axioms with the following syntax:

```
isPaintedOn (el BigGradenObj) (el Canvas)
createdBy (el BigGardenObj)(el CarlLarsson)
hasCreationDate (el BigGardenObj) (el (year 1937))
```

A couple of clarifying remarks about the GF encoding are needed in order to understand better the representation of the ontology: the dependent type `Ind` is used to encode class information of instances, and the wrapper function `el` is used to make the above-mentioned coercion, where the two types, along with the inheritance object that represents the proof that the coercion is valid are not visible here, since GF features implicit arguments.

In GF, the natural language generation is based on composeable templates. We obtain the verbalization of classes and templates automatically, mainly based on their Camel-Case representation. For the relations, more work is needed, since a grammatically correct verbalization is not possible based only on the ontology information.

Below follow a few English sentence examples that we are able to generate:

- *B*ig Garden is a painting
- *B*ig Garden is painted on canvas
- *B*ig Garden is painted by Carl Larsson
- *B*ig Garden was created in 1937

Below we provide examples for ontology relations in the shape of *O1 is painted by O2* and feed these to the GF parser which will build an abstract syntax tree, from which we abstract over the placeholders *O1* and *O2*, replacing them with function arguments. For example, the relations `hasCurrentLocation` and `hasCreationDate` have the following abstract syntax representation:

```
fun hasCurrentLocation : El Painting
 -> El Place -> Formula ;

fun  Painting_hasCreationDate :
 El Painting_Artwork
 -> El Painting_TimePeriod -> Formula ;
```

Their English representation in the concrete syntax is:

```
lin hasCurrentLocation o1 o2 =
 mkPolSent (mkCl o1
 (mkVP (passiveVP locate_V2)
      (mkAdv at_Prep o2))) ;

lin  Painting_hasCreationDate o1 o2 =
 mkPolSentPast (S.mkCl o1 (S.mkVP
   (S.passiveVP create_V2)
   (S.mkAdv in_Prep o2))) ;
```

Since the parser uses the resource library grammars, the result sentence will be syntactically correct, regardless of the arguments we use it with. Also, one does not need extensive knowledge of the GF library or GF programming in order to build verbalization. This might not make a difference for English, which is morphologically simple, but future work involves building such a representation for French, German, Finnish and Swedish, where it would be more difficult to achieve correct agreement, without grammatical tools.

Below follows an example of how the construct *owl:intersectionOf* is represented in the GF abstract syntax:

```
Equiv_TimePeriod =  Equivalent TimePeriod
 (both E52_TimeSpan Sumo.YearDuration) ;
```

Equivalent Class Class is a dependent type that encodes type equivalence.

11.5 Related Work

Museum Data Integration with semantic technologies as proposed in this paper is intended to enable efficient sharing of museum and cultural heritage information. Initiatives for developing such sharing museum data infrastructures have emerged in the recent years. Only a few of them rely on semantic technologies.

The Museum Data Exchange 2010 project has developed a metadata publishing tool to extract data in XML.[86] Brugman, Malaisé and Hollink (2008) have developed an Annotation Meta Model providing a way of defining annotation values and anchors in an annotation for multimedia resources. The difference between these approaches and our approach is that we chose to reuse many of the concepts and the relationships that are already defined in the CIDOC-CRM model.

Other related initiatives in the Web of structured data is the Amsterdam Museum Linked Open Data project,[87] aiming at producing Linked Data within the Europeana data model (Dekkers, Gradmann and Meghini 2009; Haslhofer and Isaac 2011), and the National Database Project of Norwegian University Museums (Ore 2001) who developed a unified interface for digitalizing cultural material.[88]

In Sweden, as well as other countries, semantic technologies enter the cultural heritage field increasingly and there have been some suggestions describing the tools and techniques that should be applied to digitalize the Swedish Union Catalogue (Malmsten 2008). Following these ideas and other experiences with museum data (Bryne 2009) that have shown that conversion of museum databases is best approached through integration of existing models, we decided to invest in a manual design step to built a framework that captures specific characteristics of museum databases.

[86] http://www.oclc.org/research/activities/museumdata/default.htm
[87] http://www.europeana.eu/portal/thoughtlab_linkedopendata.html
[88] http://www.muspro.uio.no/engelsk-omM.shtml

To our knowledge, we made the first attempt of using CIDOC-CRM to produce museum linked data with connections to external sources like DBpedia. Our attempt to generate natural language sentences from ontologies, and more precisely from the structured results of SPARQL queries are the novelty of the work presented in this paper.

11.6 Conclusions

We presented a framework for integrating and accessing museum linked data, and a method to present this data using natural language generation technology.

A series of upper-level and domain specific ontologies have been used to transform Gothenburg museum data from a relational database into RDF and build a Museum Reason-able View. We showed how federated results to SPARQL queries using predicates from multiple ontologies can be obtained. Consequently, we demonstrated how templates are automatically obtained in GF to generate the query results in natural language.

Future work includes extending the museum data in the Museum Reason-able View, running several queries, and increasing the coverage of the GF grammar. We intend to have a grammatical coverage for at least five languages. Other directions for future work, also include fluent discourse generation from the ontology axioms, as well as paraphrasing of the existing patterns for verbalization.

Part IV

FrameNet in the context of the Semantic Web and multilingual natural language generation

12 Applying semantic frame theory to the SW

Dannélls, Dana 2010b. Applying semantic frame theory to automate natural language templates generation from ontology statements. *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)*, 179–184. Dublin: ACL.

12.1 Introduction

Existing open-source multilingual natural language generators such as NaturalOWL (Galanis and Androutsopoulos 2007) and MPIRO (Isard et al. 2003) require a large amount of manual linguistic input to map ontology statements onto semantic and syntactic structures, as exemplified in Table 12.1. In this table, each statement contains a property and two instances; each template contains the lexicalized, reflected property and the two ontology classes (capitalized) the statement's instances belong to.

Consider adapting such systems to museum visitors in multilingual environments: as each statement is packaged into a sentence through a fixed sentence template, where lexical items, style of reference and linguistic morphology have already been determined, this adaptation process requires an extensive amount of manual input for each language, which is a labour-intensive task.

One way to automate this natural language mapping process, avoiding manual work is through language-specific resources that provide semantic and syntactic phrase specifications that are, for example, presented by means of lexicalized frames. An example of such a resource in which frame principles have been applied to the description and the analysis of lexical entries from a variety of semantic domains is the Berkeley FrameNet (FN) project (Fillmore, Johnson and Petruck 2003). The outcome of the English FN has formed the basis for the develop-

Table 12.1: MPIRO ontology statements and their corresponding sentence templates.

Ontology statement	**Sentence template**
painted-by (ex14, p-Kleo)	VESSEL *was decorated by* PAINTER
exhibit-depicts (ex12, en914)	PORTRAIT *depicts* EXHIBIT-STORY
current-location (ex11, wag-mus)	COIN *is currently displayed in* MUSEUM

ment of more sophisticated and computationally oriented multilingual FrameNets that today are freely available (Boas 2009).

This rapid development in computational lexicography circles has produced a growing number of framenet-like resources that we argue are relevant for natural language generators. We claim that semantic and syntactic information, such as that provided in a FrameNet, facilitates mapping of ontology statements to natural language. In this paper we describe the kind of information which is offered by modern computational lexical resources and discuss how template-based natural language generation (NLG) systems can benefit from them.

12.1.1 Semantic frames

A frame, according to Fillmore's frame semantics, describes the meaning of lexical units with reference to a structured background that motivates the conceptual roles they encode. Conceptual roles are represented with a set of slots called frame elements (FEs). A semantic frame carries information about the different syntactic realizations of the frame elements (syntactic valency), and about their semantic characteristics (semantic valency).

A frame can be described with the help of two types of frame elements that are classified in terms of how central they are to a particular frame, namely: core and peripheral. A core element is one that instantiates a conceptually necessary component of a frame while making the frame unique and different from other frames. A peripheral element does not uniquely characterize a frame and can be instantiated in any semantically appropriate frame.

12.1.2 The language generation module

The kind of language generation system discussed here consists of a language generation module that is guided by linguistic principles to map its non-linguistic input (i.e. a set of logical statements) to syntactic and semantic templates. This kind of generation system follows the approaches that have been discussed elsewhere (Reiter 1999; Busemann and Horacek 1998; Geldof and van de Velde 1997; Reiter and Mellish 1993).

The goal of the proposed module is to associate an ontology statement with relevant syntactic and semantic specifications. This generation process should be carried out during microplanning (cf. Reiter and Dale (2000)) before aggregation and referring expression generation take place.

12.1.3 The knowledge representation

The knowledge representation which serves as the input to the language generator is a structured ontology specified in the Web Ontology Language (OWL) (W3C 2009) on which programs can perform logical reasoning over data.

Ontological knowledge represented in OWL contains a hierarchical description of classes (concepts) and properties (relations) in a domain. It may also contain instances that are associated with particular classes, and assertions (axioms), which allow reasoning about them. Generating linguistic output from this originally non-linguistic input requires instantiations of the ontology content, i.e. concepts, properties and instances by lexical units.

12.2 From ontology statements to template specifications

Our approach to automatic template generation from ontology statements has three major steps: (1) determining the *base lexeme* of a statement's property and identifying the frame it evokes,[89] (2) matching the statement's associated concepts with the frame elements, and (3) extracting the syntactic patterns that are linked to each frame element.

[89] Base lexemes become words after they are subjected to morphological processing which is guided by the syntactic context.

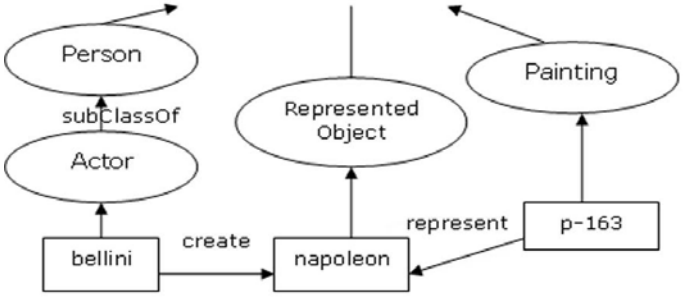


Figure 21: A fragment of the ontology.

The remainder of this section describes how base lexemes are chosen and how information about the syntactic and semantic distribution of the lexemes underlying an ontological statement are acquired.

12.2.1 Lexical units' determination and frame identification

The first, most essential step that is required for recognizing which semantic frame is associated with an ontology statement is lexicalization. Most Web ontologies contain a large amount of linguistic information that can be exploited to map the ontology content to linguistic units automatically (Mellish and Sun 2006a). However, direct verbalization of the ontology properties and concepts requires preprocessing, extensive linguistic knowledge and sophisticated disambiguation algorithms to produce accurate results. For the purposes of this paper where we are only interested in lexicalizing the ontology properties, we avoid applying automatic verbalization; instead we choose manual lexicalization.

The grammatical categories that are utilized to manifest the ontology properties are verb lexemes. These are determined according to the frame definitions and with the help of the ontology class hierarchy. For example, consider the statement *create (bellini, napoleon)*. In this domain, i.e. the cultural heritage domain, the property *create* has two possible interpretations: (1) to create a physical object which serves as the representation of the presented entity, (2) to create an artifact that is an iconic representation of an actual or imagined entity or event. FrameNet contains two frames that correspond to these two definitions, namely: *Create Representation* and *Create physical artwork*.

By following the ontological representation departing from the given instances, as illustrated in Figure 21, we learn that *bellini* is an instance

Table 12.2: Frame *Create_representation*.

Create_representation		
Def	A Creator produces a physical object which is to serve as a Representation of an actual or imagined entity or event, the Represented.	
LUs	carve.v, cast.v, draw.v, paint.v, photograph.v, sketch.v	
core FEs	Creator (C)	(1) Since [Frans]$_C$ PHOTOGRAPHED [them]$_R$ ten years ago the population has increased. (2) [Picasso]$_C$ DREW [some violent-looking birds]$_R$.
	Represented (R)	(3) When [Nadar]$_C$ PHOTOGRAPHED [her]$_R$, Desbordes-Valmore was sixty-eight. (4) [Munch]$_C$ PAINTED [himself]$_R$ as a ghost.

of the class *Actor*, *napoleon* is an instance of the class *Represented_Object*, and that *napoleon* is the represented entity in the painting *p-163*. Thus, in this context, an appropriate lexicalization of the property *create* is the verb *paint* which evokes the *Create Representation* frame.

For clarity, we specify in Table 12.2 part of the information that is coded in the frame. In this table we find the name of the frame, its definition, the set of lexical units belonging to the frame, the names of its core elements and a number of sentences annotated with these core FEs.

12.2.2 Matching the ontology concepts with frame elements

In this step, the set of core frame elements which function as the obligatory arguments of the required lexeme are matched with their corresponding ontology concepts. The algorithm that is applied to carry out this process utilizes the FE Taxonomy and the ontology class hierarchy.[90]

Matching is based on the class hierarchies. For example: *Actor*, which is a subclass of *Person* is matched with the core element *Creator*, which is a subclass of *Agent* because they are both characterized as animate objects that have human properties. Similarly, *Represented_Object*, which is a subclass of *Conceptual_Object*, is matched with the core element *Repre-*

[90]The Frame Element Taxonomy: http://www.clres.com/db/feindex.html

Table 12.3: Syntactic realizations of the lexical entry *paint*.

FEs	Syntactic Pattern
[C, R]	[[NP $_{Ext}$], [NP$_{Obj}$]]
Example 1: [Leonardo da Vinci]$_C$ *painted* [this scene]$_R$	
[R, T]	[[[NP$_{Ext}$], PP[in]$_{Dep}$]]
Example 2: [The lovely Sibyls]$_R$ *were painted* in [the last century]$_T$.	
[R, C, T]	[[NP $_{Ext}$] , [PP[by]$_{Dep}$], [PP[in]$_{Dep}$]]
Example 3: [The Gerichtsstube]$_R$ *was painted by* [Kuhn]$_C$ in [1763]$_T$.	

sented, which is a subclass of *Entity* because they are both characterized as the results of a human creation that comprises non-material products of the human mind.

This matching process leads to consistent specifications of the semantic roles specifying sentence constituents which are not bound to the input ontology structure.[91]

12.2.3 Semantic and syntactic knowledge extraction

Semantic frames, besides providing information about a lexeme's semantic content, provide information about the valency pattern associated with it, i.e. how semantic roles are realized syntactically and what are the different types of grammatical functions they may fulfill when occurring with other elements. An example of the syntactic patterns and possible realizations of the semantic elements that appear in the *Create_representation* frame (Table 12.2) are summarized in Table 12.3.[92] From this information we learn the kind of syntactic valency patterns that are associated with each semantic element. For example, we learn that in active constructions *Creator* appears in the subject position while in passive constructions it follows the preposition *by*. It can also be eliminated in passive constructions when other peripheral elements appear (Table 12.3, Example 2), in this case it is the FE *Time* (T). Although it is a peripheral element, it plays an important role in this context.

This knowledge is extracted automatically from the FN database and is converted to sentence specifications with the help of a simple

[91] One of the basic assumptions of our approach is that semantically, languages have a rather high degree of similarity, whereas syntactically they tend to differ.

[92] FN's abbreviations: Constructional Null Instantiation (CNI), External Argument (Ext), Dependent (Dep).

Perl script. Below is a template example which specifies the sentence construction of the sentence in Table 12.3, Example 3:

```
(template ( type: passive)
  (( head: |paint|) (feature:  (tense: past) )
  ( arg1 (Represented (head: |gerichtsstube|)
                          (determiner: |the|))
    arg2 (Creator (head: |kuhn|) (mod: |by|))
    arg3 (Time (head: |1763|) (mod: |in|))))
```

12.3 Testing the method

To test our approach, we employ the MPIRO domain ontology content.[93] Table 12.4 illustrates some of the results, i.e. examples of the ontology statements, the frame that matched their property lexicalization, and their possible realization patterns that were extracted from the English FrameNet.

The results demonstrate some of the advantages of the syntactic and semantic valency properties provided in FN that are relevant for expressing natural language. These include: Verb collocations, Examples (1) and (2). Intransitive usages, Example (4). Semantic focus shifts, Examples (3) and (5). Lexical variations and realizations of the same property, Examples (1), (2) and (3).

12.4 Discussion and related work

Applying frame semantics theory has been suggested before in the context of multilingual language generation (De Bleecker 2005; Stede 1996). However, to our knowledge, no generation application has tried to extract semantic frame information directly from a framenet resource and integrate the extracted information in the generation machinery. Perhaps because it is not until now that automatic processing of multilingual framenet data become available (Boas 2009). Moreover, the rapid increase of Web ontologies has only recently become acknowledged in the NLG community, who started to recognize the new needs for establishing feasible methods that facilitate generation and aggregation of natural language from these emerging standards (Mellish and Sun 2006a).

[93] <http://users.iit.demokritos.gr/~eleon/ELEONDownloads.html>

Table 12.4: Ontology statements and their possible realization patterns extracted from frames. Each instance is annotated with the three first letters of the core frame element it has been associated with.

Nr	Ontology statement	Frame	Possible realization patterns
(1)	depict (portrait$_{MED}$, story$_{ITE}$)	Communicate_ categorization	MEDIUM *depict* CATEGORY. MEDIUM *depict* ITEM *of* CATEGORY.
(2)	depict (modig$_{CRE}$, portrait$_{REP}$)	Create_physical_artwork	CREATOR *paint* REPRESENTATION. CREATOR *paint* REPRESENTATION *from* REFERENCE *in* PLACE.
(3)	depict (kuhn$_{CRE}$, flower$_{REP}$)	Create_representation	CREATOR *paint* REPRESENTED. REPRESENTED *is painted by* CREATOR *in* TIME.
(4)	locate (portrait$_{THE}$, louvre$_{LOC)}$	Being_located	THEME *is located* LOCATION.
(5)	copy (portrait$_{ORI}$, portrait$_{COP)}$	Duplication	COPY *replicate* ORIGINAL. CREATOR *replicate* ORIGINAL.

Authors who have been experimenting with NLG from Web ontologies (Bontcheva and Wilks 2004; Wilcock and Jokinen 2003) have demonstrated the usefulness of performing aggregation and applying some kind of discourse structures in the early stages of the microplanning process. As mentioned in Section 12.1.1, peripheral elements can help in deciding on how the domain information should be packed into sentences. In the next step of our work, when we proceed with aggregations and discourse generation we intend to utilize the essential information provided by these elements.

Currently, the ontology properties are lexicalized manually, a process which relies solely on the frames and the ontology class hierarchies. To increase efficiency and accuracy, additional lexical resources such as WordNet must be integrated into the system. This kind of integration has already proved feasible in the context of NLG (Jing and McKeown 1998) and has several implications for automatic lexicalization.

12.5 Conclusions

In this paper we presented on-going research on applying semantic frame theory to automate natural language template generation.

The proposed method has many advantages. First, the extracted templates and syntactic alternations provide varying degrees of complexity of linguistic entities which eliminate the need for manual input of language-specific heuristics. Second, the division of phases and the separation of the different tasks enables flexibility and re-use possibilities. This is in particular appealing for modular NLG systems. Third, it provides multilingual extension possibilities. Framenet resources offer an extended amount of semantic and syntactic phrase specifications that are only now becoming available in languages other than English. Because non-English framenets share the same type of conceptual backbone as the English FN, the steps involved in adapting the proposed method to other languages mainly concern lexicalization of the ontology properties.

Future work aims to enhance the proposed method along the lines discussed in Section 12.4 and test it on the Italian and Spanish framenets. We intend to experiment with the information about synonymous words and related terms provided in FN (which we haven't taken advantage of yet) and demonstrate how existing NLG applications that are designed to accommodate different user needs can benefit from it.

13 Exploring framenet for MLG

Dannélls, Dana and Lars Borin 2012. Toward language independent methodology for generating artwork descriptions – exploring framenet information. *EACL workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, 18–23. Avignon: ACL.

13.1 Introduction

Today museums and other cultural heritage institutions are increasingly storing object descriptions using structured information representation formats, such as semantic web domain ontologies. To make such cultural heritage content accessible to different groups and individuals in a multilingual world, this information will need to be conveyed in textual or spoken form in many languages, a language generation task which is domain specific and language dependent.

Generating multilingual natural language texts from domain specific semantic representations, such as semantic web domain ontologies, is a task which involves lexicalization and syntactic realization of the discourse relations. This paper deals with the syntactic realization problem, which is best illustrated with an example. Consider the possible formulations of the semantic relation *Create_representation* that has been lexicalized with the English verb *paint*:

1. Leonardo da Vinci *painted* this scene.
2. The lovely Sibyls *were painted* in the last century.
3. The Gerichtsstube *was painted* by Kuhn in 1763.

The syntactic structure of each sentence differs in terms of the semantic roles of the verb arguments and other constituents of the sentence. The first sentence contains the semantic roles *Creator* and *Represented*, the second sentence contains *Represented* and *Time*, and in the third sentence we find *Creator*, *Represented* and *Time*.

As the examples show there are several ways of semantically characterizing the situation expressed by a verb, with implications for the syntactic realization of that verb. When generating natural language from semantic web ontologies it is important to find generic strategies that allow us to identify the semantic elements of a verb and associate them with the appropriate argument realization of that verb. This is particularly relevant in multilingual settings because the semantic and syntactic behavior of verbs will vary depending on the target language, both in the constructions found and in their distribution.

Previous work on natural language generation of cultural heritage information from semantic web ontologies has relied on a large amount of specially tailored manual linguistic information to produce descriptions that are targeted to a specific group of readers (Androutsopoulos et al. 2001; Dannélls 2008a; Konstantopoulos, Karkaletsis and Bilidas. 2009). Although valuable information for generating natural languages is found in computational lexical-semantic resources such as the Berkeley FrameNet (Section 13.3) which exist today in several languages (Erk et al. 2003; Subirats and Petruck 2003; Ohara et al. 2003; Borin et al. 2010), there has been little emphasis on how to manage digitized data from digital libraries using these open source resources. In this paper we demonstrate how the information available in such electronically available resources can be exploited for generating multilingual artwork descriptions.

In the remainder of this paper we describe a case study on English and Swedish that underscores the importance of using a lexical resource such as a framenet (Section 13.2). We present the kind of information that is offered by two existing framenets (Section 13.3). We demonstrate how a domain specific natural language generator can benefit from the information that is available in both framenets (Section 13.4). We end with a discussion and pointers to future work (Section 13.5).

13.2 Data collection and text analysis

13.2.1 Corpus data

To identify the semantic and syntactic constructions that characterize object descriptions in the cultural heritage domain, we have collected parallel texts from Wikipedia in two languages: English and Swedish. In total, we analyzed 40 parallel texts that are available under the cate-

gory *Painting*. Additionally, we selected object descriptions from digital libraries that are available through online museum databases. The majority of the Swedish descriptions were taken from the World Culture Museum,[94] the majority of the English descriptions were collected from the Met Museum.[95]

13.2.2 Semantic analysis

The strategy we employed to analyze the texts follows the approach presented by McKeown (1985) on how to formalize principles of discourse for use in a computational process. Seven frame elements have been examined, these include: *Location* (L), *Creator* (CR), *Representation* (RE), *Represented* (R), *Descriptor* (D), *Time* (TI), *Type* (T). The text analysis has shown that the following combinations of these major frame elements are the most common:

1. RE, T, CR, TI, L, D, R
2. RE, T, CR, R, TI, L, D
3. RE, TI, T, CR, D, L, R
4. RE, TI, CR, D, R, L

The listed semantic combinations reflect the word order that we have found in the text analysis for the two languages. However, since many of the analyzed sentences that begin with the object in focus (the *Representation*) appear in the passive voice, i.e, *was painted by, was created by*, the word order of these combinations may vary. Furthermore, not all of the listed semantic elements are mandatory in the object descriptions. For example, although corresponding to the first combination of semantic elements, the sentence *De Hooch probably painted this picture in the early 1660s* only contains the frame elements CR, RE and TI.

13.2.3 Syntactic analysis

The texts have been syntactically annotated using the Maltparser (Nivre et al. 2007). Figure 22 shows two example sentences converted to constituent trees.

This small example shows that there is a difference in how syntactic trees are built for each language. While in the English sentence the verb *was painted* is followed by a preposition phrase (PP), the Swedish verb *målades* (the passive form of 'paint') is followed by a cardinal number without a preposition (which could be analyzed as an NP).

[94] <http://collections.smvk.se/pls/vkm/rigby.welcome>
[95] <http://www.metmuseum.org>

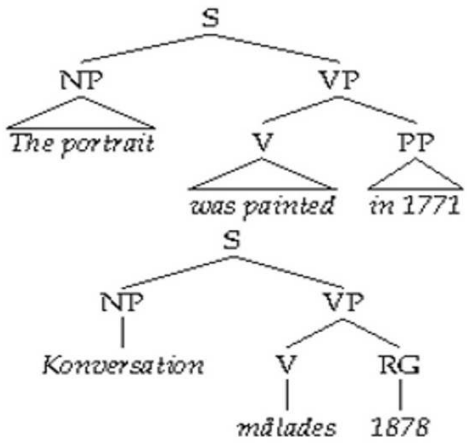


Figure 22: Parse trees for two example sentences.

13.3 Framenets

13.3.1 The Berkeley FrameNet

The Berkeley FrameNet (BFN)[96] (Fillmore, Johnson and Petruck 2003) is an electronic lexical resource based on the notion of Frame Semantics (Fillmore 1985); we know the meaning of a word through prototypical situations (scenarios) in which the word (called a lexical unit, LU) occurs. A frame can be described with the help of two types of frame elements (FEs) that are classified in terms of how central they are to a particular frame. A *core element* is one that instantiates a conceptually necessary component of a frame while making the frame unique and different from other frames. On the other hand, a *peripheral element* does not uniquely characterize a frame and can be instantiated in any semantically appropriate frame.

For example, table 13.1 describes the lexical units and the frame elements appearing in the frame *Create_representation*, which has the following definition (from the BFN website):

> A *Creator* produces a physical object which is to serve as a *Representation* of an actual or imagined entity or event, the *Represented*.

[96] http://framenet.icsi.berkeley.edu/

Table 13.1: LUs and FEs in the frame *Create_representation* in BFN.

Create_representation		
LUs	carve.v, cast.v, draw.v, paint.v, photograph.v, sketch.v	
FEs	Core	Creator (C), Represented (R)
	Peripheral	Depictive (D), Depictive_of_represented (DR), Means (ME), Instrument (IN), Iteration (I), Material (MA), Manner (M), Place (P), Purpose (PU), Representation (RE), Role (RO), Time (T)

Each lexical unit appearing in the frame carries information about its related frame elements (semantic valency) and their syntactic realizations (syntactic valency). Examples of the valency patterns that are found for the verb *paint* are listed in table 13.2.[97]

Examples of sentences that can be formed with these semantic and syntactic representations are:

1. The Gerichtsstube was painted by Kuhn in 1763.
2. The youngest girl had her portrait painted by him .
3. He painted her at least fourteen times.

Table 13.2: FEs and their syntactic realizations found in the *Create representation* frame for the verb *paint*.

Creator (CR)	Represented (R)	Time (TI)
NP.Ext	NP.Obj	PP[at].Dep
PP[by].Dep	NP.Ext	PP[in].Dep

[97] The abbreviations in table 13.2 and table 13.4 follow the BFN annotation scheme: Dependent (Dep), External Argument (Ext), Object (Obj), Constructional null instantiation (CNI).

13.3.2 The Swedish FrameNet

BFN has formed the basis for the development of computationally oriented freely available framenets for a number of languages (Boas 2009), among these the Swedish FrameNet (SweFN) (Borin et al. 2010).[98]

SweFN takes its conceptual backbone from BFN, i.e., the core and peripheral elements are exactly the same for frames appearing in both framenets. Each frame also contains semantically annotated example sentences from which we can extract syntactic information. The most notable differences between the frames can be seen from a comparison of table 13.1 and table 13.3.

The lexical units in each SweFN frame are linked to the Swedish lexical-semantic resource SALDO (Borin, Forsberg and Lönngren 2008). SweFN is also organized into a domain hierarchy, with a general domain and at present the two specialized domains *Art* and *Medicine*. In addition, each frame in SweFN is associated with a semantic type and a list of compounds instantiating part of a frame configuration.

Syntactic valency information is obtained from the Swedish Simple and Parole lexicons (Lenci et al. 2000). The encoding of this valency information is different from the one provided in BFN. For example, for the verb *avbilda* 'depict' we find the following syntactic valency:

```
S_NP_A/x [vb] DO_NP_B/y
```

S denotes the subject of the sentence, *DO* denotes direct object. Both are realized as either animate (*A, B*) or inanimate (*x, y*) NPs.

In addition, it is possible to extract almost the same information about semantic and syntactic valency from the example sentences for the verb *avbilda* (Table 13.4). It is important to note that the syntactic annotation in SweFN does not follow the BFN model, although we use the same annotation scheme here to facilitate comparison.

Examples of sentences that can be formed using the semantic and syntactic representations listed in table 13.4 are:

1. Det förra århundradet hade han avbildat konstnärinnan Anna Maria Ehrenstrahl.
   'The previous century had he depicted the-female-artist Anna Maria Ehrenstrahl.'
2. Här avbildas Gustav Adolf.
   'Here is-depicted Gustav Adolf.'

[98] http://spraakbanken.gu.se/swefn/

Table 13.3: LUs and FEs in the frame *Create_representation* in SweFN.

Create_representation	
LUs	vb: avbilda..1, avporträttera..1, filma..1, fotografera..1, knäppa..5, plåta..1, porträttera..1, skissa..1, skissera..1, skulptera..1;; vbm: måla_av..1;; nn: framställning..1, teckning..1, pennteckning..1, skiss..1, skämtteckning..1, tuschteckning..1, frihandsteckning..1
Domain	Gen/Art
Sem Type	Symbolic_creation
Compound	Manner+LU, Representation+LU

Table 13.4: FEs and their syntactic realizations found in the *Create representation* frame for the verb *avbilda* 'depict'.

Creator (CR)	Represented (R)	Time (TI)
NP.Ext	NP.Obj	AVP.Dep
CNI	NP.Ext	

13.4 Multilingual language generation of museum object descriptions

13.4.1 The language generator tool

We have developed a domain specific grammar application to generate multilingual artwork descriptions from domain specific ontologies. The application is developed in the Grammatical Framework (GF) (Ranta 2004). The key feature of GF is the distinction between an abstract syntax, which acts as a semantic interlingua, and concrete syntaxes, representing linearizations in various target languages, natural or formal. The grammar comes with a resource library which aids the development of new grammars for specific domains by providing syntactic op-

erations for basic grammatical constructions (Ranta 2009).
The information available in BFN and SweFN on semantic elements and their possible syntactic realizations with specific lexical units has guided the (manual) development of the generation grammars. Below we present the abstract and the concrete grammars of English and Swedish for the semantic elements RE, CR, TI and R.

In the abstract grammar we have a list of discourse patterns (DPs), encoded as functions that specify the semantic roles appearing in the pattern.

```
DP1: representation creator time
DP2: creator represented time
```

In the concrete grammars, patterns are linearized differently for each language. Semantic elements listed in each DP are expressed linguistically with the resource grammar constructors. In the examples below we find six of the GF constructors: `mkPhr` (Phrase), `mkS` (Sentence), `mkCl` (Clause), `mkNP` (Noun Phrase), `mkVP` (Verb Phrase), `mkAdv` (Verb Phrase modifying adverb). The lexicons which we use to lexicalize the verbs and the semantic elements are the OALD for English and SALDO for Swedish.

```
DP1
representation creator time =
str : Phr =  mkPhr
(mkS pastTense (mkCl (mkNP representation)
(mkVP (mkVP (passiveVP paint_V2)
(mkAdv by8agent_Prep (mkNP creator))
(mkAdv in_Prep (mkNP time))))));

DP1
representation creator time =
str :  Phr = mkPhr
(mkS pastTense
(mkCl (mkNP representation)
(mkVP (mkVP (passiveVP maala_vb_1)
(mkAdv by8agent_Prep (mkNP creator))
(mkAdv noPrep (mkNP time))))));
```

When used for generating sentences, the above grammatical representations will yield syntactic trees with the structures exemplified in figure 22 above.

13.4.2 Linguistic realisations from framenets

The advantage of the implementation strategy presented in section 13.4.1 is that we can build different syntactic trees for each language to form a description regardless of the order of the semantic elements.

Let us consider the lexical-semantic information provided in tables 13.2 and 13.4. This information could be embedded in the application grammar to compute the following linguistic specifications.

```
DP2
creator represented time =
str : Phr = mkPhr (mkS
(mkCl  (mkNP represented)
(mkVP (mkVP (mkVP  paint_V2))
(mkAdv by8agent_Prep (mkNP creator))
(mkAdv in_Prep (mkNP time)))));

DP2
creator represented time =
 str : Phr = mkPhr (mkS
(mkCl (mkNP creator)
(mkVP (mkVP avbilda_vb_1_1_V)
(mkNP (mkCN represented
(mkAdv noPrep (mkNP time)))))));
```

These specifications can in turn be used to generate sentences like the following:

1. [Captain Frans Banning Cocq]$_R$ painted [by Rembrandt van Rijn]$_{CR}$ [in 1642]$_{TI}$.
2. [Rembrandt van Rijn]$_{CR}$ har avbildat [Kapten Frans Banning Cocq]$_R$ [1642]$_{TI}$.
   'Rembrandt van Rijn has depicted Captain Frans Banning Cocq 1642.'

The discourse patterns can be automatically modified to compute a variety of linguistic specifications that are acquired from lexical-semantic frames.

13.5 Summary

This paper has demonstrated the differences in the syntactic realization of verbs in two languages. We described what kind of semantic and syntactic valency can be obtained from the information given in two framenets to improve syntactic realizations of object descriptions from particular sets of semantic elements.

The cultural heritage domain is a potential application area of a framenet, which we argue is an essential open source resource for generating multilingual object descriptions. We believe it is possible to establish more efficient processing if the framenet is domain-specific and thereby offers linguistic structures that are specific to the domain, in our case the art domain. Even though our generation grammars at the moment have been manually constructed using the framenet information, we hope that we have shown the utility of being able to draw on a framenet in developing such applications. The next logical step will be to attempt to generate (partial) grammars automatically from the framenet information directly. We also intend to increase the grammars to handle a larger set of semantic frames.

Part V

Coherent multilingual generation from the SW

14 ONLINE MLG FROM THE SEMANTIC WEB

Dannélls, Dana, Ramona Enache, Damova Mariana and Milen Chechev 2012. Multilingual online generation from semantic web ontologies. *Proceedings of the world wide web conference (WWW2012), European project track*, 239–242.

14.1 Introduction

The work described in this paper is developed within the Multilingual Online Translation (MOLTO) project.[99] More specifically, we present workpackage 8 (WP8): Case Study: Cultural Heritage. The objective of this workpackage is to build an ontology-based multilingual grammar for museum information using natural language generation technologies.

We have developed a Web application that applies natural language generation techniques to generate multilingual descriptions about museum objects from ontologies. Our approach is to utilize discourse structures that capture how concepts and relationships are realized linguistically. We have been experimenting with museum data to test our approach and find that it performs well for the examined languages.

The remainder of this document presents the motivation and the goals of our workpackage (Section 14.2). We describe the knowledge representation framework (Section 14.3). In section 14.4, we describe the grammar implementation and present some generation results. We end with conclusions and directions for future work (Section 14.5).

[99] http://www.molto-project.eu/

14.2 The motivation and goals

The general motivation of this work is the increase of Cultural Heritage (CH) information on the Semantic Web. Today there exist millions of collections and thousands of applications providing a wide range of users direct access to cultural heritage material. This has brought up a need to develop tools that are capable of searching and presenting different kinds of information to end-users in their language of preference.

The goals of our WP are to:

- build an ontology-based multilingual grammar for museum information for artefacts at Gothenburg City Museum (GCM) starting from the Conceptual Reference Model (CIDOC-CRM);
- build a prototype of a cross-language retrieval and representation system to be tested with objects in the museum, and automatically generate Wikipedia articles for museum artefacts in 5 languages;
- cover 15 languages for baseline functionality and 5 languages with a more complete coverage.

This paper describes the implementation of the prototype for retrieving and representing information about museum objects on the Web. It also describes the grammar that has been developed to automatically generate coherent object descriptions in two languages: English and Swedish.

14.3 The Museum Reason-able View

The Museum Reason-able View is an assembly of independent datasets, which are used as a single body of knowledge with respect to reasoning and query evaluation. Each data set in the Museum Reason-able View is aiming at lowering the cost and the risks of using specific linked datasets for specific purposes. This approach to linked data techniques has been discussed and implemented as a Reason-able View of the web of data (Kiryakov et al. 2009).

The Museum Reason-able View environment, described by Damova and Dannélls (2011) is built as an instance of BigOWLIM triple store (Bishop et al. 2011). It contains: DBPedia 3.6,[100] Geonames,[101] PROTON

[100] http://dbpedia.org/

[101] Geonames website: http://www.geonames.org/

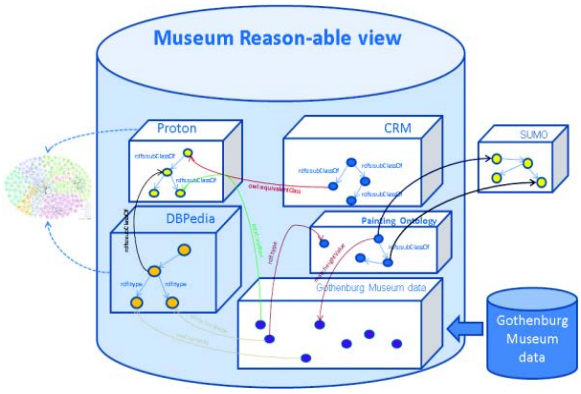


Figure 23: The Museum Reasonable View.

(Terziev et al. 2005), CIDOC-CRM (Crofts et al. 2009),[102] the painting ontology (Dannélls 2012a),[103] their mappings, and the triplified Gothenburg City Museum data (Dannélls et al. 2011).

Figure 23 shows the architecture of the Museum Reason-able View, which includes interconnected schemata and links to external datasets of the Gothenburg City museum data, such as the entire DBpedia. The Museum Reason-able View contains 245,365,883 explicit statements and 70,704,053 entities of which close to 10 thousand are museum artifacts from the Gothenburg city museum database.

14.3.1 Integrating museum data

Integrating datasets into linked data in RDF usually takes place by indicating that two instances from two datasets are the same by using the built in Web Ontology Language (OWL) predicate: `owl:sameAs`.[104] However, recent research (Damova 2011; Damova et al. 2011; Jain et al. 2011) has shown that interlinking the models according to which the datasets are described is a more powerful mechanism of dealing with large amounts of data in RDF, as it exploits inference and class assignment.

We have adopted this approach when creating the infrastructure for the museum linked data, including several layers of upper-level on-

[102] http://www.cidoc-crm.org/
[103] http://spraakdata.gu.se/svedd/painting-ontology/painting.owl
[104] http://www.w3.org/TR/owl-ref/

tologies. They provide a connection to different sets of linked data, for example PROTON for the Linked Open Data (LOD) cloud (Bizer et al. 2011). They also provide an extended pool of concepts that can be referred to in museum linked data that do not directly pertain to the expert descriptions of the museum objects, and the strictly expert museum knowledge is left to CIDOC-CRM. This model of interlinked ontologies offers a flexible access to the data with different conceptual access points.

14.3.2 Accessing museum linked data

The data in the Museum Reason-able View is accessible via SPARQL (Eric and Andy 2008) end-point and keywords.[105] The queries can be formulated by combining predicates from different datasets and ontologies in a single SPARQL query, retrieving results from all different datasets that are part of the Reason-able View.

A query example about museum objects from Swedish museums is given below.

```
select ?museumObject ?museum where {
 ?museumObject
 core:P109_has_current_or_former_curator ?museum .
 ?museum ptop:locatedIn ?location .
 ?location ptop:subRegionOf dbpedia:Sweden }
```

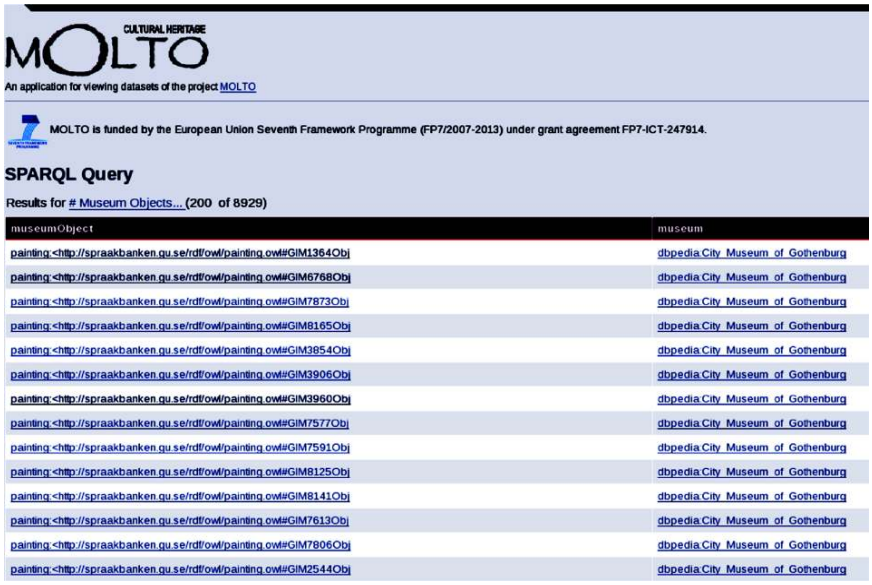
The above query returns the results that are depicted in figure 24. Note that the returned location is the DBpedia resource about the city of Gothenburg.

Other queries can be asked about the types of artwork preserved in the museum, their material, or about artwork from a certain period of time, etc.

14.4 Natural language generation

The grammar formalism utilized for generating natural language descriptions from semantic web ontologies is the Grammatical Framework GF (Ranta 2004). It is a grammar formalism, based on Martin-Löf's type theory (Martin-Löf 1982). The key feature of the grammar

[105] The data is available at: http://museum.ontotext.com


CULTURAL HERITAGE
MOLTO
An application for viewing datasets of the project MOLTO

MOLTO is funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement FP7-ICT-247914.

SPARQL Query

Results for # Museum Objects... (200 of 8929)

museumObject	museum
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM1364Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM6768Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM7873Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM8165Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM3854Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM3906Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM3960Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM7577Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM7591Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM8125Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM8141Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM7613Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM7806Obj	dbpedia:City_Museum_of_Gothenburg
painting:<http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM2544Obj	dbpedia:City_Museum_of_Gothenburg

Figure 24: Query results about museum objects from Swedish museums.

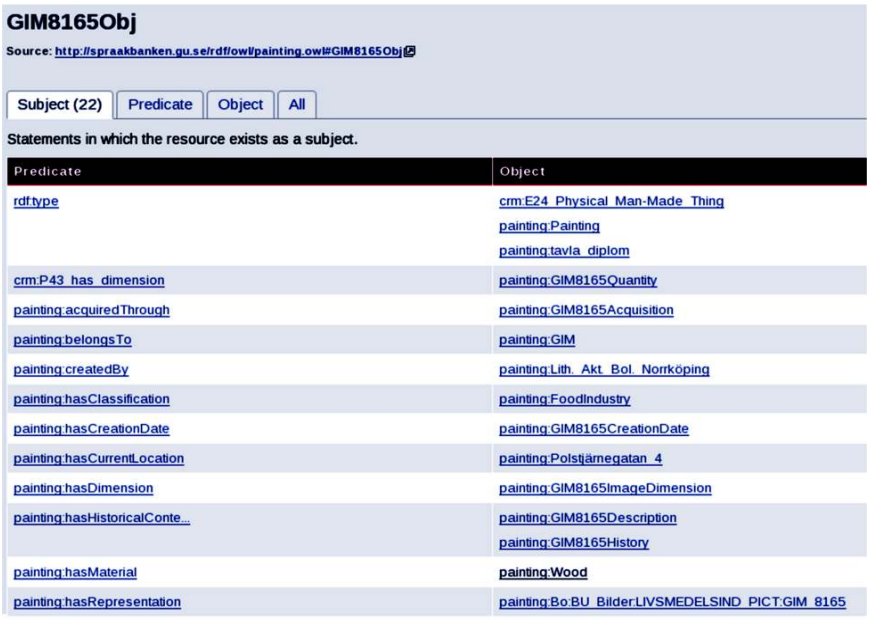
is the division of an abstract syntax, which acts as a semantic interlingua and concrete syntaxes, representing linearizations in various target languages (natural or formal).

GF comes with a resource library (Ranta 2009), covering the syntax of more than 20 languages.[106] The resource library aids the development of new grammars for specific domains by providing the operations for basic grammatical constructions, and thus making it possible for users without linguistic background to generate syntactically correct natural language.

14.4.1 Translation of the Museum Reason-able View to GF

The output result from BigOWLIM is a set of triples in the form of Resource Description Framework (RDF) statements consisting of constructs of the shape `<subject, predicate, object>` describing a resource. Each resource in the museum reason-able view is linked to its corresponding lexical unit in the GF lexicon. For example, GIM8165Obj is defined as a painting object in the abstract syntax and it is linearized as a person name with its title in the English dictionary:

[106] www.grammaticalframework.com



GIM8165Obj

Source: http://spraakbanken.gu.se/rdf/owl/painting.owl#GIM8165Obj

Subject (22) | Predicate | Object | All

Statements in which the resource exists as a subject.

Predicate	Object
rdf:type	crm:E24_Physical_Man-Made_Thing painting:Painting painting:tavla_diplom
crm:P43_has_dimension	painting:GIM8165Quantity
painting:acquiredThrough	painting:GIM8165Acquisition
painting:belongsTo	painting:GIM
painting:createdBy	painting:Lith._Akt._Bol._Norrköping
painting:hasClassification	painting:FoodIndustry
painting:hasCreationDate	painting:GIM8165CreationDate
painting:hasCurrentLocation	painting:Polstjärnegatan_4
painting:hasDimension	painting:GIM8165ImageDimension
painting:hasHistoricalConte...	painting:GIM8165Description painting:GIM8165History
painting:hasMaterial	painting:Wood
painting:hasRepresentation	painting:Bo:BU_Bilder:LIVSMEDELSIND_PICT:GIM_8165

Figure 25: A set of triples describing the painting object GIM8165Obj.

```
GIM8165Obj : PPainting ;
GIM8165Obj = mkPN "Kliché";
```

Some lexical units such as paper, wood, etc. are already available in existing lexicons that have been imported to GF. Two of the lexicons that we are currently utilizing are the Oxford dictionary for English and the Swedish Association Lexicon (SALDO) (Borin, Forsberg and Lönngren 2008), which is also available in LMF (Francopoulo et al. 2006), for Swedish.

Painting resources are encoded in the abstract grammar as a sequence of semantic categories instead of a set of statements. For example the description of the GIM8165Obj depicted in figure 25 has the following semantic representation in GF.

```
fun GIM8165ObjDescription : PaintingDescription
  GIM8165Obj LithAktBol Y1916 NoMuseum GIM
  NoColour NoSize GIM8165ObjRepresented Wood ;
```

In the above example, the function *PaintingDescription* contains the following semantic concepts: painting, painter, year, museum, collection, colour, size, represented, and material. We should note that the

retrieved information from the SPARQL query (Figure 25) contains additional semantic concepts that are not covered by the discourse patterns yet.

14.4.2 Discourse structures

Through linguistic analysis we have observed how the domain representation is encoded in a large set of well-formed object descriptions. We then followed the discourse structure to learn how the ontology statements are composed in English and Swedish (Dannélls 2011). Below we summarize some of the discourse patterns and the semantic concepts presented as functions in the GF abstract grammar.

- DP0 : painting painter year -> Text
- DP1 : painting museum painter size -> Text
- DP2 : painting painter represented museum -> Text
- DP3 : painting material year painter -> Text
- DP4 : painting painter year museum colour size -> Text

The discourse patterns are manually encoded in the application's abstract grammar. By optimizing the grammar we are able to generate several examples for each description.

```
def GenDP4 NoPainting _ _ _ _ _ _ _ _ _ = noText ;
def GenDP4 _  NoPainter _ _ _ _ _ _ _ _ = noText ;
def GenDP4 _ _  NoYear _ _ _ _ _ _ _     = noText ;
def GenDP4 painting painter year _  _  _  _  _  _ _
                     = DP0 painting painter year ;
```

The basic idea behind the above implementation rules is that although there are several semantic concepts available for a certain object we can match its description with simpler patterns containing fewer semantic concepts.

14.4.3 Generation results

Using the above discourse pattern constructions we are able to generate the following descriptions:

- (DP1-eng) Sommer Joy was painted in 1886. It measures 349 by 776 cm.
- (DP1-swe) Sommarnöje blev målad år 1886. Den är av storlek 349 och 776 cm.
- (DP2-eng) Sommer Joy is a painting made by Anders Zorn. The work depicts a view from Lilla Bommen at Hisingen.
- (DP2-swe) Sommarnöje är en målning av Anders Zorn.
  Den föreställer en utsikt från Lilla Bommen mot Hisingen.
- (DP3-eng) Sommer Joy is painted on paper in 1886 by Anders Zorn.
- (DP3-swe) Sommarnöje blev målad på papper år 1886 av Anders Zorn.
- (DP4-eng) Sommer Joy was painted by Anders Zorn in the year 1886. It is of size 349 by 776 cm and is painted on paper. The painting is displayed at the Museum of World Culture.
- (DP4-swe) Sommarnöje blev målad av Anders Zorn år 1865. Den är av storlek 349 och 776 cm och är målad på paper. Målningen återfinns på Världskulturmuseet.
- (DP4-eng) Sommer Joy was painted by Anders Zorn.
- (DP4-swe) Sommarnöje blev målad av Anders Zorn.

14.5 Summary and future work

In this paper we present a prototype developed in the context of the MOTLO project. We outline the entire infrastructure of the Museum Reason-able View and show its connection to existing infrastructures such as Dbpedia. We present the multilingual grammar application that is being developed to generate multilingual museum object descriptions from the described resources and demonstrate how the generation results are obtained.

This work is about an automatic work-flow of sharing data infrastructures that is explicitly targeted towards the Semantic Web. The primary goal of this effort is to support question answering and automatically generate short Wikipedia-like articles for museum artifacts in 5 languages with extensive coverage. We are currently extending the grammar to support more patterns and more languages including Finnish, French and German. The generation results will be evaluated using native speakers of the language.

15 On generating coherent descriptions

Dannélls, Dana 2012b. On generating coherent multilingual descriptions of museum objects from semantic web ontologies. *Proceedings of the Seventh International Natural Language Generation Conference (INLG 2012)*, 76–84. Utica, IL: ACL.

15.1 Introduction

During the last decade, there has been a shift from developing natural language generation systems to developing generic systems that are capable of producing natural language descriptions directly from Web ontologies (Schwitter and Tilbrook 2004; Fuchs, Kaljurand and Kuhn 2008; Williams, Third and Power 2011). These systems employ controlled language mechanisms and Natural Language Generation (NLG)

Table 15.1: A natural language description generated from a set of ontology statements.

createdBy (Guernica, PabloPicasso)
currentLocation (Guernica, MuseoReinaSofía)
hasColor (Guernica, White)
hasColor (Guernica, Gray)
hasColor (Guernica, Black)
Guernica is created by Pablo Picasso. **Guernica** has as current location the Museo Reina Sofía. **Guernica** has as color White, Gray and Black.

Table 15.2: A museum object description generated in English and Swedish.

Guernica is created by Pablo Picasso.
It has as current location the Museo Reina Sofía.
It has as color White, Gray and Black.
Guernica målades av Pablo Picasso.
Den finns på Museo Reina Sofía.
Den är målad i vitt, svart och grått.

technologies such as discourse structures and simple aggregation methods to verbalise Web ontology statements, as exemplified in table 15.1.

If we want to adapt such systems to the generation of coherent multilingual object descriptions, at least three language dependent problems must be faced, viz. lexicalisation, aggregation and generation of referring expressions. The ontology itself may contain the lexical information needed to generate natural language (McCrae et al. 2012) but it may not carry any information either about the aggregation of semantic concepts or the generation of a coherent discourse from referring expressions. Halliday and Hasan (Halliday and Hasan 1976), and other well known theories such as Centering Theory (Grosz, Weinstein and Joshi 1995), propose establishing a coherent description by replacing the entity referring to the Main Subject Reference (MSR) with a pronoun – a replacement which might result in simple descriptions such as illustrated in table 15.2. Although these descriptions are coherent, i.e. they have a connectedness that contributes to the reader's understanding of the text, they are considered non-idiomatic and undeveloped by many readers because of consecutive pronouns – a usage which in this particular context is unacceptable.

Since previous theories do not specify the types of linguistic expressions different entities may bear in different languages or domains, there remain many open questions that need to be addressed. The question addressed here is the choice of referential forms to replace a sequence of pronouns, which makes the discourse coherent in different languages. Our claim is that different languages use different linguistic expressions when referring to a discourse entity depending on the semantic context. Hence a natural language generator must employ language dependent co-referential strategies to produce coherent descriptions. This claim is based on cross-linguistic investigations into how

coreference is expressed, depending on the target language and the domain (Givón 1983; Hein 1989; Ariel 1990; Prince 1992; Vallduví and Engdahl 1996).

In this paper we present a contrasting study conducted in English, Swedish and Hebrew to learn how coreference is expressed. The study was carried out in the domain of art, more specifically focusing on naturally-occurring museum object descriptions. As a result of the study, strategies for generating coreference in three languages are suggested. We show how these strategies are captured in a grammar developed in the Grammatical Framework (GF).[107] We evaluated our method by experimenting with lexicalised semantic web ontology statements which were structured according to particular organizing principles. The result of the evaluation shows language-dependent coreference strategies lead to better generation results.

15.2 Related work

Also Prasad (Prasad 2003) employed a corpus-based methodology to study the usage of referring expressions. Based on the results of the analysis, he developed an algorithm to generate referential chains in Hindi. Other algorithms for characterizing referential expressions based on corpus studies have been proposed and implemented in Japanese (Walker, Cote and Iida 1996), Italian (Di Eugenio 1998), Catalan and Spanish (Potau 2008), and Romanian (Harabagiu and Maiorano 2000).

Although there has been computational work related to Centering for generating a coherent text (Power and Scott 2000; Barzilay and Lee 2004; Karamanis et al. 2009), we are not aware of any methodology or NLG system that employs ontologies to guide the generation of referential chains depending on the language considered.

15.3 Data collection, annotations and analysis

15.3.1 Material

To study the domain-specific conventions and the ways of signalling linguistic content in English, Swedish and Hebrew, we collected object descriptions written by native speakers of each language from digital

[107] http://www.grammaticalframework.org/

Table 15.3: Statistics of the text collections.

Number of	Eng.	Swe.	Heb.
Descriptions	394	386	110
Tokens	42792	27142	5690
Sentences	1877	2214	445
Tokens/sentence	24	13	13
Sentences/description	5	6	4

libraries that are available through on-line museum databases. The majority of the Swedish descriptions were taken from the World Culture Museum.[108] The majority of the English descriptions were collected from the Metropolitan Museum.[109] The majority of the Hebrew descriptions were taken from Artchive.[110] Table 15.3 gives an overview of the three text collections. In addition, we extracted 40 parallel texts that are available under the sub-domain *Painting* from Wikipedia.

15.3.2 Syntactic annotation

All sentences in the reference material were tokenised, part-of-speech tagged, lemmatized, and parsed using open-source software. We used Hunpos, an open-source Hidden Markov Model (HMM) tagger (Halácsy, Kornai and Oravecz 2007) and Maltparser, version 1.4 (Nivre et al. 2007). The English model for tagging was downloaded from the Hunpos web page.[111] The model for Swedish was trained on the Stockholm Umeå Corpus (SUC) and is available to download from the Swedish Language Bank web page.[112] The Hebrew tagger and parsing models are described in Goldberg and Elhadad (Goldberg and Elhadad 2010).

15.3.3 Semantic annotation

The texts were semantically annotated by the author. The annotation schema for the semantic annotation is taken from the CIDOC Concep-

[108] http://collections.smvk.se/pls/vkm/rigby.welcome
[109] http://www.metmuseum.org
[110] http://www.artchive.com/
[111] http://code.google.com/p/hunpos/downloads/list
[112] http://spraakbanken.gu.se/

Table 15.4: The semantic concepts for annotation.

Actor	Man-Made_Object
Actor Appellation	Material
Collection	Place
Dimension	Time-span
Legal Body	Title

tual Reference Model (CRM) (Crofts et al. 2009).[113] Ten of the CIDOC-CRM concepts were employed to annotate the data semantically. These are given in table 15.4.

15.3.4 Referential expressions annotation

The task of identifying referential instances of a painting entity, which is our main subject reference, requires a meaningful semantic definition of the concept *Man-Made Object*. Such a fine-grained semantic definition is available in the ontology of paintings (Dannélls 2012a),[114] which was developed in the Web Ontology Language (OWL) to allow expressing useful descriptions of paintings.[115] The ontology contains specific concepts of painting types, examples of the hierarchy of concepts that are specified in the ontology are listed below.

subClassOf(Artwork, E22_Man-Made_Object)

subClassOf(Painting, Artwork)

subClassOf(PortraitPainting, Painting and
depicts(Painting, AnimateThing))

subClassOf(OilPainting, Painting and
hasMaterial(Painting, OilPaint))

When analysing the corpus-data, we look closer at two linguistic forms of reference expressions: definite noun phrases and pronouns, focusing on three semantic relations: direct hyperonym (for example *Painting* is direct hyperonym of *Portrait Painting*), higher hyperonym (for example, both *Artwork* and *Man-Made Object* are higher hyperonyms of *Portrait Painting*) and synonym, i.e. two different linguistic units of reference expressions belonging to the same concept.

[113] http://cidoc.ics.forth.gr/
[114] http://spraakdata.gu.se/svedd/painting-ontology/painting.owl
[115] http://www.w3.org/TR/owl-features/

15.3.5 Data analysis and results

The analysis consisted of two phases: (1) analyse the texts for discourse patterns, and (2) analyse the texts for patterns of coreference in the discourse.

15.3.5.0.1 Discourse patterns

A discourse pattern (DP) is an approach to text structuring through which particular organizing principles of the texts are defined through linguistic analysis. The approach follows McKeown (McKeown 1985) to formalize principles of discourse for use in a computational process. Following this approach, we have identified three discourse patterns for describing paintings that are common in the three languages. These are summarised below.

- **DP1** Man-Made_Object, Object-Type, Actor, Time-span, Place, Dimension
- **DP2** Man-Made_Object, Time-span, Object-Type, Actor, Dimension, Place
- **DP3** Man-Made_Object, Actor, Time-span, Dimension, Place

15.3.5.0.2 Patterns of coreference

In the analysis for coreference, we only considered entities appearing in subject positions. Below follows examples of the most common types of coreference found in the corpus-data.

As seen in (12b) and in many other examples, the first reference expressions are the definite noun phrase *the painting*, i.e. coreference is build through the direct hyperonym relation. The choice of the reference expression in the following sentence (12c) is the definite noun phrase *the work*, which is a higher hyperonym of the main subject of reference *The Old Musician*.

Sentence (13b) shows a noun is avoided; the linguistic unit of the reference expression is a pronoun preceding a conjunction, followed by an ellipsis.

(12) a. The Old Musician is an 1862 painting by French painter, Édouard Manet.

b. **The painting** shows the influence of the work of Gustave Courbet.

c. **This work** is one of Manet's largest paintings and **Ø** is now conserved at the National Gallery of Art in Washington.

(13) a. The Birth of Venus is a painting by the French artist Alexandre Cabanel.

b. **It** was painted in 1863, and **Ø** is now in the Musée d'Orsay in Paris.

In the Swedish texts we also find occurrences of pronouns in the second sentence of the discourse, as in (14b). We learn that the most common linguistic units of the reference expressions also are definite noun phrases given by the direct hyperonym relation.

(14) a. Stjärnenatten är en målning av den nederländske postimpressionistiske konstnären Vincent van Gogh från 1889.

b. Sedan 1941 har **den** varit med i den permanenta utställningen vid det moderna museet i New York.

c. **Tavlan** har allmänt hyllats som hans magnum opus och har reproducerats många gånger.

((a) The Starry Night is a painting by the dutch artist Vincent van Gogh, created in 1889. (b) Since 1941 **it** was in the permanent exhibition of the museum in New York. (c) **The picture** is widely hailed as his magnum opus and has been reproduced many times.)

Similar to English, the most common linguistic units of the reference expressions are definite noun phrases, as in (15b). However, the relation of these phrases with respect to the main subject of reference is either a direct hyperonym or a synonym, such as *tavlan* in (14c) and (16b).

(15) a. Wilhelm Tells gåta är en målning av den surrealistiske konstnären Salvador Dalí.

b. **Målningen** utfördes 1933 och **Ø** finns idag på Moderna museet i Stockholm.

((a) Wilhelm Tell's Street is a painting by the artist Salvador Dali. (b) **The painting** was completed in 1933 and today it is stored in the modern museum in Stockholm.)

(16) a. Baptisterna är en målning av Gustaf Cederström från 1886, och **Ø** föreställer baptister som samlats för att förrätt dop.

b. **Tavlan** finns att beskåda i Betel folkhögskolas lokaler.

((a) The Baptists is a painting by Gustaf Cederström from 1886, and depicts baptists that have gathered for a bad. (b) **The picture** can be seen in Betel at the people's high school premises.)

(17) a. lila 'ohavim hyno stiyor śhemen śel hasayar haholandi vincent van gogh, hametoharac lesnat 1889.

b. **hastiyor** mosag kayom bemozehon lehomanot modernit sebahir new york.

c. **ho exad hastiyorim** hayedoyim beyoter sel van gogh.

((a) The Starry Night is an oil painting by the dutch painter Vincent van Gogh, created in 1899. (b) **The painting** is stored in the Museum of Modern Art in New York. (c) **It** is one of the most famous works of Vincent van Gogh.)

(18) a. hahalmon nehaviyon ho stiyor sel pablo picasso hametaher hames zonot.

b. **hayestira** sestzoyra ben ha sanyim 1906-1907 nehsevet lehahat min heyestirot hayedohot sel picasso vesel hahomanot hamodernit.

c. **hayestira** mosteget kayom bemostehon lehomanot modernitt sebe new york.

((a) The Young Ladies of Avignon is a painting by Pablo Picasso that portrays five prostitutes. (b) **The artwork** that was painted during 1906-1907 is one of the most known works by Picasso in the modern art. (c) **The artwork** can today be seen in the Museum of Modern Art in New York City.)

The Hebrew examples also include definite noun phrases determined by the direct hyperonym relation, as *hastiyor* in (17b). Pronouns only occur in a context that contains a comparison, for example (17c). In other cases, e.g. (18b), (18c), the relation selected for the reference expression is higher-hyperonym.

The synonym relation occurs when giving the dimensions of the painting, as in (19b).

(19) a. Soded haken (1568) ho stiyor semen al luax est meet hastayar hapalmi peter broigel haav.

b. **hatmona** hi begodel 59 al 68 centimeter, ve **Ø** motseget bemozeon letoldot haaomanot bevina.

((a) The Nest thief (1568) is an oil painting made on wood by the painter Peter Brogel Hav. (b) **The picture** measures 59 x 68 cm, and is displayed in the art museum in Vienna.)

15.3.6 The results of the analysis

The above examples show a range of differences in the way chains of coreference are constructed. Table 15.5 summarizes the results the analysis revealed. 1st, 2nd and 3rd correspond to the first, second and third reference expression in the discourse.

In summary, we found:

- Pronoun is common in Swedish and English, and rare in Hebrew
- Direct-hyperonym is common in English, Swedish and Hebrew
- Higher-hyperonym is rare in English and Swedish, and common in Hebrew
- Synonym is common in Swedish, less frequent in English, and rare in Hebrew

Although the identified strategies are constrained by a relatively simple syntax and a domain ontology, they show clear differences between the languages. As table 15.5 shows, consecutive pronouns occur commonly in English, while consecutive higher hyperonym noun phrases are common in Hebrew.

Table 15.5: Coreference strategies for a painting object realisation. Pronoun (P), Synonym (S), Direct Hyperonym (DH), Higher Hyperonym (HH), Ellipsis (Ø).

DP	English			Swedish			Hebrew		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
1	DH	P		DH	P		DH	Ø	
1	DH	HH	Ø	DH	Ø		DH		
1	P	Ø		P	Ø				
1	P	P	Ø	Ø	DH				
1				Ø	P	DH			
1,2	P	DH		P	S	Ø			
2							HH	HH	
2							HH	Ø	HH
3	P	DH		P	DH				

15.4 Generating referential chains from Web ontology

15.4.1 Experimental data

We made use of the data available in the painting ontology presented in section 15.3.4 to generate multilingual descriptions by following the domain discourse patterns. The data consists of around 1000 ontology statements and over 250 lexicalised entities extracted from the Swedish National Museums of World Culture and the Gothenburg City Museum.

15.4.2 The generation grammar

The grammar was implemented in GF, a grammar formalism oriented toward multilingual grammar development and generation (Ranta 2004). It is a logical framework based on a general treatment of syntax, rules, and proofs by means of a typed λ-calculus with dependent types (Ranta 1994). Similar to other logical formalisms, GF separates between abstract and concrete syntaxes. The abstract syntax reflects the type theoretical part of a grammar. The concrete syntax is formulated as a set of linearization rules that can be superimposed on an abstract syntax to generate words, phrases, sentences, and texts of a desirable language. In addition, GF has an associated grammar library (Ranta 2009); a set of parallel natural language grammars that can be used as a resource for various language processing tasks.

Our grammar consists of one abstract module that reflects the domain knowledge and is common to all languages, plus three concrete modules, one for each language, which encode the language dependent strategies. Rather than giving details of the grammatical formalism, we will show how GF captures the constraints presented in section 15.3.6.

English

```
painting paintingtype painter
              year museum = let
str1 : Phr = mkPhr
(mkS (mkCl (mkNP painting) (mkVP
(mkVP  (mkNP
(mkNP a_Art paintingtype) make_V2))
(mkAdv by8agent_Prep
(mkNP (mkNP  painter)
(mkAdv in_Prep year.s))))));
str2 : Phr = mkPhr (mkS
(mkCl (mkNP the_Art paintingtype)
(mkVP (passiveVP display_V2)
(mkAdv at_Prep  museum.s))))
in mkText str1 (mkText str2) ;
```

Swedish

```
painting paintingtype painter
              year museum = let
str1 : Phr = mkPhr
(mkS (mkCl (mkNP  painting)
(mkVP (mkVP
(mkNP a_Art paintingtype))
(mkAdv by8agent_Prep
(mkNP (mkNP painter)
(mkAdv from_Prep (mkNP year)))))));
str2 :  Phr = mkPhr
(mkS (mkCl (mkNP the_Art
(mkN "tavla" "tavla"))
(mkVP (mkVP (depV  finna_V))
(mkAdv on_Prep (mkNP museum)))) )
in mkText str1 (mkText str2) ;
```

Hebrew

```
painting paintingtype painter
              year museum = let
str1 : Str = ({s =  painting.s ++
paintingtype.s ++  "sl " ++
painter.s ++ "msnt " ++ year.s}).s;
str2 : Str = ({s = artwork_N.s ++
(displayed_V ! Fem) ++ at_Prep.s ++
museum.s}).s in
ss (str1 ++ " ." ++ str2 ++ " ." );
```

The above extracts from the concrete modules follow the observed organization principles concerning the order of semantic information in a discourse and the generation of language-dependent referential chains (presented in the right-hand column of table 15.6). In these extracts, variations in referential forms are captured in the noun phrase of *str2*. In the English module, the *paintingtype* that is the direct hyperonym of the painting object is coded, while in the Swedish module, a synonym word of the painting concept is coded, e.g *tavla*. In the Hebrew module, a higher concept in the hierarchy of paintings, *artwork_N.s* is coded.

15.4.3 Experiments and results

A preliminary evaluation was conducted to test how significant is the approach of adapting language-dependent coreference strategies to produce coherent descriptions. Nine human subjects participated in the evaluation, three native speakers of each language.

The subjects were given forty object description pairs. One description containing only pronouns as the type of referring expressions and one description that was automatically generated by applying the language dependent coreference strategies. Examples of the description pairs the subjects were asked to evaluate are given in table 15.6. We asked the subjects to choose the description they find most coherent based on their intuitive judgements. Participant agreement was measured using the kappa statistic (Fleiss 1971). The results of the evaluation are reported in table 15.7.

On average, the evaluators approved at least half of the automatically generated descriptions, with a considerably good agreement. A

Table 15.6: Examples of object description-pairs from experiment 1.

English	
The Little White Girl is a painting by James Abbott McNeill Whistler. **It** is held in the Gotheburg Art Museum. The Long Winter is a painting by Peter Kandre from 1909. **It** measures 102 by 43 cm. **It** is displayed in the Museum Of World Culture.	The Little White Girl is a painting by James Abbott McNeill Whistler. **The painting** is held in the Gotheburg Art Museum. The Long Winter is a painting by Peter Kandre from 1909. **It** measures 102 by 43 cm. **The painting** is displayed in the Museum Of World Culture.
Swedish	
Den lilla vita flickan är en målning av James Abbott McNeill Whistler. **Den** återfinns på Göteborgs Konstmuseum. Den långa vintern målades av Peter Kandre 1909. **Den** är 102 cm lång och 43 cm bred. **Den** återfinns på Världskulturmuseet.	Den lilla vita flickan är en målning av James Abbott McNeill Whistler. **Målningen** återfinns på Göteborgs Konstmuseum. Den långa vintern målades av Peter Kandre 1909. **Målningen** är 102 cm lång och 43 cm bred. **Tavlan** återfinns på Världskulturmuseet.
Hebrew	
הילדה הקטנה הלבנה היא תמונה של קימס אבוט מקניל. היא מוצגת במוזיאון האומנות של גוטנבורג . חורף ארוך צויר על ידי פיטר קנדרה בשנת 1909. הוא בגודל 102 על 43 ס ״ מ. הוא מוצג במוזיאון האומנות של עולם התרבות.	הילדה הקטנה הלבנה היא תמונה של קימס אבוט מקניל. היצירה מוצגת במוזיאון האומנות של גוטנבורג . חורף ארוך צויר על ידי פיטר קנדרה בשנת 1909. היצירה בגודל 102 על 43 ס ״ מ. היצירה מוצגת במוזיאון האומנות של עולם התרבות.

Table 15.7: A summary of the human evaluation.

	Pronouns	Pronouns/NPs	$\mathcal{K}$
English	17	18	0.66
Swedish	9	29	0.78
Hebrew	6	28	0.72

closer look at the examples where chains of pronouns were preferred revealed that these occurred in English when a description consisted of two or three sentences and the second and third sentences specified the painting dimensions or a date. In Swedish, these were preferred whenever a description consisted of two sentences. In Hebrew, the evaluators preferred a description containing a pronoun over a description containing the higher hyperonym *Man-made object*, and also preferred the pronoun when a description consisted of two sentences, the second of which concerned the painting dimensions.

15.5 Conclusions and future work

This paper has presented a cross-linguistic study and demonstrated some differences in how coreference is expressed in English, Swedish and Hebrew. As a result of the investigation, a set of language-specific coreference strategies were identified and implemented in GF. This multilingual grammar was used to generate object descriptions which were then evaluated by native speakers of each language. The evaluation results, although performed with a small number of descriptions and human evaluators, indicate that language-dependent coreference strategies lead to better output. Although the data used to compare the co-referential chains was restricted in size, it was sufficient to determine several differences between the languages for the given domain.

Future work aims to extend the grammar to cover more ontology statements and discourse patterns. We will consider conjunctions and ellipsis in these patterns. We intend to formalize and generalize the strategies presented in this paper and test whether there exist universal co-referential chains, which might result in coherent descriptions in more than three languages.

REFERENCES

Abney, Steven P. 1991. Parsing by chunks. *Principle-Based Parsing*, 257–278. Kluwer Academic Publishers.

Adler, Meni and Michael Elhadad 2006. An unsupervised morpheme-based HMM for Hebrew morphological disambiguation. *Proceedings of the international conference on computational linguistics (COLING-ACL)*. Sydney: ACL.

Ahlberg, Malin 2010. Towards a wide-coverage grammar for Swedish. Master's thesis, Chalmers University of Technology.

Allemang, Dean and James Hendler 2008. *Semantic Web for the working ontologist: Effective modeling in RDFS and OWL*. San Francisco: Morgan Kaufmann.

Amato, G., J. Cigarran, J. Gonzalo, C. Peters and P. Savino 2008. MultiMatch – multilingual/multimedia access to cultural heritage. *Proc. of ECDL*, 505–508. Berlin: Springer.

Androutsopoulos, I., J. Oberlander and V. Karkaletsis 2007. Source authoring for multilingual generation of personalised object descriptions. *Natural Language Engineering* 13 (3): 191–233.

Androutsopoulos, Ion, Spyros Kallonis and Vangelis Karkaletsis 2005. Exploiting OWL ontologies in the multilingual generation of object descriptions. *The 10th European Workshop on Natural Language Generation (ENLG)*, 150–155. Aberdeen, UK.

Androutsopoulos, Ion, Vassiliki Kokkinaki, Aggeliki Dimitromanolaki, Jo Calder, Jon Oberl and Elena Not 2001. Generating multilingual personalized descriptions of museum exhibits: the M-PIRO project. *Proceedings of the International Conference on Computer Applications and Quantitative Methods in Archaeology.*

Angelov, Krasimir 2011. The Mechanics of the Grammatical Framework. Ph.D. diss., Chalmers University of Technology, University of Gothenburg.

Ariel, Mira 1988. Referring and accessibility. *Journal of Linguistics* 24 (1): 65–87.

Ariel, Mira 1990. *Accessing noun phrase antecedents*. London: Routledge.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak and Zachary Ives 2007. DBpedia: A nucleus for a web of open data. *Lecture Notes in Computer Science (LNCS)*, Volume 4825, 722–735. Berlin: Springer.

Baader, Franz, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi and Peter F. Patel-Schneider (eds) 2003. *The Description Logic handbook: Theory, implementation, and applications*. Cambridge: Cambridge University Press.

Barzilay, Regina and Lillian Lee 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. *Proceedings of HLT-NAACL*, 113–120. Boston: ACL.

Bateman, John A. 1997. Enabling technology for multilingual natural language generation: The KPML development environment. *Journal of Natural Language Engineering* 3 (1): 15–55.

Bechhofer, Sean, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider and Lynn Andrea Stein 2004. OWL Web Ontology Language Reference. W3C Recommendation.

Belz, Anja and Eric Kow 2010. The GREC challenges 2010: Overview and evaluation results. *Proceedings of the Sixth International Natural Language Generation Conference (INLG)*. Ireland: ACL.

Berners-Lee, Tim 1998. *Semantic Web road map. W3C design issues*. http://www.w3.org/DesignIssues/Semantic.html.

Berners-Lee, Tim 2006. Design Issues: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, Tim, James Hendler and Ora Lassila 2001. The Semantic Web. *Scientific American*, pp. 34–43.

Binding, Ceri 2010. Implementing Archaeological Time Periods Using CIDOC CRM and SKOS. Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral and Tania Tudorache (eds), *The semantic web: Research and applications*, Volume 6088 of *Lecture Notes in Computer Science*, 273–287. Berlin: Springer.

Bishop, B., A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev and R. Velkov 2011. OWLIM: A family of scalable semantic repositories. *Semantic Web Journal, Special Issue: Real-World Applications of OWL*.

Bizer, Christian, Tom Heath, Tim Berners-Lee and Michael Hausenblas

2011. 4th linked data on the web workshop (LDOW2011). *WWW*, 303–304.

Boas, Hans C. (ed.) 2009. *Multilingual FrameNets in Computational Lexicography* . Berlin: Mouton de Gruyter.

Bontcheva, Kalina 2005. Generating tailored textual summaries from ontologies. *Second European Semantic Web Conference (ESWC)*, 531–545.

Bontcheva, Kalina and Yorick Wilks 2004. Automatic report generation from ontologies: The MIAKT approach. *Proceedings of the nineth international conference on applications of natural language to information systems (NLDB)*, 324–335.

Borin, Lars 2005. Mannen är faderns mormor: Svenskt associationslexikon reinkarnerat. *LexicoNordica* 12: 39–54.

Borin, Lars, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj and Dimitrios Kokkinakis 2010. The past meets the present in the Swedish FrameNet++. *Proceedings of EURALEX*, 269–281. Leeuwarden: EURALEX.

Borin, Lars and Markus Forsberg 2009. All in the family: A comparison of SALDO and WordNet. *Proceedings of the 17th Nordic conference of computational linguistics (NODALIDA 2009)*, NEALT Proceedings Series, Vol. 4 (2009). Odense: NEALT.

Borin, Lars, Markus Forsberg and Lennart Lönngren 2008. The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. Joakim Nivre, Mats Dahllöf and Beata Megyesi (eds), *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia no. 7, 21–32. Uppsala: Uppsala University, Department of Linguistics and Philology.

Borin, Lars, Markus Forsberg and Lennart Lönngren 2008. SALDO 1.0 (Svenskt associationslexikon version 2). Technical Report, Språkbanken, Göteborg universitet.

Bouayad-Agha, Nadja, Richard Power and Donia Scott 2000. Can text structure be incompatible with rhetorical structure? *Proceedings of INLG*, 194–200.

Bowen, J.P. and S. Filippini-Fantoni 2004. Personalization and the Web from a museum perspective. *Museum and the Web*.

Brants, Thorsten 2000. TnT – a statistical part-of-speech tagger. *Proceedings of the 6th applied natural language processing conference (ANLP*. Seattle, Washington: ACL.

Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen and Eve Maler 2008. *Extensible Markup Language (XML) 1.0 (fifth edition)*. W3C. http://www.w3.org/TR/xml/.

Brickley, Dan and R.V. Guha 2004. *RDF vocabulary description language 1.0: RDF Schema*. W3C. http://www.w3.org/TR/rdf-schema/.

Brugman, Hennie, Véronique Malaisé and Laura Hollink 2008. A common multimedia annotation framework for cross linking cultural heritage digital collections. *International Conference on Language Resources and Evaluation (LREC)* . Marrakech: ELRA.

Brun, C., M. Dymetman and V. Lux 2000. Document structure and multilingual authoring. *In Proc. of First International Natural Language Generation Conference (INLG)*. Mitzpe Ramon, Israel.

Bryne, Kate 2008. Having Triplets – Holding Cultural Data as RDF. *Proceedings of IACH workshop at ECDL2008 (European Conference on Digital Libraries)*. Berlin: Springer.

Bryne, Kate 2009. Populating the Semantic Web – combining text and relational databases as RDF graphs. Ph.D. diss., University of Edinburgh.

Burden, H. and R. Heldal 2011. Natural Language Generation from Class Diagrams. *Proceedings of the 8th International Workshop on Model-Driven Engineering, Verification and Validation (MoDeVVa 2011)*. Wellington: ACM.

Busemann, Stephan and Helmut Horacek 1998. A flexible shallow approach to text generation. *Proceedings of the 9th International Workshop on Natural Language Generation (IWNLG 98)*, 238–247. Niagara-on-the-Lake, Ontario.

Cahill, Lynne J., John Carroll, Roger Evans, Daniel S. Paiva, Richard Power, Donia Scott and Kees van Deemter 2001. From RAGS to RICHES: Exploiting the Potential of a Flexible Generation Architecture. *ACL*, 98–105. Morgan Kaufmann Publishers.

Camilleri, John J. 2012. An IDE for the Grammatical Framework. *Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*. Gothenburg, Sweden.

Capra, R., G. Marchionini, J. S. Oh, F. Stutzman and Y. Zhang 2007. Effects of structure and interaction style on distinct search tasks. *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, 442–451. New York: ACM Press.

Chiarcos, Christian and Manfred Stede 2004. Salience-driven text planning. Anja Belz, Roger Evans and Paul Piwek (eds), *Natural*

language generation, third international conference, INLG 2004, Volume 3123 of *Lecture Notes in Computer Science*, 21–30. Brockenhurst: Springer.

Cimiano, P., P. Buitelaar, J. McCrae and M. Sintek 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Web Semantics* 9: 29–51.

Clough, Paul, Jennifer Marlow and Neil Ireson 2008. Enabling semantic access to cultural heritage: A case study of tate online. *Proceedings of the ECDL. Workshop on Information Access to Cultural Heritage*. Berlin: Springer.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20: 37–46.

Crofts, Nick, Martin Doerr, Tony Gill, Stephen Stead and Matthew Stiff 2009. *Definition of the CIDOC Conceptual Reference Model*. http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.1_Nov09.pdf (Last accessed 2012-12-12).

Croitoru, M. and K. van Deemter 2007. A conceptual graph approach to the generation of referring expressions. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Cruse, D.A. 1986. *Lexical semantics*. Cambridge: Cambridge university press.

Dale, R., S. J Green, M. Milosavljevic, C. Paris, C. Verspoor and S. Williams 1998. The realities of generating natural language from databases. *Proceedings of the Applications Track of the 11th Australian Joint Conference on Artificial Intelligence*, 62–74. Brisbane, Australia.

Dale, Robert and Ehud Reiter 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 18 (2): 233–263.

Dale, Robert and Jette Viethen 2009. Referring expression generation through attribute-based heuristics. *Proceedings of the 12th European workshop on natural language generation (ENLG 2009)*, 58–65. Athens: ACL.

Damova, Mariana 2011. *Data models and alignment*. Deliverable 4.2. MOLTO FP7-ICT-247914 .

Damova, Mariana and Dana Dannélls 2011. Reason-able View of Linked Data for cultural heritage. *Proceedings of the third International Conference on Software, Services and Semantic Technologies (S3T)*.

Damova, Mariana, Atanas Kiryakov, Maurice Grinberg, Michael K.

Bergman, Frederik Giasson and Kiril Simov 2011. Creation and integration of reference ontologies for efficient LOD management. *Semi-automatic ontology development: Processes and resources*. Hershey: IGI Global.

Dannélls, Dana 2008a. Generating tailored texts for museum exhibits. *The 2nd Workshop on Language Technology for Cultural Heritage (LaTeCH 2008)*, 17–20. Marrakech: ELRA.

Dannélls, Dana 2008b. A System Architecture for Conveying Historical Knowledge to Museum Visitors. *Workshop on Information Access to Cultural Heritage (IACH)*, Lecture Notes in Computer Science. Berlin: Springer.

Dannélls, Dana 2008c. The production of documents from ontologies. *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI)*, 36–38. Patras, Greece: IOS Press.

Dannélls, Dana 2009. Improving information access to cultural content through discourse strategies. *Proceedings of the eleventh in a series of international scientific Conferences on Advances in Artificial Intelligence held bi-annually by the Italian Association for Artificial Intelligence (AI*IA)*. Reggio Emilia, Italy.

Dannélls, Dana 2010a. Discourse generation from formal specifications using the Grammatical Framework, GF. *Special issue of the journal Research in Computing Science* 46: 167–178.

Dannélls, Dana 2010b. Applying semantic frame theory to automate natural language templates generation from ontology statements. *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)*, 179–184. Dublin: ACL.

Dannélls, Dana 2011. *D.8.1 Ontology and corpus study of the cultural heritage domain*. Deliverable of EU Project MOLTO Multilingual Online Translation.

Dannélls, Dana 2012a. An ontology model of paintings. *Submitted to the Journal of Applied Ontology*.

Dannélls, Dana 2012b. On generating coherent multilingual descriptions of museum objects from semantic web ontologies. *Proceedings of the Seventh International Natural Language Generation Conference (INLG 2012)*, 76–84. Utica, IL: ACL.

Dannélls, Dana and Lars Borin 2012. Toward language independent methodology for generating artwork descriptions – exploring framenet information. *EACL workshop on Language Technology*

for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), 18–23. Avignon: ACL.

Dannélls, Dana and John J. Camilleri 2010. Verb morphology of Hebrew and Maltese – towards an open source type theoretical resource grammar in gf. *Proceedings of the language resources (LRs) and human language technologies (HLT) for semitic languages: Status, updates, and prospects*, 57–61. Valletta: ELRA.

Dannélls, Dana, Mariana Damova, Ramona Enache and Milen Chechev 2011. A framework for improved access to museum databases in the Semantic Web. *Language technologies for digital humanities and cultural heritage (LaTeCH)*. Portland, Oregon: ACL.

Dannélls, Dana, Ramona Enache, Mariana Damova and Milen Chechev 2012. Multilingual online generation from Semantic Web ontologies. *Proceedings of the World Wide Web Conference (WWW2012), European project track*, 239–242.

De Bleecker, Inge M. R. 2005. Towards an optimal lexicalization in a natural-sounding portable natural language generator for dialog systems. *ACL '05: Proceedings of the ACL Student Research Workshop*, 61–66. Ann Arbor: ACL.

Declerck, Thierry, Paul Buitelaar, Tobias Wunner, J. McCrae, Elena Montiel-Ponsoda and G. Aguado de Cea 2010. Lemon: An ontology-lexicon model for the multilingual semantic web. *Proceedings of the w3c workshop: The multilingual web - where are we?* Madrid, España: Universidad Politécnica de Madrid.

De Coi, Juri Luca, Norbert E. Fuchs, Kaarel Kaljurand and Tobias Kuhn 2009. Controlled English for reasoning on the Semantic Web. François Bry and Jan Małuszyński (eds), *Semantic techniques for the Web – the REWERSE perspective*, Volume 5500 of *Lecture Notes in Computer Science*, 276–308. Berlin: Springer.

van Deemter, K. and R. Power 2003. High-level authoring of illustrated documents. *Natural Language Engineering* 9 (2): 101–126.

van Deemter, K. and E. Krahmer M. Theune 2005. Real versus template-based natural language generation: a false opposition? *Computational Linguistics* 31 (1): 15–23.

van Deemter, Kees 2002. Generating referring expressions: boolean extensions of the incremental algorithm. *Computational Linguistics* 28 (1): 37–52.

Degerstedt, Lars and Pontus Johansson 2003. Evolutionary develop-

ment of phase-based dialogue systems. *The 8th scandinavian conference on artif. intell*, 59–67.

Dekkers, Makx, Stefan Gradmann and Carlo Meghini 2009. *Europeana outline functional specification for development of an operational european digital library*. Europeana Thematic Network Deliverable 2.5.

Di Eugenio, B. 1998. *Centering in Italian*. M. A. Walker, A. K. Joshi and E. F. Prince (eds). Oxford: Clarendon Press.

Doerr, Martin 2005. The CIDOC CRM, an ontological approach to scheme heterogeneity. Dagstuhl Seminar (ed.), *Semantic interoperability and integration*, 1862–4405. Germany: In Dagstuhl Seminar, editor, Semantic Interoperability and Integration.

Doerr, Martin, Christian-Emil Ore and Stephen Stead 2007. The CIDOC conceptual reference model: A new standard for knowledge sharing. *ER '07: Tutorials, posters, panels and industrial contributions at the 26th international conference on conceptual modeling*.

Dymetman, M., V. Lux and A. Ranta 2000. XML and multilingual document authoring: Convergent trends. *In proceedings of the international conference on computational linguistics (COLING)*. Saarbrücken: ACL.

Enache, Ramona 2009. Reasoning and language generation in the SUMO ontology. Master's thesis, Chalmers University of Technology.

Enache, Ramona and Krasimir Angelov 2010. Typeful ontologies with direct multilingual verbalization. *LNCS post-proceedings of the controlled natural languages workshop (CNL 2010)*. Marettimo: Springer.

Eric, Prud'hommeaux and Seaborne Andy 2008. *SPARQL. The query language for RDF*. W3C Recommendation.

Erk, Katrin, Andrea Kowalski, Sebastian Padó and Manfred Pinkal 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. *Proceedings of the ACL*. Sapporo: ACL.

Fellbaum, C. 1998. *WordNet: An electronical lexical database*. Cambridge: MIT Press.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *In Quaderni di Semantica Sign Language Studies* 6 (2): 222—254.

Fillmore, Charles J., Christopher R. Johnson and Miriam R.L. Petruck 2003. Background to Framenet. *International Journal of Lexicography* 16 (3): 235–250.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (5): 378–382.

Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet and Claudia Soria 2006. LMF for multilingual, specialized lexicons. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 233–236. Genoa: ELRA.

Fuchs, N. and R. Schwitter 1996. Attempto controlled english (ACE). *Proceedings of the first international workshop on controlled language applications*. Belgium.

Fuchs, Norbert E., Kaarel Kaljurand and Tobias Kuhn 2008. Attempto controlled English for knowledge representation. *Reasoning Web, Fourth International Summer School 2008*, Lecture Notes in Computer Science no. 5224, 104–124. Berlin: Springer.

Galanis, Dimitrios and Ion Androutsopoulos 2007. Generating multilingual descriptions from linguistically annotated OWL ontologies: the NaturalOWL system. *Proceedings of the 11th European workshop on natural language generation*. Merzig-Wadern, Saarland: ACL.

Gatt, Albert, Anja Belz and Eric Kow 2008. The TUNA challenge 2008: Overview and evaluation results. Michael White, Crystal Nakatsu and David McDonald (eds), *International Natural Language Generation Conference (INLG)*. Fork, Ohio: ACL.

Gatt, Albert, Anja Belz and Eric Kow 2009. The TUNA-REG challenge 2009: Overview and evaluation results. *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, 174–182. Athens: ACL.

Gawronska, B. and B. Erlendsson 2005. Syntactic, semantic and referential patterns in biomedical texts: Towards in-depth text comprehension for the purpose of bioinformatics. *Proceedings of the 2nd international workshop on natural language understanding and cognitive science NLUCS*, 68–77. Miami, USA.

Geldof, Sabine and Walter van de Velde 1997. An architecture for template-based (hyper)text generation. *Proceedings of the Sixth European Workshop on Natural Language Generation*, 28–37. Duisburg.

Givón, T. (ed.) 1983. *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam and Philadelphia: John Benjamins.

Goldberg, Eli, Norbert Driedger and Richard Kittredge 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and Their Applications* 9: 45 – 53.

Goldberg, Yoav, Meni Adler and Michael Elhadad 2006. Noun phrase chunking in Hebrew: Influence of lexical and morphological features. *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics*, 689–696. Sydney: ACL.

Goldberg, Yoav, Meni Adler and Michael Elhadad 2008. EM can find pretty good HMM POS taggers (when given a good start). *Proceedings of ACL-08: HLT*, 746–754. Columbus, Ohio: ACL.

Goldberg, Yoav and Michael Elhadad 2010. An efficient algorithm for easy-first non-directional dependency parsing. *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Los Angeles: ACL.

Gómez-Pérez, Asunción, Mariano Fernández-López and Oscar Corcho 2004. *Ontological engineering : With examples from the areas of knowledge management, e-commerce and the Semantic Web*. Berlin: Springer.

Grosz, Barbara J., Scott Weinstein and Aravind K. Joshi 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21 (2): 203–225.

Gruber, Thomas R. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies* 43 (5–6): 907–928.

Gruzitis, Normunds, Peteris Paikens and Guntis Barzdins 2012. FrameNet Resource Grammar Library for GF. *Controlled Natural Language (CNL)*, Volume 7427, 121–137. Berlin: Springer.

Guarino, N. 1998. Formal ontology and information systems. *Proceedings of formal ontology in information systems (FOIS)*, 3–15. Trento: IOS Press.

Halácsy, Péter, András Kornai and Csaba Oravecz 2007. HunPos: an open source trigram tagger. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 209–212. Prague: ACL.

Halliday, Michael A. K. and R. Hasan 1976. *Cohesion in English*. Longman Pub Group.

Halliday, Michael A. K. and R. Hasan 1989. *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Harabagiu, S. and S. Maiorano 2000. Multilingual coreference resolution. *Proceedings of ANLP*.

Hardcastle, David and Donia Scott 2008. Can we evaluate the qual-

ity of generated text? *Proceedings of the 6th edition of the Language Resources and Evaluation Conference (LREC08)*. Marrakech: ELRA.

Hartley, A., D. Scott, J. Bateman and D. Dochev 2001. AGILE – A System for Multilingual Generation of Technical Instructions. *MT Summit VIII, Machine Translation in the Information Age*, 145–150.

Hasan, R. 1985. *Linguistics, language and verbal art.* Geelong: Deakin University Press.

Haslhofer, Bernhard and Antoine Isaac 2011. The Europeana Linked Open Data pilot. *Proceedings of the international conference on Dublin Core and metadata applications*.

Hein, Anna Sågvall 1985. On different kinds of knowledge in medical text comprehension. *Logic programming in knowledge engineering*.

Hein, Anna Sågvall 1989. Definite NPs and background knowledge in medical text. *Computer and Artificial Intelligence* 8 (6): 547–563.

Hielkema, Feikje, Chris Mellish and Peter Edwards 2008. Evaluating an Ontology-Driven WYSIWYM Interface. *Proceedings of the Fifth International Natural Language Generation Conference*.

Hobbs, Jerry R. 1979. Coherence and coreference. *Cognitive Science* 3 (1): 67–90.

Horrocks, I., P.F. Patel-Schneider and F. van Harmelen. 2003. From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics* 1 (1): 7–26.

Hovy, E. 2005. Methodologies for the reliable construction of ontological knowledge. *Proceedings of the International Conference on Conceptual Structures (ICCS)*, 91–106. Berlin: Springer.

Hunston, Susan 2006. *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Isard, Amy 2007. Choosing the best comparison under the circumstances. *Proceedings of the International Workshop on Personalization Enhanced Access to Cultural Heritage (PATCH07)*.

Isard, Amy, Jon Oberlander, Ion Androutsopoulos and Colin Matheson 2003. Speaking the users' languages. *IEEE Intelligent Systems Magazine* 18 (1): 40–45.

Jain, Prateek, Peter Z. Yeh, Kunal Verma, Reymonrod G., Mariana Damova, Vasquez Pascal Hitzler and Amit P. Sheth 2011. Contextual ontology alignment of LOD with an upper ontology: A case study with Proton. *Proceedings of 8th ESWC, extended semantic web conference*. Heraklion.

Jing, Hongyan and Kathleen McKeown 1998. Combining multiple, large-scale resources in a reusable lexicon for natural language generation. *Proceedings of the 17th international conference on computational linguistics*, 607–613. Montreal: ACL.

Johannisson, Kristofer 2005. Formal and informal software specifications. Ph.D. diss., Chalmers University of Technology.

Johansson, Pontus, Lars Degerstedt and Arne Jönsson 2002. Iterative development of an information-providing dialogue system. *Proceedings of 7th ercim workshop*. Uppsala: ACL.

Judgem, J., M. Sogrin and A. Troussov 2007. Galaxy: IBM ontological network miner. In Sören Auer, Christian Bizer, Claudia Müller and Anna V. Zhdanova (eds), *Proceedings of the 1st conference on social Semantic Web*, Volume 113, 157–160. Bonner Köllen Verlag.

Jurafsky, Daniel and James H. Martin 2008. *Speech and language processing (2nd edition)*. New Jersey: Prentice Hall.

Karamanis, Nikiforos, Massimo Poesio, Chris Mellish and Jon Oberlander 2009. Evaluating Centering for Information Ordering using Corpora. *Computational Linguistics*.

Karkaletsis, V., A Valarakos and C.D. Spyropoulos 2005. Populating ontologies in biomedicine and presenting their content using multilingual generation. *AIME*, 256–265.

Kelly, Colin, Ann Copestake and Nikiforos Karamanis 2009. Investigating content selection for language generation using machine learning. *Proceedings of the 12th European Workshop on Natural Language Generation*, 130–137.

Khegai, J., B. Nordström and A. Ranta 2003. Multilingual syntax editing in GF. *Intelligent text processing and computational linguistics (CICLing-2003)*, 453–464. Berlin: Springer.

Kilgarriff, Adam 2001. Comparing corpora. *Corpus Linguistics* 6 (1): 1–37.

Kilgarriff, Adam 2010. Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project. *Proceedings of the 3rd workshop on building and using comparable corpora, LREC10*, 1–5. Valletta: ELRA.

Kiryakov, A., D. Ognyanoff, R. Velkov, Z. Tashev and I. Peikov 2009. LDSR: Materialized reason-able view to the web of linked data. *Proceedings of OWL: Experiences and Directions (OWLED) 2009*. Chantilly, USA.

Komsell, Lina and Hanna Melén 2007. Museum på webben – en

undersökning om användbarhet och åtkomst. Master's thesis, Högskolan i Borås/Institutionen Biblioteks – och informationsvetenskap (BHS).

Konstantopoulos, Stasinos, Vangelis Karkaletsis and Dimitris Bilidas. 2009. An intelligent authoring environment for abstract semantic representations of cultural object descriptions. *Workshop on language technology and resources for cultural heritage, social sciences, humanities, and education*, 10–17. Athens: ACL.

Krahmer, E., S. van Erk and A. Verleg 2003. Graph based generation of referring expressions. *Computational Linguistics* 29 (1): 53–72.

Krahmer, Emiel and Kees van Deemter 2012. Computational generation of referring expressions: A survey. *Computational Linguistics* 38 (1): 173–218.

Krahmer, Emiel and Theune Mariet 2002. Efficient context-sensitive generation of referring expressions. Kees van Deemter and Rodger Kibble (eds), *Information sharing: Reference and presupposition in language generation and interpretation*, 223–264. Stanford, CA: CSLI Publications.

Lassila, O. and D.L. McGuinness 2001. The role of frame-based representation on the Semantic Web. Technical Report, Knowledge Systems Laboratory, Stanford University. Report KSL-01-02.

Lassila, Ora and Ralph R. Swick 1999. *Resource description framework (RDF). model and syntax specification*. W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax.

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas and Antonio Zampolli 2000. SIMPLE: A general framework for the development of multilingual lexicons. *Lexicography* 13 (4): 249–263.

Levin, Beth 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Ljunglöf, Peter 2004. Expressivity and complexity of the Grammatical Framework. Ph.D. diss., Chalmers University of Technology, University of Gothenburg, Gothenburg, Sweden.

Ljunglöf, Peter and Staffan Larsson 2008. A grammar formalism for specifying ISU-based dialogue systems. *Advances in natural language processing, 6th international conference, GoTAL 2008, Gothenburg, Sweden*, Volume 5221 of *Lecture Notes in Computer Science*, 303–314. Berlin: Springer.

Lönngren, Lennart 1989. Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi. Rapport ucdl-r-89-1, Centrum för datorlingvistik, Uppsala universitet.

Malmsten, Martin 2008. Making a library catalogue part of the semantic web. *Proceedings of the international conference on Dublin Core and Metadata Applications*. Berlin: Springer.

Mann, W. C. 1983. An overview of the Penman text generation system. *The third national conference on artificial intelligence*, 261–265. Washington, DC: ACL.

Martin-Löf, Per 1975. An Intuitionistic Theory of Types: Predicative Part. H. E. Rose and J. C. Shepherdson (eds), *Logic colloquium 1973*, 73–118. Amsterdam: North-Holland.

Martin-Löf, Per 1982. Constructive mathematics and computer programming. Cohen, Los, Pfeiffer and Podewski (eds), *Logic, methodology and philosophy of science VI*, 153–175. Amsterdam: North-Holland.

Martin-Löf, Per 1984. *Intuitionistic Type Theory*. Naples: Bibliopolis.

McCoy, Kathleen F. and Michael Strube 1999. Taking time to structure discourse: Pronoun generation beyond accessibility. *Proceedings of the twenty-third annual conference of the cognitive science society*, 378–383. Vancouver, CA.

McCrae, John, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asuncion Gomez-Perez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr and Tobias Wunner 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*. Berlin: Springer.

McGuinness, Deborah L. and Frank van Harmelen 2004. *OWL Web Ontology Language overview*. W3C Recommendation.

McKeown, Kathleen R. 1985. *Text generation: Using discourse strategies and focus constraints to generate natural language text*. Cambridge: Cambridge University Press.

Megyesi, Beata 2009. The open source tagger HunPoS for Swedish. *Proceedings of the 17th Nordic conference on computational linguistics (NODALIDA)*. Odense: NEALT.

van der Meij, Lourens, Antoine Isaac and Claus Zinn 2010. A Web-based repository service for vocabularies and alignments in the cultural heritage domain. *Proceedings of the 7th extended Semantic Web conference, (ESWC 2010)*.

Mellish, Chris 2010. Using Semantic Web technology to support NLG.

Case study: OWL finds RAGS. *Proceedings of the sixth international natural language generation conference (INLG)*. Dublin: ACL.

Mellish, Chris, Alistair Knott and Jon Oberlander 1998. Experiments using stochastic search for text planning. *Proceedings of international conference on natural language generation*. Ontario: ACL.

Mellish, Chris, Jon Oberlander and Alistair Knott 1998. An architecture for opportunistic text generation. *Proceedings of the ninth international workshop on natural language generation.* Ontario: ACL.

Mellish, Chris and Jeff Z. Pan 2008. Natural language directed inference from ontologies. *Artifical Intelligence* 172 (10): 1285–1315.

Mellish, Chris, Donia Scott, Lynne Cahill, Daniel Paiva, Roger Evans and Mike Reape 2006. A reference architecture for natural language generation systems. *Natural Language Engenering* 12 (1): 1–34.

Mellish, Chris and Xiantang Sun 2006a. The Semantic Web as a linguistic resource: Opportunities for natural language generation. *Knowledge-Based Systems* 19 (5): 298–303.

Mellish, Chris and Xiantang Sun 2006b. Domain independent sentence generation from RDF representations for the Semantic Web. *Combined workshop on language – Enabled educational technology and development and evaluation of robust spoken dialogue systems, European conference on AI*. Riva del Garda.

Melnik, Nurit 2007. Extending partial pro-drop in Modern Hebrew: A comprehensive analysis. *Proceedings of the HPSG07 conference*.

Meteer, Marie W 1990. Abstract linguistic resources for text planning. *Proceedings of the Fifth International Workshop on Natural Language Generation*, 62–69. Dawson, PA: ACL.

Milosavljevic, Maria 1997. Content selection in comparison generation. *6th European workshop on natural language generation (EWNLG)*, 72–81. Merzig–Wadern: ACL.

Murray, Gabriel, Giuseppe Carenini and Raymond T. Ng 2010. Generating and validating abstracts of meeting conversations: A user study. *Proceedings of the sixth international natural language generation conference*. Dublin: ACL.

Nardi, Daniele and Ronald J. Brachman 2003. An Introduction to Description Logics. F. Baader, D. Calvanese, D. McGuinness, D. Nardi and P. Patel-Schneider (eds), *The description logic handbook. theory, implementation and applications*, 1–40. Cambridge: Cambridge University Press.

Nivre, Joakim 2005. Dependency grammar and dependency parsing. Technical Report, Växjö University: School of Mathematics and Systems Engineering.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13 (2): 95–135.

Nordström, Bengt, Kent Petersson and Jan M. Smith 1990. *Programming in Martin-Löf's Type Theory: An Introduction*. Oxford: Oxford University Press.

Novello, A. and C. Callaway 2003. Porting to an Italian surface realizer: A case study. *Proc. of the 9th European Workshop on NLG*, 71–78.

O'Donnell, Michael J., Chris Mellish, Jon Oberlander and Alistair Knott 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering* 7 (3): 225–250.

Ohara, Kyoko Hirose, Seiko Fujii, Hiroaki Saito, Shun Ishizaki, Toshio Ohori and Ryoko Suzuki 2003. The Japanese framenet project: A preliminary report. *Proceedings of pacific association for computational linguistics*, 249–254. Halifax, Nova Scotia: ACL.

Ordan, Noam, Bar Ilan, Noam Ordan and Shuly Wintner 2007. Hebrew WordNet: a test case of aligning lexical databases across languages. *International Journal of Translation* 19: 39–58.

Ordan, Noam and Shuly Wintner 2005. Representing natural gender in multilingual lexical databases. *International Journal of Lexicography* 18 (3): 357–370.

Ore, Christian-Emil Smith 2001. The Norwegian museum project, access to and interconnection between various resources of cultural and natural history. *European conference on research and advanced technology for digital libraries ECDL*.

Paiva, Daniel S. 1998. A Survey of Applied Natural Language Generation Systems. Technical Report, Information Technology Research Institute(ITRI), University of Brighton, UK. ITRI-98-03.

Paraboni, Ivandré, Kees van Deemter and Judith Masthoff 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics* 33 (2): 229–254.

Paris, Cécile L., William R. Swartout and William C. Mann 1991. *Natural language generation in artificial intelligence and computational linguistics*. Berlin: Springer.

Passonneau, R. J. 1996. Using centering to relax Gricean informational

constraints on discourse anaphoric noun phrases. *Language and Speech* 32: 229–264.

Pedersen, Bolette Sandford, Lars Borin, Markus Forsberg, Krister Lindén, Heili Orav and Eírikur Rögnvaldsson 2012. Linking and validating Nordic and Baltic wordnets. *Proceedings of the 6th International Global Wordnet Conference*, 254–260.

Pianta, Emanuele, Luisa Bentivogli and Christian Girardi 2002. MultiWordnet: Developing and aligned multilingual database. *Proceedings of the first international conference on global wordnet*, 293–302. Mysore.

Potau, Marta Recasens 2008. Towards coreference resolution for Catalan and Spanish. Ph.D. diss., University of Barcelona.

Power, Richard 2010. Complexity assumptions in ontology verbalisation. *Proceedings of the Association for Computational Linguistics (ACL)*, 132–136. Uppsala: ACL.

Power, Richard and Donia Scott 1998. Multilingual authoring using feedback texts. *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics*, ACL '98, 1053–1059. Montreal: ACL.

Power, Richard and Donia Scott 2000. Can text structure be incompatible with rhetorical structure? *Proceedings of INLG*, 194–200.

Power, Richard, Donia Scott and Robert Evans 1998. What you see is what you meant: direct knowledge editings with natural language feedback. *13th European Conference on Artificial Intelligence (ECAI).*, 677–681. John Wiley and Sons, Chichester,England.

Prasad, Rashmi 2003. Constraints on the generation of referring expressions, with special reference to Hindi. Ph.D. diss., University of Pennsylvania.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. P. Cole (ed.), *Syntax and semantics: Vol. 14. radical pragmatics*, 223–255. New York: Academic Press.

Prince, Ellen F. 1992. The ZPG letter: Subjects, definiteness, and information-status. *Discourse description. diverse linguistic analyses of a fund-raising text*, Volume 10, 159–173.

Prince, Ellen F. 1994. Subject-prodrop in Yiddish. *Focus and natural language processing. volume I. intonation and syntax.*, 159–173. Heidelberg: University Press.

Ranta, Aarne 1994. *Type-theoretical grammar: A type-theoretical grammar formalism*. Oxford: Oxford University Press.

Ranta, Aarne 2004. Grammatical Framework, a type-theoretical grammar formalism. *Journal of Functional Programming* 14 (2): 145–189.

Ranta, Aarne 2009. The GF resource grammar library. *The online journal Linguistics in Language Technology (LiLT)* 2, no. 2. http://elanguage.net/journals/index.php/lilt/article/viewFile/214/158.

Ranta, Aarne 2011. *Grammatical Framework: Programming with multilingual grammars*. CSLI Studies in Computational Linguistics. Stanford: CSLI.

Ranta, Aarne and Krasimir Angelov 2010. Implementing Controlled Languages in GF. *Lecture notes in computer science. CNL-2009 Controlled Natural Languages*, 82–101.

Ranta, Arne 1991. Intuitionistic Categorial Grammar. *Linguistics and Philosophy* 14: 203–239.

Reiter, Ehud 1994. Has a consensus NL generation architecture appeared and is it psycholinguistically plausible? *Proceedings of the Seventh International Workshop on Natural Language Generation*, 163–170. Kennebunkport: ACL.

Reiter, Ehud 1999. Shallow vs. deep techniques for handling linguistic constraints and optimisations. Tilman Becker and Stephan Busemann (eds), *Workshop at the 23d German Annual Conference for Artificial Intelligence (KI99)*, 7–12. Bonn: DFKI-D-99-01.

Reiter, Ehud and Anja Belz 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics* 35 (4): 529–558.

Reiter, Ehud and Robert Dale 2000. *Building Natural Language Generation Systems*. MIT Press and The McGraw-Hill Companies, Inc.

Reiter, Ehud and Chris Mellish 1992. Using classification to generate text. *Proceedings of the 30th annual meeting of the association for computational linguistics (ACL92)*. Newark: ACL.

Reiter, Ehud and Chris Mellish 1993. Optimizing the costs and benefits of natural language generation. *Proceedings of the 13th international joint conference on artificial intelligence (IJCAI 93)*, 1164–1169. Chambery, France: Morgan Kaufmann Publishers.

Reiter, Ehud, Roma Robertson and Liesl. Osman 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence* 144: 41–58.

Rosell, Mgnus 2009. Text clustering exploration - Swedish text representation and clustering results unraveled. Ph.D. diss., KTH.

Schalley, Andrea C. and Dietmar Zaefferer (eds) 2007. *Ontolinguistics*.

Trends in Linguistics. Studies and Monographs. Berlin: Mouton de Gruyter.

Schreiber, Guus, Alia Amin, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Laura Hollink, Zhisheng Huang, Janneke Kersen, Marco Niet, Borys Omelayenko, Jacco Ossenbruggen, Ronny Siebes and Jos T 2006. Multimedia E-Culture Demonstrator. Isabel F. Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Michael Uschold and Lor Aroyo (eds), *International Semantic Web Conference*, Volume 4273, 951–958. Berlin: Springer.

Schwitter, R. and M. Tilbrook 2004. Controlled Natural Language meets the Semantic Web. *Proceedings of the Australasian Language Technology Workshop*, 55–62. Macquarie University.

Scott, D. 1999. The Multilingual Generation Game: Authoring Fluent Texts in Unfamiliar Languages. *Proc. of the 16th International Joint Conference on Artificial Intelligence (IJCAI)*, 1407–1411.

Siddharthan, Advaith 2011. Text simplification using typed dependencies: A comparison of the robustness of different generation strategies. *Proceedings of the European Workshop on Natural Language Generation (ENLG)*.

Sinclair, John 2001. *English Dictionary for Advanced Learners*. Collins Cobuild.

Sowa, F.J. 2000. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks Cole Publishing Co.

Sripada, S., E. Reiter, J. Hunter and J. Yu 2003. Summarising neonatal time-series data. *Proceedings of the research note sessions of the EACL*, 167–170. Budapest: ACL.

Staab, S. and R. Studer 2004. *Handbook on Ontologies*. International Handbooks on Information Systems. Berlin: Springer.

Stede, Manfred 1996. Lexical semantics and knowledge representation in multilingual sentence generation. Ph.D. diss., Department of Computer Science, University of Toronto.

Strzalkowski, Tomek (ed.) 1994. *Reversible Grammar in Natural Language Processing*. Boston: Kluwer Academic Publishers.

Studer, R., V.R. Benjamins and D. Fensel. 1998. Knowledge engineering: Principles and methods. *Data Knowledge Engineering* 25 (1): 161–197.

Subirats, Carlos and Miriam R. L. Petruck 2003. Surprise: Spanish

FrameNet. *Workshop on Frame Semantics, International Congress of Linguists.* Prague, Czech Republic.

Teich, Elke 1999. *Systemic Functional Grammar in Natural Language Generation: Linguistic Description and Computational Representation*. London: Cassell Academic.

Terziev, I., A. Kiryakov, and D. Manov 2005. *D.1.8.1 Base upper-level ontology (bulo) guidance.* Deliverable of EU-IST Project IST .

Tsarfaty, R., J. Nivre and E. Andersson 2012. Joint Evaluation of Morphological Segmentation and Syntactic Parsing. *Proceedings of the international meeting of the Association of Computational Linguistics (ACL 12)*. Jeju Island, Korea: ACL.

Uschold, M. and M. King 1995. Towards a methodology for building ontologies. *In Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*. Montreal.

Vallduví, Enric and Elisabet Engdahl 1996. The linguistic realization of information packaging. *Linguistics* 34: 459–519.

W3C 2009. *OWL 2 Web Ontology Language: Document overview*. W3C. OWL Working Group.

Walker, Marilyn, Sharon Cote and Masayo Iida 1996. Japanese discourse and the process of centering. *Computational Linguistics* 20 (2): 193–232.

Wilcock, Graham 2003. Talking OWLs: Towards an ontology verbalizer. *Human Language Technology for the Semantic Web and Web Services*, 109–112. Sanibel Island.

Wilcock, Graham and Kristiina Jokinen 2003. Generating responses and explanations from RDF/XML and DAML+OIL. *Knowledge and Reasoning in Practical Dialogue Systems IJCAI*, 58–63. Acapulco.

Williams, Sandra, Allan Third and Richard Power 2011. Levels of organisation in ontology verbalisation. *Proceedings of the 13th European Workshop on Natural Language Generation*, European Workshop on Natural Language Generation (ENLG '11), 158–163. Nancy: ACL.

Wintner, Shuly 2000. Definiteness in the Hebrew noun phrase. *Journal of Linguistics* 36: 319–363.

Yeh, Ching-Long and Chris Mellish 1997. An empirical study on the generation of anaphora in Chinese. *Computational Linguistics* 23 (1): 171–190.

Young, Michael R. 1999. Using Grice's maxim of quantity to select the content of plan descriptions. *Artificial Intelligence* 115 (9): 215–256.

A Appendix: PoS tag sets

A.1 English

Table A.1: English PoS categories.

Tag	Meaning
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to

UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

A.2 Swedish

Table A.2: Swedish PoS categories.

Tag	Meaning
AB	Adverb
DT	Determiner
HA	WH-adverb
HD	WH-determiner
HP	WH-pronoun
HS	WH-possessive
IE	Infinitival marker
IN	Interjection
JJ	Adjective
KN	Coordinating conjunction
NN	Noun
PC	Participle
PL	Particle
PM	Proper Noun
PN	Pronoun
PP	Preposition
PS	Possessive pronoun
RG	Cardinal number
RO	Ordinal number
SN	Subordinating conjunction
VB	Verb
UO	Foreign word
MAD	Major delimiter
MID	Minor delimiter

A.3 Hebrew

Table A.3: Hebrew PoS categories.

Tag	Meaning
AGR-gn	Agreement particle
AT	Accusative marker
AUX	Auxiliary verb
CC	Coordinating conjunction
CD-gn-(H)	Numeral (definite)
CDT-gn-(H)	Numeral determiner (definite)
COM	Complementizer
DT	Determiner
IN	Preposition
JJ-gn-(H)	Adjective (definite)
JJT-gn	Construct state adjective
H	Definiteness marker
HAM	Yes/No question word
MD-gnpt	Modal
MOD	Modifier
NN-gn-(H H-gnp)	Noun (definite definite-genitive)
NNG-gn-(H H-gnp)	Gerund noun (definite definite-genitive)
NNGT-gn	Construct state gerund
NNP-gn	Proper noun
NNT-gn	Construct state noun
POS	Possessive item
PRP-gnp	Personal pronoun
QW	Question/WH word
RB	Adverb
RBR	Adverb, comparative
REL	Relativizer
VB-gnpt	Verb, finite
VB-M	Verb, infinite
WDT-gn	Determiner question word
ZVL	Garbage
yy*	various symbols

B Appendix: Dependency category sets

B.1 English

Table B.1: English dependency categories.

Tag	Meaning
appos	appositional modifier
attr	attribute
aux	auxiliary
auxpass	passive auxiliary
cc	coordination
ccomp	clausal complement with internal subject
comp	complement
compl	complementizer
conj	conjunct
cop	copula
csubj	clausal subject
det	determiner
dobj	direct object
infmod	infinitival modifier
iobj	indirect object
mod	modifier
mod	modifier
nn	noun compound modifier
nsubj	nominal subject
nsubjpass	passive nominal subject
obj	object
partmod	participial modifier
pobj	object of preposition
poss	possession modifier
prep	prepositional modifier

rcmod	relative clause modifier
ref	referent
subj	subject
tmod	temporal modifier
xcomp	clausal complement with external subject

B.2 Swedish

Table B.2: MAMBA dependency categories.

Tag	Meaning
+A	Conjunctional adverbial
+F	Coordination at main clause level
AA	Other adverbial
AG	Agent
AN	Apposition
AT	Nominal (adjectival) pre-modifier
CA	Contrastive adverbial
DB	Doubled function
DT	Determiner
EF	Relative clause in cleft
EO	Logical object
ES	Logical subject
ET	Other nominal post-modifier
FO	Dummy object
FP	Free subjective predicative complement
FS	Dummy subject
FV	Finite predicate verb
I?	Question mark
IC	Quotation mark
IG	Other punctuation mark
IK	Comma
IM	Infinitive marker
IO	Indirect object
IP	Period
IQ	Colon
IR	Parenthesis
IS	Semicolon
IT	Dash
IU	Exclamation mark
IV	Nonfinite verb

JC	Second quotation mark
JG	Second (other) punctuation mark
JR	Second parenthesis
JT	Second dash
KA	Comparative adverbial
MA	Attitude adverbial
MS	Macrosyntagm
NA	Negation adverbial
OA	Object adverbial
OO	Direct object
OP	Object predicative
PL	Verb particle
PR	Preposition
PT	Predicative attribute
RA	Place adverbial
SP	Subjective predicative complement
SS	Other subject
TA	Time adverbial
TT	Address phrase
UK	Subordinating conjunction
VA	Notifying adverbial
VO	Infinitive object complement
VS	Infinitive subject complement
XA	Expressions like "så att säga" (so to speak)
XF	Fundament phrase
XT	Expressions like "så kallad" (so called)
XX	Unclassifiable grammatical function
YY	Interjection phrase
CJ	Conjunct (in coordinate structure)
HD	Head
IF	Infinitive verb phrase minus infinitive marker
PA	Complement of preposition
UA	Subordinate clause minus subordinating conjunction
VG	Verb group

B.3 Hebrew

Table B.3: Hebrew dependency categories.

Tag	Meaning
acomp	adjectival complement
advcl	adverbial clause modifier
amod	adjectival modifier
appos	appositional modifier
attr	attribute
ccomp	clausal complement with internal subject
comp	complement
compl	complementizer
conj	conjunction
cop	copula
def	definiteness
dobj	direct object
gen	genitive
hd	head daughters
infmod	infinitival modifier
iobj	indirect object
mod	modifier
nn	noun compound modifier
num	number
obj	object
partmod	participial modifier
pobj	object of preposition
possmod	possession modifier
prep	preposition
punct	punctuation
rcmod	relative clause modifier
subj	subject
tmod	temporal modifier
xcomp	clausal complement with external subject

C Appendix: Semantic categories set

Table C.1: Semantic categories.

Tag	Meaning
AAP	Actor Appellation
ACT	Actor
COL	Collection (like an exhibition)
CON	Conceptual object
DIM	Dimension
ENT	Entity
EVT	Event
LEB	Legal Body (like museum or institute)
MAT	Material
MMO	Man-Made object
PLC	Place
TIT	Title
TMP	Time-Span

D Appendix: Hebrew character sets

D.1 Transliteration and transcription letters

There are two transliteration tables provided in table D.1: one supported by the parser and one supported by the grammar.

Table D.1: Transliteration and transcription letters.

Hebrew	Transliteration		Transcription
Letter	Mila	GF	
א	A	A	a
ב	B	b	b̆ or b
ג	G	g	g
ד	D	d	d
ה	H	h	h
ו	W	w	v
ז	Z	z	z
ח	X	H	h
ט	J	T	t
י	I	y	y
כ	K	k	k̆ or k
ך	K	K	k̆
ל	L	l	l
מ	M	m	m
ם	M	M	m
נ	N	n	n
ן	N	N	n
ס	S	S	s
ע	E	O	ʼ
פ	P	p	p̆ or p
ף	P	P	p̆

צ	C	Z	c
ץ	C	Z.	c
ק	Q	q	q
ר	R	r	r
ש	F	s	š or ś
ת	T	t	t

E Appendix: The RGL categories and functions

E.1 Categories

Table E.1: Phrasal and lexical categories.

Category	Explanation
Text	text
Phr	phrase
S	declarative sentence
NP	noun phrase
VP	verb phrase
Cl	declarative clause
CN	common noun (without determiner)
PN	person name
Pron	personal pronoun
Prep	preposition
Adv	verb phrase modifying adverb
N	common noun
V	one place verb
V2	verb with an NP complement

E.2 Functions

Table E.2: Functions in the RGL.

Function	Type	Example
mkText	S –> Text	she slept
mkPhr	Cl –> Phr	she sleeps
mkS	Conj –> S –> S –> S	she sleeps and I run
mkCl	NP –> VP –> Cl	she always sleeps
mkCl	NP –> V2 –> NP –> Cl	she loves him
mkNP	Det –> CN –> NP	the old man
mkNP	PN –> NP	Paris
mkNP	Pron –> NP	we
mkVP	V2 –> NP –> VP	to love him
mkCN	N –> CN	house
mkAdv	Prep –> NP –> Adv	in the house