

Uncovering Prepositional Senses

Tine Lassen



Copyright © 2010

Tine Lassen

Computer Science
Department of Communication,
Business and Information Technologies



Roskilde University
P. O. Box 260
DK-4000 Roskilde
Denmark

Telephone: +45 4674 3839
Telefax: +45 4674 3072
Internet: http://www.ruc.dk/dat_en/
E-mail: datalogi@ruc.dk

All rights reserved

Permission to copy, print, or redistribute all or part of this work is granted for educational or research use on condition that this copyright notice is included in any copy.

ISSN 0109-9779

Research reports are available electronically from:
http://www.ruc.dk/dat_en/research/reports/

Uncovering Prepositional Senses

by

Tine Lassen

A dissertation presented to the faculties of Roskilde University
in partial fulfillment of the requirement for
the PhD degree

Department of Communication, Business and Information Technologies
Roskilde University, Denmark, 2010

For my parents



Betty Lassen

11.26.1928 - 12.26.2002

Jørgen Varde Lassen

02.26.1930 - 10.09.2008

Abstract

This dissertation is concerned with the semantics of Danish prepositions in an ontology-based information retrieval framework. In such a framework, conceptual indexing of texts is needed and, for us, the goal of this indexing process is to index texts based on the conceptual content of larger text chunks – ideally based on the conceptual content of sentences. The conceptual content of text chunks is mapped into a so-called generative ontology, which is to be understood as a non-finite set of concepts. Basically, a generative ontology consists of a given finite ontology ordered by the ISA relation called the skeleton ontology, and a set of production rules (cf. generative grammars) that allows for production of compound concepts. We represent such compound concepts in the ontology language ONTOLOG. In this language, compound concepts are represented as conceptual feature structures of the form $c[rl:cl]$. The attributions consist of pairs of relations and concept arguments which function as conceptual restrictions on the core concept. However, the generative ontology should not admit arbitrary combinations of relations and concepts: We thus propose to introduce ontological affinities that may specify ontologically admissible ways of combining concepts. The main focus of the dissertation is to identify such ontological affinities for semantic relations denoted by a selection of Danish prepositions. We describe two experiments: The first small-scale experiment concerns a domain-specific corpus which includes texts from the domain of nutrition. For this corpus, sentences containing syntactic structures in the form of NP-PREP-NP are annotated with information about e.g. semantic types for heads of the noun phrases and the relation denoted by the preposition. The relations used in the annotation of these data stem from a small pre-defined set of relations, and the ontological type information stems from the SIMPLE ontology. The resulting data set was used as input to a machine-learning algorithm, and the result was a set of rules that predict the semantic relation of a preposition based on the ontological types of its arguments. Based on encouraging results of this first experiment, the second and larger experiment was launched. This experiment concerns a general language corpus for which the same type of syntactic structures were annotated. This time, the annotation used the newly released Danish language wordnet, DanNet, as a source of ontological type information, while the relations stem from a larger set of relations which were the result of an analysis of dictionary entries and

corpus evidences containing prepositions. Again, machine learning was applied, and the result was a set of rules. These rules were transformed into a dictionary of prepositional senses, where, given a preposition and a sense, ontological affinities are expressed as restrictions on the ontological types of the arguments. Thus, the essential results of this research is knowledge about the relations that subset of Danish prepositions can denote and the ontological affinities for these relations.

Dansk Resumé

Denne afhandling beskæftiger sig med danske præpositioners semantik inden for rammerne af ontologibaseret informationssøgning. Inden for sådanne rammer er begrebsbaseret indeksering nødvendig, og for os er målet for en sådan proces at indeksere tekster i forhold til det begrebsmæssige indhold af så store tekststykker som muligt – ideelt set af hele sætninger. Det begrebsmæssige indhold af tekststykker afbildes ind i en såkaldt generativ ontologi, som skal forstås som en ikke-finit mængde af begreber. Grundlæggende består en generativ ontologi af en given finit ontologi, hvor begreberne er ordnet ved ISA-relationen, som kaldes skeletonontologien, og en mængde af produktionsregler (jf. generative grammatikker) som muliggør en produktion af sammensatte begreber. Vi repræsenterer sådanne sammensatte begreber ved hjælp af ontologibeskrivelsessproget ONTOLOG. I dette sprog repræsenteres sammensatte begreber som begrebsmæssige trækstrukturer af formen $c[r1:c1]$. Træktilskrivelser består af par af relationer og argumenter i form af begreber, og disse fungerer som begrebsmæssige restriktioner på kernebegrebet. Den generative ontologi skal imidlertid ikke tillade tilfældige kombinationer af relationer og begreber: Derfor foreslår vi at der introduceres ontologiske affiniteter som kan specificere lovlige begrebssammensætninger. Hovedfokus for denne afhandling er identifikationen af sådanne ontologiske affiniteter for semantiske relationer der denoteres af en mængde af danske præpositioner. Vi beskriver således to eksperimenter: Det første mindre eksperiment behandler et domænespecifikt korpus som består af tekster fra ernæringsdomænet. For dette korpus opmærkes syntaktiske konstruktioner af formen NP-PREP-NP med information om fx. semantisk type for NP-kernerne, samt relationen der denoteres af præpositionen. De relationstyper, der benyttes i opmærkningen stammer fra en mindre prædefineret mængde, og den ontologiske typeinformation stammer fra SIMPLE-ontologien. Det resulterende datasæt blev siden brugt som input til en maskinindlæringsalgoritme, og resultatet af dette var en mængde regler som kan forudsige den semantiske relation for en given præposition baseret på argumenternes ontologiske typer. På grundlag af et opmuntrende resultat af dette første eksperiment blev det andet og mere omfattende eksperiment sat i gang. Dette eksperiment behandler et almensprogligt korpus, for hvilket samme typer af syntaktiske konstruktioner blev opmærket. Denne gang blev det nyligt offentliggjorte

danske ordnet, DanNet, benyttet som kilde til de ontologiske typer, mens relationerne stammer fra en større mængde af relationer som er resultatet af en analyse af ordbogsindgange og korpusbelæg indeholdende præpositioner. Igen blev maskinindlæring anvendt og resulterede i en mængde regler. Disse regler blev omsat til en præpositionsordbog hvor, givet en præposition og en semantisk relation, de ontologiske affiniteter udtrykkes som restriktioner på de ontologiske typer af argumenterne. Således er de væsentligste resultater af denne forskning viden om hvilke relationer en delmængde af danske præpositioner kan denotere, samt viden om ontologiske affiniteter for disse relationer.

Acknowledgements

I would like to thank everybody who has supported and encouraged me in finishing this dissertation.

Thank you to Roskilde University and the OntoQuery project (funded by the Danish Research Agency under the Information Technology programme) for funding my fellowship and to the SIABO project (funded by the Danish Strategic Research Council under the NABIIT programme) for funding my employment as a research assistant. Also, thanks to Copenhagen Business School for letting me use an office there.

Especially, I would like to thank my supervisors, Troels Andreasen and Per Anker Jensen, everybody involved in the SIABO-project and my colleague Thomas Vestskov Terney for collaboration and fruitful discussions. I would also like to thank my family and friends, and last but not least, Dorthe for encouragement.

In addition, my sister Merete Bert Lassen deserves a big thank you for proof-reading the dissertation, as well as my niece, Mathilde, for the drawings for figures 36 and 37.

Table of Contents

Chapter 1 Introduction	1
1.1 Research Question	2
1.2 Outline of the Dissertation	3
1.3 Notational conventions	5
Chapter 2 Prepositions	7
2.1 Word Classes and Criteria for Word Classification	8
2.2 Views on the Essence of the Class of Prepositions	11
2.2.1 About a Quantitative Delimitation	12
2.2.2 About a Qualitative Delimitation	12
2.2.2.1 Morphology	12
2.2.2.2 Syntax	21
2.2.2.3 Semantics	31
2.2.2.4 Delimitation From other Word Classes	46
2.3 Our Definition of the Class of Prepositions	49
Chapter 3 Ontologies	55
3.1 What is an Ontology	56
3.2 Types of Ontologies	57
3.2.1 An Ontology Spectrum	60
3.2.2 An Ontology of Ontologies	63
3.3 Lexical Ontologies	66
3.3.1 Princeton WordNet	68
3.3.1.1 WordNet Principles and Evolution	68
3.3.1.2 Structure and Contents of the Database	71
3.3.2 EuroWordNet	78
3.3.2.1 Structure of the Language-independent Part of EWN	79
3.3.3 DanNet	87

3.3.3.1 Den Danske Ordbog.....	87
3.3.3.2 The Semantic Lexicon SIMPLE	89
3.3.3.3 Contents and Structure of DanNet	91
3.4 Summary	96
Chapter 4 Linguistic Expressions, Concepts and Semantic Relations	99
4.1 Concepts and Relations as Signs	99
4.1.1 What is a Concept.....	100
4.1.1.1 Concepts, Linguistic Expressions and Linguistic Signs	101
4.1.2 Semantic Relations	104
4.1.2.1 Relational Signs	105
4.2 Representation	107
4.2.1 Generative Ontologies and ONTOLOG	108
4.2.2 The Relation between the Sign and the Ontology	112
4.2.3 Atomic and Compound Concepts	112
4.2.4 Treatment of Unknown Words and Concepts.....	115
4.3 Relation Denoting Words	117
4.3.1 Verbs.....	118
4.3.1.1 Limited Set of Relations	120
4.3.1.2 Multifaceted Conceptual Content	120
4.3.1.3 Implications of Adding a Relation Hierarchy	122
4.3.1.4 Only Binary Relations.....	127
4.3.1.5 Modeling the Conceptual Content of Sentences.	127
4.3.1.6 Congruous Meaning of Nouns and Verbs.....	130
4.3.2 Prepositions.....	130
4.3.2.1 Finite Set of Relations Denoted by Prepositions.....	131
4.3.2.2 Relations Denoted by Prepositions are Binary	132
4.3.2.3 The Conceptual Content of a Preposition is not Multifaceted	132
4.3.2.4 Compound Concepts Reflecting the Conceptual Content of Sentences.....	135
4.4 Pluralities as Arguments.....	138
4.4.1 Concepts Denoted by Collective and Mass Nouns.....	139
4.4.2 Concepts Denoted by Coordinated Phrases	140
4.4.3 Concepts Denoted by Preposition Complements.....	142
4.4.4 Concepts Denoted by Verb Arguments	144
4.5 Summary	147

Chapter 5 A Machine Learning Approach to Disambiguation of Semantic Relations.....	149
5.1 Content-based Information Retrieval	150
5.2 Word Sense Disamiguation vs. Relation Disambiguation	150
5.3 Semantic Role Information - Available Resources	152
5.4 The Task	156
5.5 Semantic Relations	157
5.6 The Corpus	161
5.6.1 Annotation	161
5.6.2 The Ontological Type Annotation	163
5.6.3 The Set of Relations.....	164
5.7 Experiments.....	166
5.7.1 Analyzing the Rules.....	172
5.8 Summary	177
Chapter 6 Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions.....	179
6.1 Selection of Prepositions for Further Analysis.....	181
6.2 Semantic Relations denoted by Prepositions.....	184
6.3 Investigating Preposition Senses	185
6.4 Final Relation Set – The Relations One by One.....	189
6.4.1 ADD – ADDITION	190
6.4.2 AGT AGENT.....	190
6.4.3 BMO ‘BY MEANS OF’	191
6.4.4 CHR HAS CHARACTERISTIC	192
6.4.5 RCH INVERSE CHARACTERISTIC	192
6.4.6 CMP COMPRISING, has part.....	193
6.4.7 POF PART OF.....	194
6.4.8 COM COMPARISON	195
6.4.9 CUM CUM (with, accompanying)	197
6.4.10 SRC SOURCE (of event)	197
6.4.11 TAR TARGET (of event).....	198
6.4.12 RST RESULT	198
6.4.13 CAU CAUSES.....	199
6.4.14 CBY CAUSED BY.....	199
6.4.15 INH INHERENT RELATION	200
6.4.16 LOC LOCATIVE	201
6.4.17 RLO INVERSE LOCATIVE.....	203

6.4.18 MEA MEASURE	204
6.4.19 MNR MANNER	204
6.4.20 MTH MATH	205
6.4.21 PNT PATIENT	206
6.4.22 PRP PURPOSE	207
6.4.23 QUA QUA	208
6.4.24 SBT SUBSTITUTION	209
6.4.25 HPR HYPERNYMY	209
6.4.26 HPO HYPONYMY	210
6.4.27 SUP SUPERIORITY	210
6.4.28 TMP TEMPORAL	211
6.4.29 WRT WITH RESPECT TO	212
6.5 Investigations in Korpus 2000	213
6.5.1 Compiling a subcorpus	214
6.5.2 Annotation of Corpus-evidences	214
6.5.3 Ontological Type Annotation	215
6.5.4 Semantic Relation Annotation	217
6.6 Prepositions and the Relations they Denote	220
6.6.1 Af	220
6.6.2 Efter	222
6.6.3 For	223
6.6.4 Fra	224
6.6.5 Gennem/igennem	225
6.6.6 Hos	226
6.6.7 I	228
6.6.8 Med	229
6.6.9 Mellem/imellem	231
6.6.10 Over	232
6.6.11 På	233
6.6.12 Til	235
6.6.13 Under	236
6.6.14 Ved	237
6.7 Rules	239
6.7.1 Frequency-based Rule Deduction	239
6.7.2 Learning Rules with WEKA	243
6.8 Evaluation and Analysis of Selected Rules	250
6.8.1 The 10 most Precise Rules	252
6.8.2 The 10 most Covering Rules	257

6.8.3 The 10 ‘best’ Rules	263
6.9 Dictionary of Prepositions	272
6.9.1 Example Entries	273
6.10 Summary	274
Chapter 7 Conclusion and Future Work	277
Appendix A A Rule-based Dictionary of Danish Prepositions.....	282
Bibliography.....	302

List of Figures

Figure 1 The Danish system of prepositions (Brøndal 1940)	43
Figure 2 The English system of prepositions (Brøndal 1940)	44
Figure 3 Kinds of ontologies.....	58
Figure 4 A revised version of 'Kinds of ontologies'.....	59
Figure 5 An ontology spectrum (Lassila & McGuinness, 2001)	60
Figure 6 A revised ontology spectrum.....	63
Figure 7 Ontology of ontologies (B. N. Madsen & Thomsen, 2009)	65
Figure 8 Structure of EuroWordNet. The figure is based on (Piek Vossen et al., 1997) and (Piek Vossen et al., 1998).	81
Figure 9 The first levels of the EuroWordNet top-ontology. The top node here labeled 'Thing', should in reality be labeled 'top'. The tool used to produce this figure (Protégé), however, did not allow for a renaming of the top node.....	82
Figure 10 An unfolding of the concept <i>1stOrderEntity</i> in the EuroWordNet top-ontology	83
Figure 11 An unfolding of the concept <i>2ndOrderEntity</i> in the EuroWordNet top-ontology	85
Figure 12 An example entry from the online version of DDO for the lemma <i>oliemaleri</i> (oil painting) (DSL, 2010)	88
Figure 13 A template for the ontological type VEHICLE+ARTIFACT+OBJECT in DanNet (B. S. Pedersen et al., 2009).....	93
Figure 14 Excerpt of the DanNet top-ontology	94
Figure 15 Orthogonal and taxonomical hyponyms of the concept MALERI (B. Pedersen et al., 2009)	95
Figure 16 A drawing illustrating the concept horse	101
Figure 17 A photograph illustrating the concept horse.....	101
Figure 18 The general linguistic sign.....	103
Figure 19 A specific linguistic sign 'horse'	103
Figure 20 A synonymy relation between the words hack and nag	105
Figure 21 A linguistic expression-realization of the signifier in a relational sign.....	106
Figure 22 The null-realization as signifier in a relational sign	106
Figure 23 Ontology fragment showing the concepts A, B, C and B[REL:C].....	113

Figure 24	Ontology fragment showing the concepts horse, black and HORSE[CHR:BLACK].....	115
Figure 25	Position of the compound concept Pedometer[PRP:ESTIMATION[WRT:DISTANCE[WRT:WALKING]]].....	117
Figure 26	Ontology fragment containing the concepts Peter and house.....	123
Figure 27	Decomposition of the compound concept BUILDING[AGT:PETER,PRD:HOUSE].....	124
Figure 28	A fragment of a relation hierarchy based on Levin's verb classes.....	126
Figure 29	The atomic concepts Peter and house related via the 'build relation'.....	128
Figure 30	The compound concept PETER[BUILD:HOUSE].....	129
Figure 31	The compound concept BUILDING[AGT:PETER, PRD:HOUSE].....	129
Figure 32	A reified LOCATION relation.....	132
Figure 33	Gray area indicates possible position of the vase relative to the table given by <i>on</i>	133
Figure 34	Gray area indicates possible position of the vase relative to the table given by <i>above</i>	134
Figure 35	Gray area indicates possible position of the vase relative to the table given by <i>below</i>	134
Figure 36	The situation described by (32) with <i>on the table</i> modifying the verb <i>broke</i>	135
Figure 37	The situation described by (32) with <i>on the table</i> modifying the noun <i>vase</i>	136
Figure 38	The PP <i>on the table</i> modifying the verb <i>broke</i>	136
Figure 39	The PP <i>on the table</i> modifying the noun <i>vase</i>	137
Figure 40	A plurality composed of the concepts A and B.....	138
Figure 41	Mapping of the text <i>Peter saw a school of herring</i> into an ontology.....	139
Figure 42	Plurality-reading of <i>Peter and Mary wrote a book</i>	141
Figure 43	Individual-reading of <i>Peter and Mary wrote a book</i>	141
Figure 44	Plurality as a relatum.....	143
Figure 45	A plurality of relata.....	144
Figure 46	Plurality as a relatum for <i>intersection</i>	145
Figure 47	A plurality of relata for <i>intersection</i>	145
Figure 48	Ontology fragment showing the path from BLOODPROP (blood clot) to the top level of the SIMPLE ontology. The grey nodes are top level nodes, and white nodes are domain level nodes.....	164

Figure 49	Frequency distribution for the 12 relations in the corpus	167
Figure 50	Frequency distribution for the 15 prepositions in the corpus	167
Figure 51	Distribution of the data set into a training, tuning and test set.	169
Figure 52	The tasks and relation sets involved in the experiments.....	180
Figure 53	The contents of KorpusDK.....	213
Figure 54	Frequency distribution for relations in the data set	218
Figure 55	Frequency distribution for the relations denoted by <i>af</i>	222
Figure 56	Frequency distribution for the relations denoted by <i>efter</i>	223
Figure 57	Frequency distribution for the relations denoted by <i>for</i>	224
Figure 58	Frequency distribution for the relations denoted by <i>fra</i>	225
Figure 59	Frequency distribution for the relations denoted by <i>gennem</i>	226
Figure 60	Frequency distribution for the relations denoted by <i>hos</i>	228
Figure 61	Frequency distribution for the relations denoted by <i>i</i>	229
Figure 62	Frequency distribution for the relations denoted by <i>med</i>	231
Figure 63	Frequency distribution for the relations denoted by <i>mellem</i>	232
Figure 64	Frequency distribution for the relations denoted by <i>over</i>	233
Figure 65	Frequency distribution for the relations denoted by <i>på</i>	234
Figure 66	Frequency distribution for the relations denoted by <i>til</i>	236
Figure 67	Frequency distribution for the relations denoted by <i>under</i>	237
Figure 68	Frequency distribution for the relations denoted by <i>ved</i>	239
Figure 69	Precision scores for the most frequent relation per preposition	241
Figure 70	Frequency distribution for the 14 prepositions in the data set.....	242
Figure 71	Frequency distribution for the 14 prepositions in the citation version of Korpus 2000. The column <i>rest</i> shows the accumulated frequency for all tokens in Korpus 2000 that are tagged as prepositions.	242
Figure 72	Rule with low coverage and high precision. Box indicates rule coverage, circle indicates correct classification	250
Figure 73	Rule with high coverage and low precision. Box indicates rule coverage, circle indicates correct classification	251
Figure 74	Rule R with high recall, which classifies 100% of the members of the class C correctly. Outer box indicates the class C, inner box indicates coverage of rule R, and circle indicates correct classification.....	251
Figure 75	Rule R' with low recall, which classifies 50% of the members of class C' correctly. Outer box indicates the class C', inner box indicates coverage of rule R', and circle indicates correct classification.....	252

Figure 76 Matrix of relations and prepositions, showing which combinations resulted in rules.	271
Figure 77 Precision scores for the individual entries in the dictionary of prepositions	272

List of Tables

Table 1 Lexical matrix (cf. (G. Miller et al., 1990))	72
Table 2 Number of words, synsets, and senses in WordNet 3.0 (Princeton_University 2010).....	72
Table 3 Some semantic relations in WordNet (Christiane Fellbaum, 1998).	74
Table 4 EuroWordNet statistics (Piek Vossen, 2001).....	78
Table 5 A SIMPLE encoding template for the ontological type PHYSICAL CREATION (A. Lenci et al., 1999)	91
Table 6 Semantic relations in DanNet (Pedersen, Braasch et al. 2009).....	95
Table 7 The levels of concern for various generative frameworks	108
Table 8 Example text annotated with all levels of information	162
Table 9 The initial relation set consisting of 16 relations cf. (Nilsson, 2001)	165
Table 10 The final relation set consisting of the 12 relations that were used in the annotation - a subset of the relation set proposed in (Nilsson, 2001).	166
Table 12 The rules produced by the JRip algorithm for the data set with ontological types in the feature space, with number of matches, number of incorrect matches and accuracy score. The table is sorted by accuracy score.	174
Table 13 Confusion matrix for the JRip algorithm applied with prepositions and ontological types in the feature space.....	175
Table 14 The rules produced by the JRip algorithm for the data set with prepositions and ontological types in the feature space, with number of matches, number of incorrect matches and accuracy score. The table is sorted by accuracy score.	176
Table 15 frequency counts for the 25 most frequent prepositions in Korpus 2000.....	183
Table 16 Prepositions present in all 7 selected sources.	184
Table 17 Relation inventory as a combination of suggestions in (Nilsson, 2001) and (Nilsson, 1999).....	185
Table 18 Matrix of prepositions and relations based on the dictionary analysis.....	189

Table 19	○ In relation set from the dictionary analysis, but not used in the annotaion ● In relation set, and used in the annotation ⊕ Not in relation set, but used in the annotation.....	219
Table 20	Relations, frequencies and percentage of all relations.....	240
Table 21	Precision scores for the most frequent relation per preposition with arithmetic and weighted average.....	243
Table 22	The percentage of correctly classified instances and K-score for the output of the JRip and PART algorithms on four different combinations of input features with the split data set as input.	245
Table 23	The percentage of correctly classified instances and K-score for the output of the JRip and PART algorithms on four different combinations of input features with the non-split data set as input.	245
Table 24	Improvement in precision for rules produced with increasingly large feature spaces	247
Table 25	Classification matrix	252
Table 26	Scores for the 10 best rules by precision, most covering first	253
Table 27	The 10 most precise rules.....	253
Table 28	Scores for the 10 best rules by correctly covered instances.....	258
Table 29	10 best rules by correctly covered instances.....	258
Table 30	Scores for 10 best rules ranked by Q(R)	264
Table 31	The 10 ‘best’ rules.....	265

Chapter 1

Introduction

Following the ever-increasing amount of electronically stored texts, better methods for retrieval of texts are needed. For search in very large text collections, keyword-based retrieval models are inadequate. The inadequacy does not lie in a fact that the models do not return relevant documents to a given query, but rather that they also, to some degree, return irrelevant documents. Thus, it becomes cumbersome to find the most relevant documents in a large query result. In addition, and not least, many potentially relevant documents are not returned because they do not contain the exact query terms but, perhaps, they contain synonyms to the query term or different syntactic forms.

One possible solution to this problem is the introduction of conceptual search technology and ontologies in information search systems. Conceptual analysis of documents and queries paired with ontologies can improve search: Given a query expressed in a natural language, such a system can translate the query into a conceptual form and match it against an index that constitutes conceptual forms of words, phrases or sentences in the documents linked to an ontology. The index may also be expanded according to the ontology. The system can then return answers that either match the conceptual form of the query exactly or, by the inclusion of an ontology combined with some similarity measure, return documents that match the query to some extent. In order for such a system to work, documents must be indexed with respect to conceptual form and not with respect to word occurrences. This means that some kind of translation mechanism is needed in order to get from the textual form to a conceptual form.

Introduction

In our framework, in a conceptual indexing process, the conceptual content of text chunks is mapped into a so-called generative ontology. A generative ontology is to be understood as a non-finite set of concepts. Basically, a generative ontology consists of a given finite ontology ordered by the ISA relation called the skeleton ontology, and a set of production rules (cf. generative grammars) that allows for production of compound concepts. We represent such compound concepts in the ontology language ONTOLOG. In this language, compound concepts are represented as conceptual feature structures in the form of $c[r1:c1]$, where attributions consist of pairs of relations and concept arguments that function as conceptual restrictions on the core concept. However, the generative ontology should not admit arbitrary combinations of relations and concepts: We thus propose to introduce ontological affinities that may specify ontologically admissible ways of combining concepts.

This dissertation is not about information search per se, but rather, it is about how to get from a textual form to a corresponding conceptual form, and more specifically, it is about how we get from prepositional form to conceptual form:

Prepositions are highly polysemous, i.e. they can potentially denote many senses, and in addition, each sense can be expressed by a number of prepositions. This means that a many-to-many relationship exists between prepositional forms and conceptual forms. People, however, are seldom in doubt of the intended meaning of a preposition in context and, thus, their high degree of polysemy/synonymy does not appear to create noise to any problematic degree in human discourse. Thus, what this dissertation seeks to examine is what it is in the context of prepositions that gives us enough information to disambiguate prepositional senses.

1.1 Research Question

The main objective for the research described in this dissertation is to uncover the senses of Danish prepositions. The senses, in this context, are semantic relations denoted by prepositions. In order to give an account of this topic, we first need to define the essence of the class of prepositions. Next, we must define a set of possible relations that prepositions can denote, and finally, we discover the senses that prepositions in Danish texts in fact

do express, and infer ontological affinity rules for these. Thus, the questions that this dissertation seeks to answer are the following:

1. What is an adequate definition of the class of prepositions?
2. Which semantic relations can a subset of Danish prepositions denote?
3. Can we infer ontological affinity rules for the relations denoted by a subset of Danish prepositions from an annotated corpus?

The first part of the research question includes a survey of how the class is defined in a number of sources on the Danish language. The definitions include syntactic, morphological and semantic criteria for a definition of the class. We conclude by phrasing our own definition of the class; a definition that adequately defines the class of prepositions for our further treatment. For the second part of the research question, we first select a subset of 14 Danish prepositions from their common inclusion in a number of Danish dictionaries, and produce a preliminary list of possible senses based on their definitions in the selection of dictionaries. Subsequently, a large number of corpus evidences are analyzed, and as a result, we produce a final list of possible senses for the subset of prepositions. For the third part, we mark up a subpart of the Danish general language corpus Korpus 2000 with various features, including ontological types and semantic relations based on our analysis of prepositional senses, and feed this dataset to a machine-learning algorithm. As a result, we get a set of rules; these rules are transformed into a dictionary of prepositions that express affinities as ontological type-restrictions on arguments of semantic relations denoted by prepositions. Thus, the essential results of this research are knowledge about the relations that this subset of Danish prepositions can denote as well as a representation of the ontological affinities for these relations.

1.2 Outline of the Dissertation

This dissertation consists of 7 chapters. The present Chapter 1 introduces the following 6 chapters: Chapter 2 *Prepositions* gives a survey of how the class of prepositions is defined in a number of sources on the Danish language, and provides an adequate definition for the further treatment of

Introduction

the members of the class. Chapter 3 introduces the notion of *Ontologies* and gives an account of what this is. Special emphasis is placed on a specific type of ontologies, namely lexical ontologies or wordnets, which are the types of ontologies that are used in the experiments described in chapters 5 and 6. Subsections are devoted to a description of the theories behind as well as to the structure and contents of each of the wordnets Princeton WordNet, EuroWordNet and DanNet. Chapter 4 *Linguistic Expressions, Concepts and Semantic Relations*, gives important background knowledge of how we view various aspects of our ontology-based framework. We assert that concepts exist in the minds of people and are abstract ideas of entities in the world, and a relation is the conceptual glue that binds concepts together in discourse, and we describe the combination of a conceptual level and the expression level as a sign, the duality of which is crucial for our treatment of text and mapping into a generative ontology. We introduce the ontology language ONTOLOG and the notion of generative ontologies, as well as our aim to construct compound concepts reflecting the conceptual content not just of individual words but of text chunks and that ideally, we represent the conceptual content of sentences as compound concepts. In addition, we give an account of our different treatments of the relation denoting word classes verbs and prepositions where relations denoted by verbs are reified and where relations denoted by prepositions are treated as associative relations, as well as of our treatment of arguments where the conceptual content is a plurality.

Chapters 5 and 6 describe experimental work: Chapter 5 *A Machine Learning Approach to Disambiguation of Semantic Relations* describes our first experiments in using machine learning for disambiguation of semantic relations denoted by prepositions. The chapter reflects a body of work that was carried out in collaboration with Thomas V. Terney within the framework of the OntoQuery project¹. This first experiment concerns a domain-specific corpus which includes texts from the domain of nutrition, where sentences containing syntactic structures in the form of NP-PREP-NP are annotated with information about e.g. semantic types for heads of the noun phrases and relations denoted by the prepositions, and used as input to a machine-learning algorithm. The result of this is a set of rules. We situate the employed notion of relation disambiguation in relation to word sense

¹ <http://www.ontoquery.dk>

disambiguation and describe the corpus, the set of semantic relations used, and the different levels of our annotation process. Finally, we describe the execution of the experiments, including the applied algorithms, and give an analysis of some of the produced rules that predict the semantic relation of a preposition based on e.g. the ontological types of its arguments. Chapter 6 *Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions*, describes a second and larger experiment carried out by the author alone. This experiment concerns a general language corpus for which the same type of syntactic structures as in the former experiment are annotated. Initially, we describe the selection of a subset of Danish prepositions from their common inclusion in a number of Danish dictionaries, and produce a preliminary list of possible senses based on their descriptions in the selection of dictionaries. Subsequently, we describe how a large number of corpus evidences are analyzed, and a resulting final list of possible senses for the subset of prepositions. We then describe the mark-up of a subpart of the Danish general language corpus Korpus 2000 with various features, including ontological types and semantic relations based on our analysis of prepositional senses. Further, we describe the application of machine learning to the dataset and the resulting rules, which are analyzed, evaluated and transformed into a dictionary of prepositional senses. Finally, Chapter 7 *Conclusion*, concludes and points to future work.

Appendix A *A Rule-based Dictionary of Danish Prepositions*, contains a dictionary of prepositional senses for a selection of 14 Danish prepositions, which is the result of the experiments described in chapter 6. In this dictionary, given a preposition and a sense, ontological affinities are expressed as restrictions on the ontological types of the arguments.

1.3 Notational conventions

Throughout this dissertation, we represent strings in *italics*, concepts in SMALL CAPS and semantic relations in CAPITALS. In chapter 4, a sign, i.e. the combination of a linguistic expression and its associated concept, is represented in **bold**.

Danish examples are normally followed by an English translation equivalent in plain text, either following the example and in parentheses or below the examples.

Introduction

Chapter 2

Prepositions

Prepositions are small, yet they have great power of expression. These small words express relations between phrases or clauses, and though limited in number, they are not limited in the number of senses they can express. In her essay *Silken, rummet, sproget, hjertet* (Christensen, 2000), Inger Christensen poetically describes the importance of these modest words:

Alle præpositioner er nærmest usynlige. De holder sproget oppe på samme måde som rummet bærer planeterne. I deres begrænsede antal, op, ned, ud ind, over under, o.s.v holder de bevidstheden i samme slags bevægelse som verden. De sætter alle substantiver på plads i forhold til hinanden og bekræfter os stiltiende i, at vi på forhånd er båret oppe af et uudtømmeligt stort, altid eksisterende sammenligningsgrundlag.
(Christensen, 2000)

All prepositions are practically invisible. They hold up the language in the same way as space carries the planets. In their limited number, up, down, out, in, over, under, etc. they keep our consciousness in the same kind of movement as the world. They put all nouns into order relative to each other, and silently confirm our belief that, from the outset, we are supported by an inexhaustibly large, always existing standard of comparison.
(my translation)

One can wonder why most languages which have prepositions have such a limited number, which makes them highly polysemous. Why don't we simply make up some new prepositions – one for each relation, one could ask. The principles of economy of language, however, prevent language users from inventing new words when words already exist that can express

Prepositions

what we wish to communicate. And the fact is that we are seldom in doubt of the intended meaning of a preposition in context, and thus, their high degree of polysemy does not create noise to any problematic degree in human discourse. What this dissertation seeks to examine is what it is in the context of prepositions that gives us enough information to disambiguate prepositional senses. Such knowledge is needed in order to disambiguate prepositional senses in automated text analysis for information search purposes.

Whether or not prepositions should be treated as a separate word class is not evident; The renowned Danish linguist Otto Jespersen (Jespersen,1924) argues that prepositions should *not* be viewed as a separate word class, but rather as members of the class of uninflectable words or particles, comprising adverbs, prepositions, conjunctions and interjections. He asserts that many of the words in these four classes overlap in form, but are traditionally classified as prepositions, adverbs or conjunctions according to their function in a given text. He compares the functions of particles to verbs that are sometimes transitive, and sometimes intransitive (e.g. *sing* in *he sings* and *he sings a song*) and, despite these differences, no one would assign them to different word classes. Particles are sometimes complete in themselves, and sometimes they must be followed by a complement (or object), in order to be complete. Thus, the ‘transitive particles’, as we may call them, or particles that need complementation, are the types of words that we are concerned with in this work.

2.1 Word Classes and Criteria for Word Classification

Different features can be used in the classification of words into word classes: *syntactic*, *morphological*, and *semantic* features. Morphological features have to do with the form of the written or spoken word, and syntactic features have to do with how the word behaves in discourse: which other types of words does it combine with? Morphological features thus concern the internal structure of words, and syntactic features concern the external structure. Semantic features have to do with the meaning of the word.

Morphological features are defined by inflectional and derivational behavior. Inflectional features initially concern whether or not a word can be inflected, and if can, which inflectional pattern does it belong to.

Derivational features concern whether or not other word classes can be derived from the word through affixing, or if it itself contains derivational affixes.

Syntactic features are defined by the ability of a word to modify or to be modified by other types of words in phrases or its ability to fill a syntactic function in a sentence.

Semantic features concern the kinds of entities in the world that the word denotes.

From a morphological viewpoint, word classes are typically being divided into two portions: those that can be inflected and those that cannot. Syntactic and morphological features differ from language (family) to language (family). Semantic features, in principle, do not differ from language (family) to language (family).

From a semantic viewpoint, word classes are often divided into general types: Content words and function words. Content words denote concrete or abstract entities in the world, and/or denote relations between them. Function words do not denote entities in the world, but have a purely syntactic and/or semantic function.

Roughly speaking, content words can be divided into object-denoting classes and relation denoting classes. The object-denoting classes count e.g. nouns and pronouns, and the relation-denoting classes count e.g. verbs, conjunctions and prepositions. To the class of function words belong e.g. determiners and numerals. However, the division into these two portions is not a straight cut; some words are what we could call object relational, i.e. words that are at the same time object denoting and relation denoting. Examples of such types of words are relational nouns, that apart from pointing out entities in the world also denote inherent relations to other entities; e.g. the relational noun *member* denotes a set of entities in the world, of which we know that they are members of something. Between such an entity and 'something', a membership relation exists. Other examples of words that can be seen as object-relational are adjectives and verbs, which denote relations, but at the same time point out events or phenomena as relata. E.g. 'a red pony', where the adjective 'red' denotes a

Prepositions

(color) relation between the entity that the word ‘pony’ points out and the phenomenon ‘redness’.

In (Brøndal, 1928), The Danish linguist Viggo Brøndal poses a universal theory of word classes, in which he suggests categories that are sufficient for defining a system of word classes for any language. This view is also adopted by (Diderichsen, 1946). The main categories, which may be combined for defining a specific class, are:

- * Relator classes (r): Classes of connectors, whose function is to express a connection or a relation, and cannot in addition express an object, or be descriptive. Pure connector classes are prepositions and conjunctions.
- * Descriptor classes (d): Classes of descriptors, whose function it is to describe objects (e.g. true adverbs, adjectives, verbs).
- * Relatum and descriptum classes (R and D): Classes of objects that can be related to or described (e.g. Proper nouns, numerals, pronouns)

We will return to Brøndal’s definition and description of a specific relator class, namely the class of prepositions, in the following section 2.2.

On a similar note, (Spang-Hanssen, 1996) describes the notion of word classes as follows: Grammatical categories to a certain extent reflect our understanding of the world, which becomes evident by the fact that our categorization of words corresponds to the categories we use in logic or computer programs:

Objects: nouns, pronouns
Quantifiers: articles
Properties: adjectives
Relations: verbs, prepositions
Processes: verbs
Connectors: conjunctions

Definitions of word classes can be extensional or intensional; extensional word class definitions enumerate the inventory of words in a language that

belong to the given word class, and intensional word class definitions use one or a combination of the features described above in order to describe the characteristics of the words belonging to the given word class. Extensional word class definitions are by definition language specific and intensional word class definitions may be language specific or not, depending on the criteria used in the definition.

Word classes are normally characterized as being open or closed. In a synchronic view, open word classes may have new members added, and closed word classes may not. In a diachronic view, however, language change causes all word classes to change. The class of prepositions is a closed word class; thus, new words are never, or rarely, added to this class, and it can, in principle, be enumerated. The extension, or the members of the class, varies from reference work to reference work. New senses or new uses of existing prepositions, however, frequently occur – probably mostly as a result of influence from other languages. These years, influence on the Danish language primarily comes from English, and examples of transfer of lexical items in expressions are seen. Probably, as an example of this phenomenon, an increasing use of the preposition *for* in collocations with the adjective *klar* (ready) can be observed. Traditionally, the typical preposition collocating with *klar* is *til* (as in *klar til skole* (ready for school), *klar til weekenden* (ready for the weekend), etc.), but in examples concerning transfer of football or handball players from one club to another, examples such as ‘*Næste svensker klar for FCK* (next Swede ready for FCK)’ are observed.

In the following sections, we provide examples of how different sources define the class of prepositions based on different criteria.

2.2 Views on the Essence of the Class of Prepositions

In the following sections, we look at how prepositions have been treated in a selection of works on the Danish language. Some sources, e.g. (Brøndal, 1940) and (Jespersen, 1924), are not exclusively concerned with the Danish language, however, since we are specifically concerned with Danish prepositions, we generally reference Danish examples in the following. While most examples derive from the sources they are given in connection with, some may be modified slightly, and many include our translations. Because this dissertation is concerned with written forms of Danish only,

Prepositions

views that include phonological features such as stress are generally not referenced from the described sources. Where a translation is not provided for an example in the given source, it has been translated into English by this author. In some cases, examples of prepositions that are given without context are not translated. This is due to our conception that, in general, prepositions out of context do not have a one-to-one translation between languages, but that they only do so in context. Some examples for which a translation does not include the exemplary syntactic or morphological structure, both a literal and a free translation are provided.

2.2.1 About a Quantitative Delimitation

(Togoby) notes that the class of prepositions is smaller than the class of finite inflectional forms, but not infinite. He claims that there are 20-30 prepositions in all. (Spang-Hanssen, 1996) notes that the languages have only about 20 prepositions that have to cover all possible conditions. Thus, they must be able to bend and stretch. He also notes that they all have one or more prototypical uses, which other uses can be related to. As described below in section 2.2.2.1, (Brøndal, 1940) notes that Danish has 18 true prepositions, but he has no mention of how large the class of prepositions in general is. (Diderichsen, 1946) does not say anything about the size of the class, but does list the 16 *most important prepositions* as a subset of Brøndal's true prepositions. (Christian Becker-Christensen & Widell, 2003) note that prepositions form a small, closed word class, but do not say anything further about the number of members of the class.

2.2.2 About a Qualitative Delimitation

In the following sections, we give an account of how prepositions are described from a morphological, a syntactic and a semantic point of view, as well as how members of the class of prepositions are distinguishable from other word classes.

2.2.2.1 Morphology

Most sources, including (Hansen & Heltoft, 2003), (Per Anker Jensen, 1985) and (Allan, Holmes, & Lundskær-Nielsen, 1995) state that prepositions are uninflectable words. This is also true of (Jespersen, 1924); however, as noted above and described in more detail below in section 2.2.2.4, he does not treat prepositions as an independent word class but rather as belonging to the group of particles. All particles, he states, are

invariable, apart from some adverbs that are able to form comparative and superlative forms. (Nielsen, 1995) concurs with Jespersen's view, and does not favor a view of prepositions as an independent word class. She prefers to delimit word classes by morphological features alone; this way the classification of a word can be treated separately from its syntactic function in a given construction. This can be done only by viewing the word's form, which then remains independent of the function.

(Diderichsen, 1946) says that prepositions normally do not inflect, but from some prepositions, adjectives in the comparative and superlative forms can be derived (*over* : *øvre* : *øverst* | *under* : *underst*) (*over* : upper : uppermost | *under* : lowermost), and from some prepositions, relational adverbs (*over* : *ovre* : *oven* | *af* : *a'e*) (*over* : over : above | *of* : off), cf. (*Proppen er a'e*) (The cork is off), can be derived.

(Allan et al., 1995) note that prepositions may be divided into four types:

1. Simple prepositions

Simple prepositions consist of a single morpheme. This type includes the most common prepositions: e.g. *af*, *efter*, *fra*, *før*, *i*,...

2. Compound prepositions

Compound prepositions are written in one word, but consists of two roots and may be of the form:

i) preposition+preposition

e.g. *igennem* (through), *imellem* (between), *imod* (against)

ii) preposition+noun

ifølge (according to) (only example)

Prepositions

3. Complex prepositions

Complex prepositions consist of two or more words forming a semantic unit which has a function similar to that of a preposition:

i) Adverb+preposition

This type is extremely common and forms an open-ended group: Members of this group consist of a positional/directional adverb and a preposition. E.g. *inde i, ned på, op ad, ...*

Some of these positional/directional adverbs may also function as prepositions (*bag, foran, inden, om, over, uden*), whereas the rest cannot. Examples of adverbs function as prepositions are:

<i>Han</i>	<i>stod</i>	<i>bag</i>	<i>døren</i>
He	was standing	behind	the door
<i>De</i>	<i>kiggede</i>	<i>over</i>	<i>muren</i>
They	looked	over	the wall

For those with double forms, *om/omme, over/ovre*, only the directional version can be used as a preposition.

(Obs. Danish distinguishes complex prepositions and equivalent compound adverbs: E.g. *opad* – adverb, whereas *op ad* – preposition)

ii) Preposition+noun+preposition

A large number of examples follow this pattern, where *af* is the most common second preposition:

<i>af</i>	<i>hensyn</i>	<i>til</i>
(out) of	consideration	for

<i>i</i>	<i>mangel</i>	<i>af</i>
for	want	of

<i>på</i>	<i>grund</i>	<i>af</i>
on	account	of

iii) Preposition+og+preposition

Some coordinated phrases consisting of a preposition+og+another preposition that go together as an idiom:

(stå)	<i>af</i>	<i>og</i>	<i>på</i>
(get)	on	and	off

<i>for</i>	<i>og</i>	<i>imod</i>
------------	-----------	-------------

Prepositions

for and against

fra *og* *med*

from and including

iv) Discontinuous prepositions

A few examples exist where the prepositional complement is positioned between two prepositional elements:

ad ... *til*

in the direction of

for ... *siden*

ago

fra ... *af*

from ... onwards

4. Prepositions derived from other word classes, e.g.:

i) Participle forms

angående (concerning)

fraregnet (not counting)

- ii) Words which function as adverbs and/or conjunctions can occur as prepositions:

bag (behind)

efter (after)

indtil (until)

- iii) Finally, the adjective *lig* (like) is sometimes used as a preposition

(Mikkelsen, 1911) notes that combinations of prepositions and adverbs, e.g. *neden for* (below), *nær ved* (close to), *bag på* (on the back of), *langs med* (along) are to be considered prepositions. Also, combinations of prepositions and nouns function as prepositions, e.g. *for Ns skyld* (for N's sake), *på Ns vegne* (on behalf of N), *ved siden af* (next to), *på denne side af* (on this side of), *ved hjælp af* (by means of), *i sammenligning med* (in comparison with), *i tilfælde af* (in the event of), etc.

(Brøndal, 1940) does not accept a definition of prepositions on the grounds of morphology; Even though the class of prepositions has been defined as uninflectable since classical antiquity, this is not true, he claims. Some Danish prepositions do in fact inflect, as exemplified by *over/ovre om/omme*. Also, he rejects the inflection criterion as such as a means of classifying words, as some languages (e.g. Chinese) have no inflection at all, but have distinct word classes notwithstanding. Further, a large group of diverse words, named *indeclinabilia* or particles, lack inflection and, thus, this criterion does not supply sufficient information in order to distinguish prepositions from adverbs, conjunctions, and interjections.

Prepositions

With a reference to Descartes' methodological rules (cf. (Descartes, 1903)), by which e.g. "*commencing with objects the simplest and easiest to know, I might ascend by little and little, and, as it were, step by step, to the knowledge of the more complex;*" (Descartes, 1903), Brøndal divides prepositions into true (*ægte*) and false (*uægte*) prepositions in order to start with the true prepositions. This division process consists of three steps, where he assumes a set of possible prepositions as a point of departure:

1. Elimination of prefixes.

This step involves a separation of prefixes from the class of prepositions. A kinship exists between prefixes and prepositions, and this close kinship has historically resulted in a prevailing mix-up between the two categories, Brøndal claims. He describes the kinship as follows:

- a) True prepositions can often appear as prefixes, or more correctly as preverbs, as e.g. in *for-stå* (under-stand).
- b) Often, double forms exist for prepositions and prefixes. Such double forms co-exist, where one is an independent preposition and the other is a bound morpheme, as e.g. the French preposition *pour* and the equivalent prefix *pro-*.
- c) Many prefixes are originally prepositions which were earlier used as preverbs, but these have now coalesced with their verb without being fully absorbed, as e.g. German *er-* and *be-*.

Brøndal thus assumes the following distinguishing features for prepositions and prefixes, respectively:

- * Prepositions can exist as independent word forms or as preverbs. They maintain their form if they appear in a preverbal position.
- * Prefixes are bound morphemes, and thus cannot exist as independent word forms. A prefix in the form of a bound morpheme may be overlapping in meaning and share its etymology with a preposition.

After all prefixes have been eliminated from the set of possible prepositions, we have a set of word forms consisting of prepositions only. However, some of these are considered false prepositions that should be eliminated:

2. Elimination of false prepositions.

This step involves a separation of *false* prepositions from the class of *true* prepositions.

Since prepositions are defined as the simplest expression in a language of a relation, they should be very simple in structure. For this reason, compound or complex word forms should be excluded as *false prepositions*. Amongst these are:

a) Sayings

The number of sayings, or free combinations of words, is unlimited and they may be formed of different kinds of words and be of different syntactic structure. Thus, they do not constitute a specific type, nor can they be included in any system. Examples of this type are *på grund af* (because of), *i anledning af* (on the occasion of), *med hensyn til* (with regard to), etc.

b) Compound word forms

Compound words such as e.g. *over-for* (opposite), *i-mod* (against), etc. should not be considered actual prepositions because they consist of two relatively independent parts; either as a combination of two complete word forms (or unbound morphemes), or as a combination of a complete word form and an affix.

c) Words from other word classes

Words from word classes other than prepositions should be excluded. These may be:

1) Unrelated classes such as nouns (e.g. German *kraft* (by virtue of), *trotz* (despite), *laut* (according to)), adjectives (e.g. English *round*, Danish *langs* (along)), and true adverbs (e.g. French *près* (close)).

2) Related classes such as participles (e.g. French *pendant* (for), *durant* (during)) and situatives (e.g. Danish *oven* (above), *neden* (below), German *ausser* (beside(s))).

Brøndal notes that all of these types may transfer into true prepositions over time, or (possibly homonymic) double forms may form where one remains a situative and the other becomes a preposition, e.g. Italian *su* (up/on), which is considered a

Prepositions

situative when it correlates with *giú* (down), and a preposition when it correlates with *a* (in).

Finally, some true prepositions are *reintegrated* into the class:

3. Reintegration of true prepositions.

While the class of prepositions in some treatments is too broadly defined, it is too narrowly defined in others, Brøndal notes. For example, when prefixes are not treated as true prepositions, when postpositions are treated as an independent class, or when specific uses of prepositions are treated as conjunctions, adverbs, adjectives or nouns. The following ‘misclassifications’ should be considered true prepositions:

a) ‘Conjunctions’

E.g. *for* (for) when used as a conjunction, and English *to*, cf. German *zu*, Swedish *till*, when used as infinitive markers, they should still be considered prepositions.

b) ‘Adverbs’

When post-positioned and governing as in *som jeg sørger over* (which I mourn (over)), post-positioned and not governing as in *se efter!* (look!), or pre-positioned and descriptive as in *for stor* (too big), the prepositions should still be considered prepositions.

c) ‘Adjectives’

For predicative uses as in *er du med?* (do you follow?) or *is it over?* or attributive uses as in *the off man*, the prepositions should still be considered prepositions.

d) ‘Nouns’

When used as a noun as in *an off*, French *le pour* (the pro) or *le contre* (the con), the prepositions should still be considered prepositions.

The result that Brøndal arrives at is that Danish has 18 true prepositions, namely *til, på, for, efter, over, under, ad, mod, om, i, gennem, mellem, af, ved, uden, fra, hos* and *med*. Brøndal, however, makes no claim that his system only applies to true prepositions or to any other relator class, for that matter. We note, however, that even though Brøndal states that a definition of the class of prepositions is not possible on the grounds of morphological features, he does include such features in the delimitation of the class - thus,

morphological features may be grounds for exclusion from a class, but not for inclusion.

2.2.2.2 Syntax

(Brøndal 1940) basically finds definitions of word classes based on syntactic criteria impossible. A class always has more possible sentence functions, which do not combine into one definition, he claims. Attempts at such definitions lead to co-classification of formations of very dissimilar character.

Others do, however, find it apt to describe syntactic features that characterize the behavior of prepositions and prepositional phrases:

Concerning complementation, (Jensen 1985) and (Christian Becker-Christensen & Widell, 2003) note that prepositions cannot occur alone as a sentence constituent. They always combine with one or more words, their complement, to form a prepositional phrase. They can combine with a noun, a pronoun, a sentence or an infinitive construction, and these are then related to other words in the sentence via the preposition. In addition, (Nielsen, 1995) notes that the preposition is characterized by the fact that it always precedes its complement, hence the name preposition.

(Hansen & Heltoft, 2003) say that prepositions have two main syntactic-semantic functions:

1. In the first function, they govern a nominal phrase, e.g. *over byen, over tolv*; In this case, the relation between a preposition and its complement resembles that between a verb and a direct object. They may also occur without a complement, in which case they are said to have an adverbial function (e.g. *hatten skal på*), but more correctly, the function is comparable to that of a verb with an omitted object. This often occurs in predicative constructions.
2. In the second function, in addition, the preposition itself is governed by its nominal complement, in which case neither the preposition nor the complement are omissible (e.g. *skyde på pianisten, *skyde på/*skyde pianisten*). Prepositions, in this case, resemble governing conjunctions (*om, at*).

In (Allan et al., 1995), a fairly detailed account of the syntactic features of prepositions is provided:

Prepositions

It is stated that the most important function of the preposition is to form a (grammatical) relation between two entities. One is represented by the prepositional complement, and the other by a different clause constituent. Such a construction is called a prepositional phrase.

The different types of prepositional complements may be a noun/pronoun, an adverbial (incl. a prepositional phrase), an infinitive or a subordinate clause.

a) Preposition+noun /pronoun

For the majority of cases, the complement is a noun or a pronoun:

For min onkel / for ham
For my uncle / for him

b) Preposition+adverbial

Adverbials such as adverbs and prepositional phrases can occur as prepositional complements:

For alltid
For ever

Tak for i dag
Thank you for today

c) Preposition+infinitive

It is also common to have an infinitive as a prepositional complement. In such cases, the subject of the finite verb is also the subject of the infinitive.

Hun har travlt at lave mad
She is busy cooking

Jeg tænker på at holde op
I'm thinking of stopping

d) Preposition+subordinate clause

i) A preposition can govern an at-clause as in e.g.:

Hun sørgede for, at jeg kom hjem
She saw to it that I got home

ii) Or it can govern other types of subordinate clauses, e.g. interrogative clause introduced by *om* (whether) or by an interrogative hv-word:

Det afhænger af, om vi kan få støtte
It depends on whether we can get support

The position of the preposition may be:

a) Preposed

The vast majority of prepositions precede their complement:

Foran huset
In front of the house

b) Discontinuous

i) A small number are discontinuous, i.e. they consist of two parts with the complement positioned inbetween:

Ad byen til
Towards the town

Prepositions

For fem år siden
Lit: For five years ago
Five years ago

- ii) For a small subgroup of the discontinuous prepositions, the second element is a noun. For this group, the complement is usually a genitive form or a possessive pronoun:

For hendes skyld

For her sake

På firmaets vegne

On behalf of the company

c) Postposed

- i) In some idiomatic expressions, the preposition is placed after its complement:

Dem foruden
Lit: Them apart
Apart from them

Hele landet over
Lit: The whole country over
Throughout the country

ii) Some prepositions constitute the second element in compound adverbs. These are characteristic of formal written Danish and most have the locative adverbs *der-* or *her-* as first element, e.g. *derefter* (thereafter) and *hertil* (to which).

iii) In some cases, the preposition occurs in clause-final position. The following are examples of this structure:

(1) When the complement is moved to front position in the clause for emphasis. This type is common when the complement is a personal pronoun:

Hende kan man regne med
 Lit: Her can you count on
 You can count on her

(2) In interrogative and relative clauses, where the prepositional complement (or part of it) is an interrogative/relative pronoun. In these cases, the pronoun may be omitted.

Landet, (som) vi bor i
 The country (that) we live in

(3) In some fixed expressions, especially when the verb and the preposition form a semantic unit (a prepositional verb), the preposition may appear without an overt complement:

Han skal sidde efter
 He's got a detention

Prepositions

Also (Mikkelsen, 1911) provides a thorough account of various aspects of the syntactic features of prepositions and prepositional phrases. The following sums up the main points of this account.

Prepositional phrases may modify:

1. Verbs, e.g.:

Rejse til Amerika
Travel to America

2. Nouns (or other nominal uses of words), e.g.:

Et slagsmål på torvet
A fight at the square

3. Adjectives (or other adjectival uses of words), e.g.:

Rig på penge
Rich in money

4. Adverbs, e.g.:

Han bor her i gaden
He lives here in this street

The prepositional phrase may have the following functions in the sentence:

1. As a subject:

Over tusinde mennesker var til stede
Over a thousand people were present

2. As an object:

Jeg tabte henimod hundrede kroner
I lost about a hundred kroner

3. As a subject complement:

Vi var henved hundrede stykker
 We were about a hundred

4. As an indirect object

den amerikanske præsident rækker
 Lit: the American president gives
over hundred mennesker hånden
 over a hundred people his hand
 the American president shakes over a hundred people's hands

5. Object complement

Han følte sig i besiddelse af store evner
 Lit: He felt himself in possession of great abilities
 He felt he possessed great abilities

Also the prepositional phrase may function as a size determiner or as a state expression:

Han løb over hundrede skridt tilbage
 He ran over a hundred steps back

I sin nød henvendte han sig til sine venner
 Lit: In his trouble appealed he himself
 to his friends
 In his troubles, he appealed to his friends

Usually, the prepositional phrase complements are nouns, nominal pronouns, infinitive clauses, sentential clauses or interrogative clauses. They may also frequently be quoted expressions, adjective or a participle forms used as nouns, genitive forms used in a neuter or plural sense or adjectival numerals or pronouns used as nouns.

Prepositions

Not frequently, they are nominal uses of adverbs, nominal uses of prepositional phrases, figurative relative clauses, conjunctive clauses or conjunctive phrases.

Sometimes, the prepositional phrase is decomposed, i.e. the complement and the preposition are separated. This may happen when:

- a. The complement is an interrogative or relative pronoun.

Hvem kan du stole på?
Lit: Who can you trust on?
Who can you trust?

- b. The complement is emphasized by promoting it to sentence initial position:

Den mand jeg talte med i søndags
The man I talked with on Sunday

- c. In main clauses following a subsidiary clause with *desto* or *jo*, or comparative clauses with *jo*:

Jo ældre han bliver,
Lit: The older he gets,
desto flere galskaber finder han på
the more madness finds he on
The older he gets, the madder his ideas

- d. In some comparative clauses with *sådan* – (som), *så* – som or *så*, that have the sense of a causal clause or an adversative causal clause, or after an expressions that denotes a judgement:

Jeg må hjælpe ham, så stor
Lit: I must help him, such great
en fare som han er i
a danger that he is in
I have to help him, since he is in such a great danger

- e. Sometimes when the complement is or contains *ingen*:

Hun vil ingen tale med
 Lit: She wants nobody talk to
 She does not want to talk to anybody

Sometimes, the prepositional phrase is incomplete, i.e. the complement is omitted. This may happen in cases where e.g.:

- a. The preposition is found in a relative clause that modifies a noun which is the actual complement:

Den båd, som jeg var i, kæntrede
 The boat that I was in capsized

- b. In some comparative clauses with *end*:

Du bruger mere end jeg kan være tjent med
 Lit: You spend more than I can be served with
 You spend more than what serves me

- c. In the first of two coordinated prepositional phrases where the complements are identical, the first mention of the complement may be omitted.

Hensigten med, og nytten af denne bestemmelse
 The purpose of, and use of this provision

- d. When a prepositional phrase modifies the last of two coordinated verbs, and when the first verb denotes an action that is a prerequisite for, or initiates the action denoted by the second verb. The actual complement is the object of the first verb.

Han tog en stok og støttede sig på
 Lit: He took a cane and leaned himself on
 He took a cane to lean on

- e. When the actual complement has a different function in the sentence, in this case the subject, and especially in connection with infinitives:

Prepositions

Bogen havde han taget for at læse i
Lit: The book had he taken for to read in
He had taken the book to read it

- f. When the actual complement is a word in a preceding sentence or clause:

Han tog flasken, men der var ingenting i
Lit: He took the bottle but there was nothing in
He took the bottle, but there was nothing in it

Expansions of prepositional phrases, indirect objects, size determinations, and adverbials are normally positioned in front of the phrase:

Han bor tre mil fra byen
He lives three miles from the city

However, in some cases, the expansion is positioned after the phrase:

I morgen tidlig
Lit: In morning early
early tomorrow morning

A preposition may be omitted in the second of two coordinated prepositional phrases, except after *men*, *som* and *end*:

Han er rejst fra Neapel og Rom
He has traveled from Naples and Rome

Du skal ikke møde i kjole men i frakke
Lit: You must not attend in tailcoat but in coat
Do not attend in a tailcoat, but rather in a coat

Other examples of prepositions that may be omitted are *af* which is sometimes omitted after another prepositional phrase, *i* and *på* may be

omitted in connection with *være*, *fra* may be omitted in connection with people names and origin, and *ved* may be omitted in exclamations. Also, in front of an infinitive, and some combinations of a verb and a noun or an adjective the preposition may be omitted:

	<i>Han</i>	<i>fik</i>	<i>lov</i>		<i>(til)</i>	<i>at</i>	<i>rejse</i>
Lit:	He	got	permission		(to)	to	travel
	He was permitted to travel						

Also, the preposition *for* may be omitted in front of the adverb *for*, and the preposition *om* may be omitted in front of the interrogative conjunction *om*:

<i>Spørg</i>	<i>(om)</i>	<i>om</i>	<i>han</i>	<i>kan</i>	<i>komme</i>
Ask	(about)	if	he	can	come

2.2.2.3 Semantics

(Christian Becker-Christensen & Widell, 2003) says that prepositions basically denote a localization in space or time, as in e.g. *Bogen ligger på bordet* (The book is on the table) or *De kommer på mandag* (They will be here on monday). However, they also denote more abstract conditions as in *Lad os se på den* (Let us have a look at it) and *Jeg tænker på om det bliver regnvejr* (Lit: I think about whether it is going to rain).

In (Hansen & Heltoft, 2003), it is said that prepositions denote abstract relations and prototypical place relations.

(Diderichsen, 1946) says that prepositions denote relations, without at the same time denoting an object (a *relatum*), such as verbs do. Thus, they are pure relators.

According to (Brøndal 1940), the only criterion that will define the class of prepositions by a positive characterization, is a definition based on semantic or conceptual criteria; i.e. a definition that defines the criteria for inclusion in the class, rather than criteria for exclusion from the class, or by criteria that do not define at all, cf. the syntactic and morphologically based criteria. The class of prepositions in general can alone be defined as expressing relations, and the individual members of the class must be defined by a sum of special relations.

As said by Brøndal, this definition is precise or specific in that it assigns general and specific relations to the class itself as well as to its individual members. It is fruitful, in that it both requires and allows for synonymics

Prepositions

work: it requires differentiation between authentic and unauthentic prepositions, definition of the class' relation to other classes and arrangement of the prepositions in their mutual semantic correlations. We return to Brøndal's synonymics system for relations later in this section.

(Jespersen, 1924) says that while prepositions should be included in dictionaries, their proper place is in a grammar that deals with the *general facts* that should be mentioned in connection with them. Such general facts include syntactic behavior, e.g. the ability to occur with certain types of complements, e.g. infinitive clauses, combinations of two prepositions as in *From behind the bush*, etc., but grammar also has to deal with general facts concerning the way prepositions express rest at a place or movement to and from a place, the relation between the local and temporal significations of the same preposition – and even more important is the uses of a preposition where it loses the local or temporal signification and moves into the category of colorless or pale words or auxiliaries, as in: *The father of the boy*, *The scoundrel of a servant*, etc. In some cases, however, it is doubtful or even arbitrary what is to be treated in a grammar, and what is to be treated in a dictionary. While grammars should deal with general facts and dictionaries with special facts of a language, these two fields sometimes overlap and certain things should be treated in both.

(Allan et al., 1995) include a large section with detailed descriptions of the use (i.e. the meaning) of 12 common Danish prepositions, as well as a brief survey of some other Danish prepositions. The purpose of the section is to provide distinctions relevant for translation purposes.

Nearly all prepositions can appear with spatial and temporal meaning. Spatial meaning can be subdivided into literal/physical (*bogen ligger på bordet*) and figurative (*vi er på randen af en katastrofe*). If a figurative meaning becomes too far from the original literal meaning, it becomes opaque (*på må og få*) and is said to have abstract meaning.

As an example, the following is a summary of the treatment of the preposition *på*.

1) Space

The spatial meanings of *på* range from 'on top of' something, whether horizontal or vertical, to 'at' or 'in' areas, buildings, institutions, etc.

a) Institutions and places in terms of function:

På apoteket (at the chemist's)

På bunden (at the bottom)

b) Areas:

På gaden (In the street)

c) Islands:

På Falster (on Falster)

d) Surfaces:

På bordet (on the table)

På højre/venstre side (on the right/left)

2) Time

a) A point in time:

På den tid (at that time)

På lørdag (on saturday)

Prepositions

b) Duration

På indicates how long a given action takes:

Gøre noget på meget kort tid (do something in a very short time)

3) Measure

Et barn på tre år (a child of three)

En lejlighed på fire værelser (a flat with four rooms)

4) Genitive

In some cases, a prepositional phrase with *på* can replace an –s genitive:

Farven på huset (the colour of the house)

Den varmeste tid på året (the warmest time of the year)

5) Manner

På is here mainly found in idiomatic expressions

På dansk (in Danish)

Gøre noget på skråmt (pretend to do something)

På ny (anew, again)

6) Attached to nouns

Abonnement på (subscription to)

Angreb på (attack on)

7) Attached to verbs with strong stress

i) *På* is used with verbs denoting four of the five senses:

Føle på (feel)

Høre på (listen)

ii) *På* is sometimes used with verbs denoting movement of parts of the body:

Falde på enden (fall on one's bottom)

Rynke på næsen (turn up one's nose)

iii) Other:

Bero på (be due to)

Hilse på (greet)

8) Attached to verbs with weak stress

In some cases verb+*på* constitutes a phrasal verb, in which *på* is stressed and has adverbial function:

Finde/hitte på (think up)

Prepositions

Skrive sig på (sign up)

9) Attached to adjectives/participles:

Gal på (angry with)

Misundelig på (envious of)

10) Idiomatic expressions with *på*:

På gensyn! (See you soon!)

Være på den (be in trouble)

(Mikkelsen, 1911) does not describe the semantics of prepositions per se, but he does describe some semantic characteristics of prepositions in various sections.

In connection with a description of noun cases, he claims that most of the relational states that can be expressed using genitive constructions may also be expressed by prepositions. He lists the following six relational states, which may be expressed by genitive forms or by the use of prepositions. The six relational states are a combination of semantic relations (1, 2, 5, 6) and grammatical relations (3, 4).

1. Possession, e.g.

<i>Hovedet</i>	<i>på</i>	<i>et barn</i>
The head	of	a child

<i>Skaftet</i>	<i>på</i>	<i>kniven</i>
The shaft	of	a knife

2. Belonging together, e.g.

En søn af snedkeren
A son of the furniture maker

En fætter til Erik
A cousin to Erik

3. Subject relation, e.g.

sang af drenge
singing by boys

undersøgelser af en videnskabsmand
investigations by a scientist

4. Object relation, e.g.

Ejeren af garden
The owner of the farm

Forlæggeren af bogen
The publisher of the book

5. Size, substance and class, e.g.

Et barn på to år
A child of two years

En lejlighed på fem værelser
An apartment of five rooms

6. Content, e.g.

Antallet af soldaterne
The number of the soldiers

Skaren af riddere
The band of knights

Prepositions

Mikkelsen claims that some relational states may only be expressed through prepositional phrases, however, the nature of these states are not defined. Examples are:

<i>En</i>	<i>mand</i>	<i>i</i>	<i>våben</i>
A	man	in	arms

<i>Lyst</i>	<i>til</i>	<i>arbejde</i>
Desire	to	work

In a section about the subordinate phrases of the sentence, Mikkelsen, among others, describes different semantic types of expressions; expressions of states, expressions of determination with respect to size, place, time, circumstances, manner, and other determinations. Some of these may be realized by a prepositional phrase:

State expressions.

State expressions describe the character or nature of a noun. The noun that it modifies may be in subject or object position, or the complement of a preposition. The state is in most cases expressed by an adjective or a participle, e.g. (a) he woke up *sound and well* (Han vågnede sund og frisk), but may in some cases be expressed by an adverb or a prepositional phrase: (b) We found the city *in a complete uproar* (Vi fandt byen i fuldstændigt oprør). Thus, in example (a), a state relation exists between *he* and *sound and well*, meaning that the person referred to by the pronoun *he* is in a state of soundness and wellness, and in example (b), a state relation exists between *the city* and *a complete uproar*, meaning that the city is in a state of complete uproar.

Size determination.

Verbs, adjectives, cardinal numbers, adverbs prepositional phrases and temporal conjunctions, may be modified by size determinators in the form of nouns, adjectival or nominal pronouns in the neuter form and adverb or prepositional phrases, e.g.:

<i>Pakken</i>	<i>vejer</i>	<i>tre</i>	<i>pund</i>
The package	weighs	three	pounds

Han løb over hundrede skridt tilbage
He ran over a hundred steps back

Size determinations have six subtypes, namely extent (quality), extent (space or time), distance (space or time), weight degree, value and quantity, difference and proportion.

Of these, Mikkelsen exemplifies the following as expressed by a prepositional phrase:

Determination of extent (space or time):

Jeg holder det næppe ud året til ende
Lit:I stand it hardly the year to end
I can hardly stand it till the end of the year

Place determination; place conditions are typically expressed by prepositional phrases, e.g.:

I Danmark
In Denmark

Ved stranden
By the beach

Time determination. Time conditions are typically expressed by prepositional phrases, e.g.:

Jeg stod op ved daggry
I got up at dawn

Manner determination. Manner conditions are typically expressed by prepositional phrases, e.g.:

Han arbejder med iver
He works with eagerness

Prepositions

Other determinations. A noun may be modified by a prepositional phrase, where the complement is identical to the noun being modified, in order to express other types of conditions:

Repetition

År efter år hører man de samme taler
Lit: Year after year hear one the same speeches
Year after year, we hear the same speeches

Succession

Han bliver bedre dag for dag
He becomes better day by day

Immediate succession

Det gik slag i slag
Lit: It went stroke by stroke
It happened in quick succession

Reciprocity

De stod ansigt til ansigt
They stood face to face

(Brøndal, 1940) describes prepositions as *'The simplest expression for relation in the language'*. The class of prepositions in general can alone be defined as expressing relations, and the individual members of the class must be defined by a sum of special relations. As noted by Brøndal, this definition is precise or specific in that it assigns general and specific relations to the class itself as well as to its individual members. This approach is fruitful in that it both requires and allows for synonymics work: it requires a differentiation between *authentic* (ægte) and *unauthentic* (uægte) prepositions, a definition of the class' relation to other classes and arrangement of the prepositions in their mutual semantic correlations. Below, we give an account of the intensional approach of defining relations and by these arranging prepositions in synonymics systems.

According to (Brøndal, 1940), prepositions can be defined by a combination of relation forms (*relationsformer*) and relation types (*relationsarter*).

Relation forms may be:

- * Polar: A relation type is polar in form if it regularly appears in two mutually complementary forms; a positive and a negative form. Polar forms are correlative or mutually dependent. This relates to the relation types described below, where polar forms may be transitive-intransitive, symmetric-asymmetric, plural-implural, etc.
- * Neutral: a relation type is neutral in form if it is neither positive nor negative, i.e. its definition includes a non-usage of a given relation type. For example, if a relation is non-symmetric, it means that the polar forms symmetric-asymmetric are not part of its definition.
- * Complex: Where polar relation forms may be described as ‘either-or’, neutral relation forms as ‘neither-nor’, complex relation forms may be described as ‘both’. For example, if a relation is both symmetric and asymmetric, it is in a complex form.
- * Complex-polar: A relation type is complex-polar in form if one relation type is at the same time complex and polar.

The necessary and sufficient number of relation types (or relational dimensions) are:

A. Abstract:

- 1) Symmetry: A relation may be symmetric, asymmetric. A relation is symmetric if xry always entails yrx . Thus, if a relation can be reversed, it is symmetric, if it cannot, it is asymmetric. This relation type defines *symmetric:asymmetric* opposites like *af:til* (of:to) and *ved:på* (by:on).
- 2) Transitivity: A relation may be transitive, intransitive or non-transitive. A relation is transitive if xry and yrz always entails xrz . Thus, if the application of a relation is transferable, then it is transitive. If not, it is intransitive. This relation type defines the opposites *på:til* (on:to) and *ved:af* (by:of).
- 3) Connexity: A relation that exists between mutually connected solidary relates is connex. A relation is connex if the existence of x

Prepositions

and y always entails xry or yrx . Thus, if solidarity exists, then the relation is connex, if it is impossible, then the relation is inconnex.

- 4) Variability: A relation can exist between groups of objects (or pluralities) or between single objects. Relations that hold between pluralities are variable, and relations that hold between single objects are invariable. The relation can be described as $n:m$ or $1:1$.
- 5) Plurality: A relation can exist between a single object and a plurality, or between a plurality and a single object. A relation that holds between a single object and a plurality is plural, and a relation that holds between a plurality and a single object is implural. The relation can be described as $1:n$ or $n:1$.
- 6) Generality: A relation can exist between a specific object and an arbitrary object, or between an arbitrary and a specific object. A relation that holds between a specific and an arbitrary object is general, and a relation that holds between an arbitrary and a specific object is ingeneral (or particular).

Composite relations:

B. Concrete

- 7) Continuity: The opposition between processes and a result/state. The relation type incorporates symmetry and transitivity; a continuous relation is symmetric and transitive, a discontinuous relation is asymmetric and intransitive.
- 8) Totality: The opposition between a whole and a part. The relation type incorporates plurality and generality; a whole presupposes both a plurality of elements and a general relation holding between them.

C. Complex

- 9) Extensionality: The opposition between intension and extension (or stative and dynamic relations). An example of *intensive:extensive* opposites are the English *at:a-* (as in *be a-singing*).
- 10) Integrity: The opposition between limitation and completeness, as illustrated by the German opposites *bis:samt* (until:including).

D. Total

11) Universality: The most far-reaching relation type that includes all other types of relations. The relation is described as being at the boundary of thought itself and describing relations of almost mythical character.

The resulting systems for Danish and English prepositions, cf. (Brøndal, 1940), are shown below in Figure 1 and Figure 2.

	Intransitive			Transitive	
Asymmetric	til		-	paa	
	for	efter		over	under
Asymmetric/ symmetric	ad	mod om	-	i	gennem mellem
Symmetric	af		-	ved	
	uden	fra		hos	med






Figure 1 The Danish system of prepositions (Brøndal 1940)

□ ST = Defined by symmetry and transitivity	<table border="1"><tr><td>inconnex</td><td>connex</td></tr></table>	inconnex	connex
inconnex	connex		
■ STC = Defined by symmetry, transitivity and connexity			
■ S ₂ T = Defined by complex symmetry and transitivity	<table border="1"><tr><td>invariable</td></tr><tr><td>variable</td></tr></table>	invariable	variable
invariable			
variable			
■ S ₂ TV = Defined by complex symmetry, transitivity and variability			

Prepositions

	Intransitive			Transitive	
Asymmetric	to		-	on	
	for	after		over	under
Asymmetric/ symmetric	against		at	a	through
	about		till		between
Symmetric	of		-	by	
	off	from		with	in

Figure 2 The English system of prepositions (Brøndal 1940)

	E = Defined extensionality	<table border="1"><tr><td>intensive</td><td>extensive</td></tr></table>	intensive	extensive
intensive	extensive			
	I ₂ = Defined by complex integrity (limitation and completeness)			
	ST = Defined by symmetry and transitivity			
	ST To = Defined by totality	<table border="1"><tr><td>partitive</td><td>total</td></tr></table>	partitive	total
partitive	total			
	S ₂ TV = Defined by complex symmetry, transitivity and variability	<table border="1"><tr><td>invariable</td></tr><tr><td>variable</td></tr></table>	invariable	variable
invariable				
variable				

In several cases, we do not agree with Brøndal's assignation of relation types and forms, as illustrated in the following examples:

Brøndal defines the Danish preposition *på* as being transitive and asymmetric, and thus, $x_{p\grave{a}}y$ and $y_{p\grave{a}}z$ should entail $x_{p\grave{a}}z$, and $x_{p\grave{a}}y$ should never entail $y_{p\grave{a}}x$. Let us take a look at example (1):

(1)

Bøffen er på tallerkenen og tallerkenen er på bordet

The steak is on the plate and the plate is on the table

It is arguably true that (1) logically entail *bøffen er på bordet* (the steak is on the table)², and if we accept this, the transitive property of the preposition is

² However, pragmatically it is not fully acceptable: If asked whether the steak is on the table if it in fact is on the plate which is on the table, most people would probably answer *no, it is on the plate!*

reasonable. However, this is not an acceptable inference for all uses of the preposition *på*, cf example (2).

- (2) *Tallerkenen er på bordet og bordet er på gulvtæppet*
The plate is on the table and the table is on the carpet

For example (2), we would never accept an assertion that the plate is on the carpet! And certainly, if we assert a contingency condition for $x_{på}_y$ to be true, as touched upon later in section 4.3.2, the relation denoted by *på* is **not** transitive.

It is reasonable, though, to define the relation denoted by *på* as asymmetric: *en bøf på tallerkenen* (a steak on the plate) does not entail *et tallerken på bøffen* (a plate on the steak).

The preposition *af* is defined as being intransitive and symmetric, and thus, x_{af}_y and y_{af}_z should never entail x_{af}_z , and x_{af}_y should always entail y_{af}_x . Let us take a look at examples (4) and (5):

- (3) *datter af et medlem af en bande*
daughter of a member of a gang

- (4) *datter af et medlem*
daughter of a member

- (5) *et medlem af en bande*
a member of a gang

It is true that (3) does not entail *datter af en bande* (daughter of a gang), and consequently the intransitive property of the preposition is plausible. However, it is not obvious that the relation is always symmetric: Example (4) does not entail *medlem af en datter* (member of a daughter), while for example (5), it is arguably acceptable to say that it entails (6).

- (6) *en bande af medlemmer*
a gang of members

However, if we accept that *af* in the context of (5) denotes a type of partitive relation, the relations denoted by (5) and (6) are in fact not identical but

Prepositions

rather inverse relations; in (5), the relation denoted by *af* could be named *part_of* and in (6) the relation could be named *has_part*. Note, however, that the Danish preposition *af* is not defined by totality in Brøndal's system.

The systems are described as synonymics systems, and while they do provide valuable information about cross-lingual synonymics, they do not provide much information about the meaning of prepositions or the language-internal synonymic properties of prepositions. As can be read from Figure 1 and Figure 2, the Danish preposition *til* and the English preposition *to* are synonymous, because they have the same definition. But what do they mean? Brøndal does not provide any answer to that question. Also, the systems are not transparent with respect to assignment of relation types and forms, as exemplified above.

2.2.2.4 Delimitation From other Word Classes

(Christian Becker-Christensen & Widell, 2003) note that, in dictionaries, most prepositions are also classified as adverbs. They are prepositions when they have a complement, and adverbs when they have no complement, cf. *Hun satte ringen på fingeren* (She put the ring on her finger) (preposition) and *Hun tog tøj på* (Lit: She put clothes on) (adverb). They are also classified as adverbs when they occur in combination with a postpositioned preposition as in *Den står bag ved den kasse* (Lit: It is behind by that box) where *bag* is considered an adverb, but in *Den står bag den kasse* (It is behind that box), where *bag* is considered a preposition. A compound consisting of an adverb and a preposition is considered an adverb when there is no complement as in *Den står bagved* (Lit: It is behind).

(Jespersen, 1924) notes that in almost all grammars, adverbs, prepositions, conjunctions and interjections are treated as four distinct parts of speech, where the difference between them seems equal to the differences between nouns, adjectives, pronouns and verbs. However, in this way the dissimilarities between them are greatly exaggerated and their evident similarities are obscured, he claims. Jespersen therefore suggests reverting to referring to the group of adverbs, prepositions, conjunctions and interjections as *particles*.

Because almost all words in the class of particles are all invariable in form, it is necessary to look at other word classes in order to find the differences between the particles. Jespersen states that many words are subject to a

distinction which is designated by different names, and that they are thus not perceived as being essentially the same. For example, when a verb has the ability to occur with or without a complement, and both cases are seen as being complete. In such cases, the verb may be intransitive in some cases, and transitive in other cases. This is e.g. the case for the verb *play* in: *he plays* and *he plays the piano*. However, despite the differences regarding complementation patterns for such verbs, no one attempts to categorize them into different word classes. Jespersen claims that the same kinds of difference in complementation patterns occur for particles. For example, for words such as *on* or *in*, they can occur with or without a complement in contexts such as *put your cap on* or *put your cap on your head* but, unlike the verbs that exhibit the same behavior, these words are termed adverbs and prepositions respectively when they occur without or with a complement.

Similarly, the differences between conjunctions and prepositions are blurred. In examples such as *after his arrival* and *after he had arrived*, where the particle *after* is traditionally treated as a preposition in the first case and a conjunction in the second, merely because the complement in the first case is a substantive and in the latter a clause. He calls such uses of conjunctions *sentence prepositions*.

For these reasons, Jespersen suggests that we do not treat the individual uses of members of the class of particles as separate word classes, any more than we treat different uses of verbs as separate word classes. The fact that he includes words which are only used as interjections as belonging to this class owes to the fact that they too are invariable in form and are thus 'most conveniently classed with other particles'.

Similarly, (Nielsen, 1995) says that by a classification into word classes by morphological criteria, a division into two main groups is achieved; the inflectable and the non-inflectable words, and the non-inflectable words cannot be further divided. This class is thus a negatively characterized and quite diffuse remainder class. This class of particles has just one common feature, namely their form invariability. In order to make the group more transparent, they can be divided into various particle functions by syntactic and semantic criteria. This way, we can achieve a division into prepositions, conjunctions/subjunctions, interjections, etc.

Prepositions

(Allan et al., 1995) also note that prepositions are closely related to adverbs and conjunctions. Many words that function as prepositions can also function as adverbs or conjunctions, e.g. *af, efter, for, i, med, på, til, ved*, etc. Traditionally, the difference between the two word classes prepositions and adverbs is that prepositions have complements, and adverbs do not. This is a reasonably clear distinction, but not entirely satisfactory. Alternative solutions include combining the two classes into one class, which is often called particles, and expanding the notion of prepositions to include those items that can *potentially* combine with a complement (i.e. including those traditionally termed adverbs). They also note that certain prepositions can function as conjunctions, e.g. *efter, for, om, til*. The distinction between prepositions and conjunctions is much more clear-cut than between prepositions and adverbs. A conjunction introduces a clause, where a preposition governs different types of complements, including a clause headed by a conjunction:

1. Preposition

Vi læste efter mørkets frembrud (we read after dusk)

Det er et spørgsmål om penge (It's a question of money)

2. Conjunction

Vi læste, efter (at) det blev mørkt (we read after it got dark)

Det er et spørgsmål, om han har nogen penge (it's doubtful whether he has any money)

2.3 Our Definition of the Class of Prepositions

Our definition of the class of prepositions draws on the cited definitions above. We define the class of prepositions by morphological, syntactic as well as by semantic criteria of which no part alone defines the class uniquely. Not every aspect of the each criterion is repeated here; we refer to the detailed descriptions above.

Morphologically based definition part

The class consists of uninflectable words. Regarding form, we accept prepositions as being simple, compound or complex.

- * Simple prepositions are one-word prepositions, such as *i*, *på*, *over*, *under*, etc.:
- * Compound prepositions are germanic compounds (i.e. written in one word), that typically consist of an adverb compounded with a simple preposition or of two simple prepositions, but are not restricted to these types. Examples of such prepositions are *foran*, *forbi*, *foruden*, *iblandt*, *igennem*, *imellem*, *imod*, *ifølge*, *jævnside*, *hinsides*, *omkring*, *udi*, etc.
- * Complex prepositions, or multiword prepositions, consist of more than one word form, but semantically comprise one unit. Examples of such prepositions include both continuous and discontinuous prepositions: *over for*, *inden i*, *ud ad*, *for ... siden*, *fra ... til*, etc.

Brøndal's argument that not all words that should be, or are, classified as prepositions are uninflectable, does not hold in our view. The pairs *over/ovre om/omme*, as Brøndal exemplifies, are in our opinion not examples of inflectional pairs, but rather of derivational pairs; *over* is a preposition, and *ovre* is an adverb. This adheres to what Diderichsen says, when he claims that some prepositions inflect, and that a few prepositions can produce adjectives in the comparative and superlative forms or relational adverbs.

An alternative explanation is that the forms *over* and *om* in the pairs *over/ovre om/omme* are in fact not prepositions, but adverbs from the onset.

Prepositions

These forms should then simply be seen as homographic with the prepositions *over* and *om*. However, no matter how one chooses to view this, the result is the same: If we view the phenomenon as derivational forms of prepositions, there is no evidence that prepositions inflect, and if we view it as inflectional forms of adverbs, there is still no evidence that prepositions inflect. All in all, there is no evidence that certain prepositions inflect, and we can maintain our definition that prepositions do not inflect.

Syntactially based definition part

Regarding syntactic properties, we assume prepositions to be transitive. Usually, the prepositional phrase complements are nouns, pronouns, or clauses. They may also be quoted expressions, adjective or a participle forms used as nouns, genitive forms used in a neuter or plural sense or adjectival numerals or pronouns used as nouns. Infrequently, they are nominal uses of adverbs, nominal uses of prepositional phrases, figurative relative clauses, conjunctive clauses or conjunctive phrases.

The complement may, in some cases, be omitted by ellipsis. In example (7), both the words *på* and *under* are considered prepositions:

- (7) *Ligger bogen på bordet? Nej, under ε*
Is the book on the table? No, under ε

In example (7), ϵ marks ellipsis where the elided element is the complement of the preposition *under*, which in this example is identical to the complement of the preposition *på*, but omitted to avoid repetition.

A preposition in combination with its complement form a prepositional phrase (a PP), which can have a number of grammatical functions in a sentence:

Indirect object:

Jeg købte blomster til min mor
I bought flowers for my mother

Subject complement:

Huset står i en smuk grøn farve
Lit: The house stands in a beautiful green farve

The house is a beautiful green color

Object complement:

Han malede huset i en rød nuance
 Lit: He painted the house in a red shade
 He painted the house in a shade of red

Adverbial

Vi går en tur i skoven
 We take a walk in the wood

Prepositions often overlap in form with adverbs and verbal particles, but we only consider the transitive variants as prepositions. This criterion serves as a means to differentiate prepositions from other particles.

Semantically based definition part

Prepositions denote binary relations that hold between two relates. They are pure relators (cf. (Brøndal, 1928, 1940)), and thus do not contain any other semantic elements than this.

As mentioned above, while Brøndal's systems do provide valuable information about cross-lingual synonymics, they do not provide much information about the *meaning* of the individual prepositions. For example, the systems provide the information that the Danish preposition *til* and the English preposition *to* are synonymous, because they have the same definition, but they do not provide any information as to what they mean. Combined with the fact that the systems are opaque, we choose not use Brøndal's detailed theory directly.

Prepositions are at the same time highly polysemous and synonymous, i.e. a given preposition has a number of senses, and a given sense may be expressed by a number of prepositions. A detailed corpus-based analysis of the sense inventory for a subset of Danish prepositions is provided in Chapter 6.

Thus, we can summarize our definition of the class prepositions:

Prepositions

- a) *The class consists of uninflectable words which may be of simple, compound or complex form.*
- b) *Prepositions are transitive. Their complement may be of various forms but is typically a noun, a pronoun or a clause (including infinitives).*
- c) *Prepositions are pure relators that denote binary relations.*

This definition adequately defines the members of the class of prepositions.

Prepositions

Chapter 3

Ontologies

Etymologically, the word *ontology* consists of two roots; *ontos* which means *being*, and *logos* which means *study*. In modern usage, the word is polysemous, i.e. it has two different but related meanings.

In a chronological view, the first sense denotes a **philosophical theory of metaphysics** concerned with the study and categorization of what exists, or can be believed to exist, in the world, as well as with the interrelations between these entities. The use of the word *Ontology*³ in this sense was introduced in the beginning of the 17th century, and first recorded in a dictionary a century later in 1721: Bailey's dictionary defines ontology as '*an Account of being in the Abstract*'. However, the ideas behind the theory go back to the classic Greek philosophers, mainly Aristotle.

In its second sense, the word is used to denote an **information theoretical artefactual construct**. Such a construct, an ontology, can be described as a formal representation of a set of concepts within a domain and of the relations that exist between them. The use of the word *ontology*⁴ in this sense was introduced in the 1980's in the field of artificial intelligence.

³ In the abstract sense described here, the noun *Ontology* is uncountable and orthographically normally capitalized.

⁴ In the concrete sense described here, the noun *ontology* is countable and not capitalized.

Ontologies

In the rest of this chapter, as well as in the dissertation as such, we are concerned with the artefactual constructs of ontologies as denoted by the second sense of the word *ontology*.

This chapter gives a foundational account of ontologies, but focuses on the types of ontologies that are part of the main work described in this dissertation, namely lexical ontologies.

In section 3.1, we seek to define the concept of an ontology, in section 3.2, we describe how it is possible to categorize ontologies, as well as describe various types of ontologies. In section 3.3 we put special emphasis on lexical ontologies or wordnets: in section 3.3.1, we describe the mother wordnet, Princeton WordNet, in section 3.3.2, we describe the European common wordnet project called EuroWordNet, in section 3.3.3 we describe the Danish wordnet DanNet, and finally in section 3.4 we give a brief summary of the chapter.

3.1 What is an Ontology

In the years that ontologies have been around, a number of definitions have been put forward. The most widely quoted definition is coined by Tom Gruber (Thomas R. Gruber, 1993; T. R. Gruber, 1995):

An ontology is an explicit specification of a conceptualization.

Gruber further defines a conceptualization as *an abstract, simplified view of the world that we wish to represent for some purpose*. Others have given other definitions of the concept of an ontology:

- * *A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. (Genesereth & Nilsson, 1987)*
- * *An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary. (Neches et al., 1991)*
- * *An ontology is a logical theory which gives an explicit, partial account of a conceptualization. (Guarino & Giaretta, 1995)*

- * *An ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. (Swartout, Ramesh, Knight, & Russ, 1997)*
- * *An ontology is a set of logical axioms designed to account for the intended meaning of a vocabulary. (Guarino, 1998)*
- * *An ontology is a formal, explicit specification of a shared conceptualization. Conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group (Studer, Benjamins, & Fensel, 1998) (an extended and explained version of Gruber's definition).*
- * *OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called an ontology. (W3C, 2004)*
- * *an ontology (...) is a representation of the types of entities existing in the corresponding domain of reality and of the relations between them. (Chen & Lonardi, 2009)*

And we could go on. Thus, there is not yet absolute consensus about how to define the concept of an ontology, but we can combine and boil down the above-mentioned definitions to the following definition:

An ontology depicts a consensual, simplified view of a domain, and is an explicit representation of the types of entities that exist in the domain as a hierarchical structure as well as of the corresponding terms. In addition to entities, an ontology comprises relations between entities and it may also comprise rules for combining entities and relations.

3.2 Types of Ontologies

Ontologies are used in a variety of fields and for a variety of purposes, ranging from e.g. document classification to complex question-answering systems. The requirements for the ontologies in the different ends of the

Ontologies

spectrum are, of course, quite different. Ontologies may be expressed in a variety of formalisms, ranging from a simple list or box notation to fully expressive logics.

We can categorize ontologies based on different features, e.g.:

- * Level of abstraction (specific \rightarrow top)
- * Domain or task (general \rightarrow specific)
- * Level of formality (informal \rightarrow formal)

Regarding the first feature, *level of abstraction*, we categorize ontologies based on the specificity of the contained concepts. A top-level ontology only contains very general concepts, e.g. *event*, *abstract entity*, *concrete entity*, *artefact*, etc., that are all common to the more specific concepts in one or more lower-level ontologies, e.g. a domain ontology, and independent of any particular task or domain.

In relation to the second feature, *domain or task*, we categorize ontologies as general ontologies if they model the general domain or are independent of a specific task. Domain-specific ontologies model a special domain (e.g. aviation, genes, etc.), and possibly, we can have ontologies that are both general domain and domain-specific if they contain parts of both. Task-specific ontologies are ontologies that are designed to be used for a specific task, e.g. e-commerce, diagnostics systems, etc.

We can illustrate these distinctions by the figure rendered in Figure 3, cf. (Guarino, 1997, 1998).

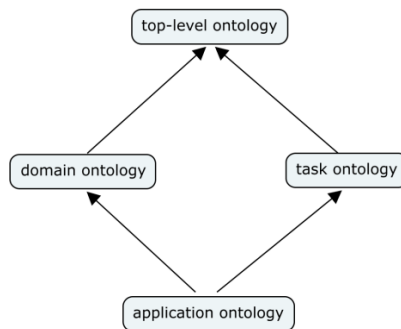


Figure 3 Kinds of ontologies

The type of ontology *application ontology* in Figure 3 is an ontology that is designed for a specific use in a specific application, and it thus contains concepts that are relevant to both a specific domain and a specific task.

Normally, in a graph representation of an ontology, directed arrows represent ISA relations. If we read the arrows in Figure 3 as such, we get at an interpretation that a *domain ontology* ISA *top-level ontology*. Surely, this is not the intended reading. As noted in (Guarino, 1997) as well as in (B. Madsen & Thomsen, 2008), the figure should be read such that individual terms in e.g. a domain ontology are specializations of individual terms in the top-level ontology. To avoid a possible misinterpretation, (B. Madsen & Thomsen, 2008) propose a revised version of Figure 3, similar to the rendition in Figure 4. For the types of ontologies in this figure, a top-level ontology describes general concepts, a domain ontology describes domain-dependent concepts, a task ontology describes task dependent-concepts, and finally, an application ontology describes concepts that depend on a given domain *and* task.

For a more detailed description of a suggestion to an ontology of ontology types, cf. (B. N. Madsen & Thomsen, 2009), see section 3.2.2.

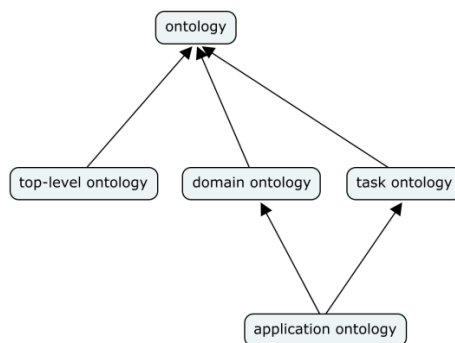


Figure 4 A revised version of 'Kinds of ontologies'

For the third feature, *level of formality*, we categorize ontologies based on the formality of description. Informal ontologies are purely vocabulary-based models with some form of structure, and formal ontologies use some logic to describe the meaning of concepts and relations between concepts. We elaborate further on this in section 3.2.1 below.

3.2.1 An Ontology Spectrum

For the ontology categorization criterion, level of formality of description, we can use the illustration of an ontology spectrum from (Lassila & McGuinness, 2001) as rendered in Figure 5, as a means of exposition.

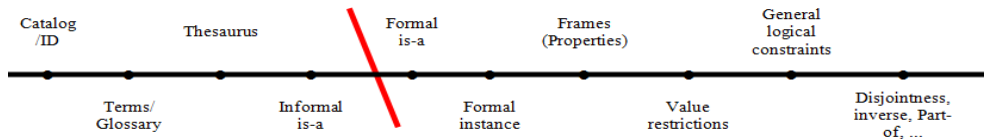


Figure 5 An ontology spectrum (Lassila & McGuinness, 2001)

On the left-hand side of the dividing line in Figure 5, we find what are often referred to as *lightweight* ontologies, and on the right-hand side, we find what are often referred to as *formal* or *heavyweight* ontologies.

In the leftmost end of the ontology spectrum from (Lassila & McGuinness, 2001), we find the point *Catalog/ID*. A catalog may be a controlled vocabulary where each uniquely identified word in a finite list of terms has a specific meaning, i.e. ambiguity is not a factor.

The next point in the spectrum is *Terms/Glossary*, where terms are glossed with a meaning statement in a natural language. Such glossaries are intended for human readers, and are typically not in a form that can be understood by a computer.

The next point, *Thesaurus*, offers additional semantics. Thesauri include information about relations between terms, e.g. synonymy and broader and narrower term relations; however, they do not provide an explicit hierarchy. Thesauri may exist in computer understandable formats.

As we near the dividing line that brings us to formal ontologies, we get to the point *Informal ISA*. Informal ISA hierarchies are for example web indices as the ones provided by e.g. search engines such as Yahoo and Google or by e-businesses such as Amazon, eBay, etc. Such indices provide a number of ordered general categories under which web pages can be categorized; however, they do not form a strict ISA hierarchy. As an example, Google directory categorizes documents concerned with trade in riding helmets under the category: Shopping > Sports > Equestrian > Apparel. While (riding) apparel *is* related to equestrian, it can hardly be said to be a *type of* Equestrian. Similarly, the

relation between sports and shopping cannot be said to be a type relation. Thus, such directories can be said to express associative relations between categories rather than strictly hierarchical relations. Informal ISA hierarchies are typically machine understandable.

In our view, ontologies must have some kind of hierarchical structure, and thus, we do not consider lists of terms such as catalogs and glossaries as ontologies. Nevertheless they can be a fruitful resource in the ontological modeling of a domain. We thus describe *lightweight ontologies* in the following manner:

- * Some hierarchical structure.
- * Include natural language terms and relations between them.

Moving to the other side of the dividing line, the next point in the spectrum is *Formal ISA*, or strict subclass hierarchies. Such hierarchies include transitivity, which can be explained such that if B ISA A, and if C ISA B, then it necessarily follows that C ISA A. For example, if “SafetyApparel” ISA “Apparel” and “Helmet” ISA “SafetyApparel”, then it follows that “Helmet” ISA “Apparel”. The next point, *Formal instance*, is similar to formal ISA, but with instances of classes. If A ISA B, then if an object is an instance of B, it necessarily follows that the object is an instance of A. Thus, if “Helmet” ISA “SafetyApparel” and “MyHelmet” is an instance of “Helmet”, then it follows that “MyHelmet” is also an instance of “SafetyApparel”.

Next in the spectrum is *Frames,properties/attributes*. Here, classes may have properties. For example, the “Apparel” class may have the properties “price” and “Material” that is inherited by its subclasses. Instances of classes can have values associated with these properties. The next point, *Value restrictions*, allows for restrictions on such values. For example, the restrictions may be on the data type (integer, string, etc.) or on the value range (e.g. that is has to be of a certain ontological type).

Next point, *General logical constraints* include constraints on combinatory capacities between classes or classes and properties. For example, we can imagine a general logic constraint that says that it is not possible for an individual to be asleep and awake at the same time. Finally on the rightmost end of the spectrum, we have *Disjointness, inverse, part-of,...* which allows

Ontologies

for the specification of more detailed relationships such as disjoint classes, inverse relationships, part-whole relationships, etc.

We can sum up the description of *heavyweight ontologies* in the following manner:

- * Strict ISA hierarchies
- * Possibility for explication of the meaning of concepts expressed in terms of some logic.
- * Possibility for explication of properties such as inheritance, value restrictions, logical constraints, inverse relationships, part-whole relationships in terms of some logic.

The ontology spectrum in Figure 5 fails to specifically mention one of the most widely used types of ontologies, namely the lexical/terminological ontology, even though they may fall under the *formal ISA*. Lexical/terminological ontologies are types of ontologies that are mainly focused on the lexical expressions (that denote concepts) and arrange them in a skeletal strict ISA hierarchy, but also comprise other types of relations. Wordnets typically comprise general language vocabulary, while terminological ontologies comprise domain-specific terms and may also include additional terminological information such as subdivision dimensions and characteristics.

Typically, such ontologies do not comprise logic forms, axioms, etc., and we would thus position them in the middle of the spectrum and name them *middleweight ontologies*. Examples of middleweight ontologies are wordnets, which are described in more detail in section 3.3, and terminological ontologies, such as the *ontology of ontologies* in Figure 7. We can sum up the description of *middleweight ontologies* in the following manner:

- * Strict ISA hierarchies
- * Focus on natural language expressions.
- * Include concepts labelled with natural language expressions and relations between them.
- * May include subdivision dimensions and characteristics

In addition, some speak of *formal lightweight ontologies*, cf. (Fausto, Biswanath, & Vincenzo, 2009) for ontologies that fall into both the heavyweight and lightweight categories. Such ontologies may be obtained by translating the natural language labels of lightweight (or middleweight) ontologies into logic forms.

Figure 6 shows our proposal for a revised ontology spectrum with the new category *middleweight ontologies*.

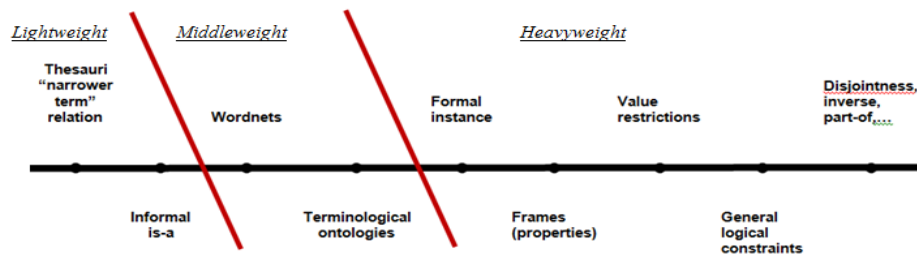


Figure 6 A revised ontology spectrum

3.2.2 An Ontology of Ontologies

In (B. N. Madsen & Thomsen, 2009), the different types of ontologies are illustrated by means of an ontology of ontologies, cf.

Figure 7. This figure presents a different way of categorizing ontologies. The figure itself is an example of a terminological ontology, where concepts are shown as boxes, type relations as full straight lines, and part-whole relations as angular lines. There are no associative relations in the ontology in

Figure 7, but these are part of the formalism, and would be shown as dotted lines labelled with the relation type. Further, subdivision dimensions are shown as boxes spanning across type relation lines, and delimiting feature specifications are shown as feature-value pairs below concepts. Features are inherited. In such an ontology, polyhierarchy is allowed, but duplicated delimiting features is not. This means that a new concept that is in a type relation to two or more concepts under the same subdivision dimension is not allowed, but a concept that is in a type relation to two or more concepts under different subdivision dimensions would be allowed. Thus, we may add a legal concept *general domain lexical ontology* and this concept would inherit the features DOMAIN:general and PRINCIPLE:lexical. However, we cannot legally add a concept *frames description logic ontology* as it

Ontologies

would inherit an illegal duplicated delimiting feature PARADIGM: PARADIGM:frames and PARADIGM:description logic.

According to the *ontology of ontologies* shown in Figure 7, there are 10 general subdivision dimensions described below:

Under the dimension POINT OF VIEW, we find *philosophical ontology* and *pragmatic ontology*. This division refers to the two senses of the word *ontology*, as described at the beginning of this chapter.

Under the dimension ISSUE OF CONCEPTUALIZATION, we find only one ontology type, namely *meta ontology*. A meta ontology is a type of ontology that contains the meta concepts that are essential for the modeling of some domain, such as the ontology of ontologies. We would assume that at least one other type of ontology should fall under this dimension, namely the ontologies that are not meta ontologies.

Under the dimension RELATION TYPES, we find ontologies with different types of relations; ordering or non-ordering: ontology with type relations, ontology with partitive relations, ontology with associative relations, and ontology with mixed relations.

Under the dimension LEVEL, we find universal ontology, top-level ontology and specific ontology which are all in a type relation to the general concept ontology. However, top-level ontology and specific ontology are also in a partitive relation to universal ontology. This structure reflects the revised view on (Guarino, 1998) as described above in section 3.2.

Next, under the dimension DOMAIN, we find *general ontology* and *domain-specific ontology*. An ontology that models the conceptual content of the vocabulary in a specific domain is a domain-specific ontology, and one that models the conceptual content of the general vocabulary is a general ontology. Since an ontology that contains both general and domain-specific concepts and thus can be classified as a domain-specific ontology as well as a general ontology, cannot legally be added as a subconcept due to the principles described above, we propose that such an ontology type is added as a sibling concept with the characteristic DOMAIN:mixed.

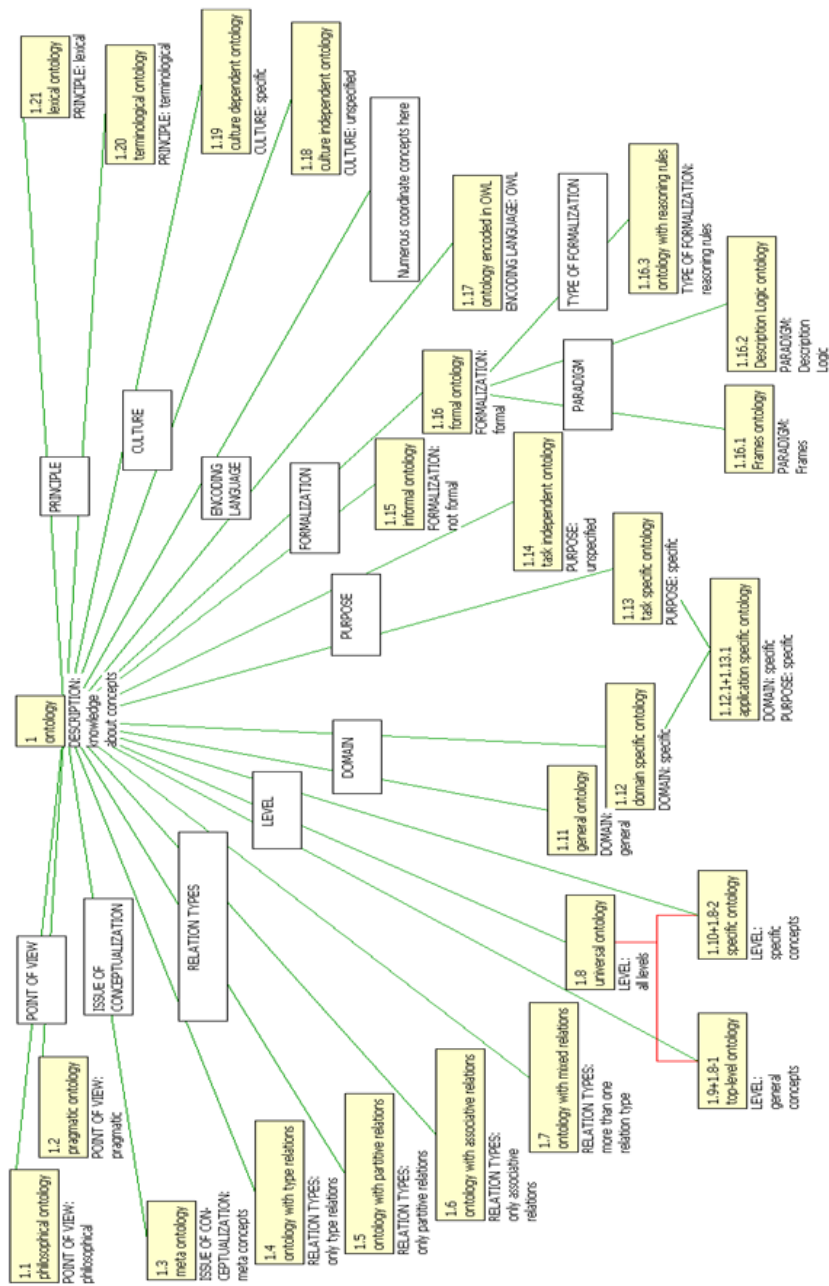


Figure 7 Ontology of ontologies (B. N. Madsen & Thomsen, 2009)

Ontologies

Under the dimension PURPOSE, we find task-specific ontology and task-independent ontology. A task-specific ontology is an ontology that has been designed for some explicit task, e.g. information retrieval, and contains only the information relevant to this task. A task-independent ontology is an ontology that has not been designed for any explicit task, but can be used and re-used in a number of different types of applications.

Under the dimension FORMALIZATION, we find informal ontology and formal ontology. The division into these two types is based on similar criteria as the ones described above in section 3.2.1. Further, formal ontologies may be subdivided according to the applied PARADIGM and TYPE OF FORMALIZATION. For the dimension PARADIGM, we find *Frames ontology* and *Description logic ontology*, and under the dimension TYPE OF FORMALIZATION, we find *ontology with reasoning rules*.

Under the dimension ENCODING LANGUAGE, we find *ontology encoded in OWL* and numerous coordinate concepts that have not been explicated. A large number of ontology encoding languages exist at this point, and more will surely emerge.

Next, under the dimension CULTURE, we find *culture dependent ontology* and *culture-independent ontology*. This division is dependent on the domain being modeled; an ontology of e.g. the legal domain is by nature culture dependent, while an ontology of plant species is not likely to be culture dependent.

Finally, under the dimension PRINCIPLE, we find *lexical ontology* and *terminological ontology*. Note that this dimension refers to the modeling principles and not to the formalization principles. However, while the principles involved in the design of terminological ontologies are well described (cf. e.g. (B. N. Madsen & Thomsen, 2006; Bodil Nistrup Madsen, Hanne Erdman Thomsen, & Carl Vikner, 2004; B. N. Madsen, Thomsen, & Vikner, 2005)), it is not obvious what specific principles are involved in the design of a lexical ontology per se. The principles behind wordnets, however, are well described, e.g. (Christiane Fellbaum, 1998; Piek Vossen et al., 1998), and we thus assume these to apply here. (Bodil Nistrup Madsen, Hanne Erdman Thomsen, & Carl Vikner, 2004) describes some differences in the modeling principles for the two types of ontologies.

3.3 Lexical Ontologies

In this section, we describe various lexical ontologies or wordnets. The special emphasis that is put on wordnets in this section stems from the fact

that the main experimental work described in this dissertation (cf. chapters 5 and 6) uses wordnets as ontologies. Wordnets are lexical databases that group words in a language into synonymous meaning sets, synsets, and specify relations that hold between synsets. They describe the order that already exists between words and the phenomena they denote, as poetically put below in (8):

(8)

Set udefra, i deres tilfældige tilstand, f.eks. i en ordbog, ligner ordene kaos. Men egentlig er de altid i orden, så at sige hjemme hos deres fænomener. Vi tror imidlertid, at det altid er op til os at ordne ordene i sætninger og modsætninger, før det hele ordner sig. Intet kan være mere forkert. Den orden vi prøver at ordne os til, findes i forvejen.

(Christensen, 2000)

From an outside perspective, in their random state, for example in a dictionary, the words resemble chaos. But, in reality, they are always ordered, at home with their phenomena, so to speak. While we believe that it is up to us to arrange the words into propositions and oppositions until everything is in order, nothing could be further from the truth. The order we attempt to arrange, is already there.

(my⁵ translation)

Lexical ontologies or wordnets are used for a variety of purposes, but are particularly useful for information search purposes. It is well known that keyword-based search has its limitations, specifically that a search for a word or a sequence of words will only return answers that contain an exact match to the query, but exclude possibly relevant answers that do not contain the exact search term. Many search systems apply natural language processing methods in order to achieve better query results; such methods include stemming or lemmatization and may even include synonym expansion of individual query terms. More advanced systems use conceptual query expansion by use of a lexical ontology, where query terms are expanded to include terms denoting subconcepts or other related concepts. For query expansion by synonyms as well as by sets of synonyms

⁵ Thanks to Stig W. Jørgensen and Merete Bert Lassen for help with this translation!

Ontologies

denoting subconcepts, lexical ontologies are key. For example, if the original query is *painting*, the query may be expanded with synonyms yielding ‘*painting* OR *picture*’ and with subconcepts⁶ in addition yielding e.g.:

- => *mural* OR “*wall painting*”
- => *nude* OR “*nude painting*”
- => “*oil painting*”
- => *portrait*
- => “*sand painting*”
- => *seascape* OR *waterscape*
- => “*still life*”
- => “*trompe l'oeil*”
- => *watercolor* OR *water-color* OR *watercolour* OR *water-colour*

Since wordnets are lexical constructs, they are language-specific by nature. Wordnets exist for a number of different languages; however, we will not go into detail with them all. In section 3.3.1, we will describe the English language wordnet, Princeton WordNet, and in section 3.3.2, we will describe a group of wordnets for various European languages, EuroWordNet. Further, in section 3.3.3, we will describe the Danish language wordnet, DanNet.

3.3.1 Princeton WordNet

Princeton WordNet, henceforth referred to as WordNet, is a lexical database for English which was originally developed at Princeton University under the direction of the psychology professor George A. Miller. The principles behind the design were inspired by psycholinguistic and computational theories of human lexical memory.

3.3.1.1 WordNet Principles and Evolution

The development of the database commenced in 1985, and is still going on. Version 1.0 of WordNet was released in 1986, and the current version, WordNet 3.0, was released in 2006. Even though we, as many others,

⁶ The following is a subset of the hyponyms of *painting* in Princeton WordNet 2.1

conceive WordNet as a lexical ontology, this was not the original idea behind the construct.

During the 1980s, the idea for WordNet emerged from trying to understand how children learn new words. Originally, the idea was to understand the learning process of children by simulating the acquisition of lexical meaning. This effort failed however, but the attempt led to new discoveries about relations between words, cf. (Miller 1990, 1995; Fellbaum 1998). According to (G. Miller & Fellbaum, 2007), there are two major approaches to the semantic analysis of words:

- Componential analysis, which is characterized by inclusion of generic concepts in more specific concepts; a relation also known as the subsumption relation. For example, the concept MURDER is said to include the superordinate concept KILL. This type of analysis is applied in class-based ontology modeling.
- Relational semantics, which relates words without assuming anything about composition or semantic inclusion. According to this approach, relations between words are reflected in a semantic network of word meanings; for example, *car* and *vehicle* can be regarded as labels for two nodes in a semantic network, and an ISA edge between these nodes simply represents the conception that a car is a kind of vehicle.

WordNet is based on relational semantics.

The ISA relation is not the only relation that relates nodes in WordNet, cf. section 3.3.1.2. Another frequent relation for nouns in WordNet is the part-of relation that for example relates *tire* and *car*, and which expresses the conception that a tire is a part of a car. The ISA and the part-of relation relations with the addition of antonymy and entailment for adjectives and verbs respectively, including their inverses, formed the initial basic semantic relations that structure WordNet.

In WordNet, each node consists of one or more words that are synonymous or *cognitively synonymous* (cf.(D. A. Cruse, 1986)), called synonym sets or just *synsets*. This conception of synonymy is not strict: That words are synonymous in this sense means that they can be substituted for one another in most contexts, but not necessarily in all, and that such a substitution may

Ontologies

not change the truth value of a proposition. Initially, WordNet contained just nouns. Later, verbs and adjectives were added, and finally, in the 1990s, adverbs were added. The synsets for each word class were added separately, and as a result, they initially formed four independent networks.

As mentioned above, WordNet is often referred to and used as a lexical ontology, but since WordNet was not originally meant to be an ontology, the constructors were not from the onset concerned with following any ontological *best practice* (G. Miller & Fellbaum, 2007). But since WordNet is being widely used as an ontology, some changes are gradually being made that makes the database more ontology-like. The goal is to improve the usefulness of WordNet for language-based problems that require both basic lexical information and reasoning, and to improve WordNet's capacity to meet the increasingly high demands by language-based applications, (Clark, Fellbaum, Hobbs, et al., 2008). As an example, WordNet was originally constructed with 25 so-called unique beginners, rather than a common top node. However, repeated wishes from users to merge the 25 trees were heard, and WordNet now provides a common top node labeled *entity*. Also, WordNet did not originally distinguish between types and instances, and consequently the relations between e.g. *a nation* and *political unit* and between *Spain* and *nation* were represented in the same way, namely as the ISA relation. However, in ontology modeling, concepts such as *nation* are often viewed as types (or classes) and individual occurrences of such types, such as *Spain*, are viewed as instances. In the latest versions of WordNet, the instance-of relation has been added for such cases, cf. (G. A. Miller & Hristea, 2006).

Other additions to Wordnet in the latest version include the following, which are further described below in section 3.3.1.2

In order to increase WordNet's effectiveness, especially for word sense disambiguation purposes which initially was limited because of the sparsity of edges, and particularly the lack of cross-part of speech edges, cf. (G. Miller & Fellbaum, 2007), *morphosemantic links* have been introduced (Christiane Fellbaum & Miller, 2003). Morphosemantic links are (morpho)semantic relations between morphologically related nouns and verbs, where e.g. the noun *employer* is linked to the appropriate senses of the verb *employ*. In addition, the morphosemantic links give the semantic type of the relationship, as for example for the relation between *employ* and *employer* which is given as an agent relation, (Princeton_University, 2010).

All glosses in WordNet 3.0 have been translated into axioms of the form:

John(x1)& work(e,x1)& present(e)

The main additions to WordNet by version is outlined below:

- **v1.0 (1986)**
 - synsets (concepts) + ISA links
- **v1.7 (2001)**
 - additional relationships
 - has-part
 - causes
 - member-of
 - entails-doing
- **v2.0 (2003)**
 - instance/class distinction
 - Paris instance-of Capital-City ISA City
 - derivational links
 - explode related-to explosion
- **v3.0 (2006)**
 - No major changes in the database per se, but additional files were added

3.3.1.2 Structure and Contents of the Database

WordNet groups English words from the word classes nouns, verbs, adjectives and adverbs into sets of synonyms, called synsets, that each represent what we refer to as a concept. A synset consists of an inventory of synonymous words or collocations from the same word class. In WordNet, a collocation is a string of two or more words, and may e.g. be multi-word compounds (eg. “fountain pen”), phrasal verbs (e.g. “take in”) or stable collocations/frozen expressions (e.g. “walk of life”). The lexical matrix from (G. Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) rendered in Table 1, illustrates the relation between word forms and word meanings, and the grouping of synonymous word forms into synsets. In Table 1, M_1 - M_m are to be understood as meanings, and for a given M , the word forms occurring on the horizontal level constitute a synset connected to the given meaning. Thus, the word forms F_1 and F_2 are synonymous, and are thus in the same synset, namely the synset connected to M_1 , and the word form F_2 is

Ontologies

polysemous; it has entries in two synsets, namely the synsets connected to M_1 and M_2 .

		Word Forms				
		F_1	F_2	F_3	...	F_n
Word Meanings	M_1	$E_{1,1}$	$E_{1,2}$			
	M_2		$E_{2,2}$			
	M_3			$E_{3,3}$		
	...					
	M_m					$E_{m,n}$

Table 1 Lexical matrix (cf. (G. Miller et al., 1990))

Words and synsets are linked together through semantic and lexical relations. In WordNet, lexical relations are relations that hold between word forms, and include synonymy for all word classes, antonymy for adjectives, and derivationally related form for verbs. Semantic relations are relations that hold between word meanings (or synsets). The inventory of semantic relations varies based on the word class, but includes the relations in Table 3.(cf. (Princeton_University, 2010)). Table 2 gives the number of words, synsets, and senses in the latest version of WordNet, WordNet 3.0.

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Table 2 Number of words, synsets, and senses in WordNet 3.0 (Princeton_University 2010)

Word class	Relation	Definition	Example
Nouns			
	Hypernym	Y is a hypernym of X if A(X is a (kind of) Y)	The generic term used to designate a whole class of specific instances.
	Hyponym	X is a hyponym of Y if A(X is a (kind of) Y)	The specific term used to designate a member of a class.
	Coordinate	X is a coordinate term of Y if Y and X share a hypernym	Coordinate nouns are nouns that have the same hypernym
	Holonym	Y is a holonym of X if A(X is a part of Y)	The name of the whole of which the meronym names a part.
	Meronym	X is a meronym of Y if A(X is a part of Y)	The name of a constituent part of, the substance of, or a member of something.
Verbs			
	Hypernym	The verb Y is a hypernym of the verb X if A(the activity X is a (kind of) Y)	
	Hyponym	The verb X is a hyponym of Y if A(the activity X is a (kind of) Y)	
	Troponym	X is a troponym of Y if A('to X' is 'to Y' in some manner)	A verb expressing a specific manner elaboration of another verb. For example, 'rewrite' is a troponym of 'write'.
	Entailment	A verb X entails Y if A(X cannot be done unless Y is, or has been, done)	A verb expressing a state/event that requires some other state/event to be true in order to be true. For example, if Z snores, it is required that Z sleeps.
	Coordinate	X is a coordinate term of Y if Y and X share a hypernym	Coordinate verbs are verbs that have the same hypernym.

Ontologies

Table 3 Some semantic relations in WordNet (Christiane Fellbaum, 1998).

As mentioned above in section 3.3.1.1, WordNet is being increasingly widely used as an ontology in natural language processing, so much so that Kilgarriff (Kilgarriff, 2000) claims that “*not using it requires explanation and justification*”. In order to meet the demands from such users, some changes are gradually being made that makes the database more ontology-like. Below, we briefly describe some additional resources to WordNet 3.0 as so-called standoff files that are part of the effort to make WordNet more of an ontology.

Core WordNet

A list of approximately 5000 core word senses in WordNet has been extracted from WordNet. The full WordNet contains tens of thousands of synsets that refer to highly specific animals, plants, chemical compounds, etc. that are less relevant to processing of general language texts. For this reason the Princeton WordNet group has compiled a core WordNet that consists of approximately 5000 synsets that all express *frequent and salient concepts*. The concepts were selected by first compiling a list with the most frequent strings from the British National Corpus, and then extracting all WordNet synsets containing these strings. Then, human raters determined which of the possible senses for the strings expressed salient concepts. The resulting top 5000 concepts comprise the core WordNet. As a result of applying this method, the core concepts constitute a mix of general and commonly used domain-specific terms, (Clark, Fellbaum, Hobbs, et al., 2008).

The core WordNet terms are distributed on word classes as follows:

Nouns: 3299

Verbs: 1000

Adjectives: 698

Gloss Corpus

All nouns, verbs, adjectives and adverbs in the glosses for all synsets have been disambiguated against WordNet senses and linked to the corresponding synsets. This work has resulted in a semantically annotated corpus consisting of the annotated glosses, aka *Princeton WordNet Gloss Corpus*.

Logical Forms of the Glosses for the Full WordNet

All glosses in WordNet 3.0 have been translated into logical forms using eventuality notation. The translation was performed in the following manner: Initially, each word and its corresponding gloss is transformed into a sentence of the form “*word is gloss*” and then parsed. Afterwards, the parse tree is converted into a logical form with variables. As syntactic relations are recognized, variables in the logical fragments are acknowledged as being equal. For example, *John works* would initially be translated into:

John(x1) & work(e,x2) & present(e)

Where *e* is the variable for a working event. Then, when the system has recognized *John* as the subject of *works*, *x1* and *x2* are made equal:

John(x1) & work(e,x1) & present(e)

In this way, all the modified WordNet glosses have been translated into axioms of the form:

;;; "*ambition#n2: A strong drive for success*"
ambition(x1) -> a(x1) & strong(x1) & drive(x1) & for(x1,x6) & success(x6)

Finally, all predicates are assigned word senses by means of sense-tagged gloss corpus described above. (Clark, Fellbaum, Hobbs, Harrison et al., 2008)

Logical Forms of the Glosses for the Core WordNet Nouns

Logical forms for the glosses of the noun senses in core WordNet were subject to a more detailed analysis, and thus, the logical forms of the glosses for the core concepts are generally of higher quality than those for all glosses described above (Princeton_University, 2010). The Logical Forms of the glosses for the core WordNet concepts exist as a separate downloadable file.

Ontologies

The Teleological⁷ Database

The teleological database contains, for approximately 350 artifacts (nouns), an encoding of the typical activity (purpose) for which that artifact was intended or designed for, e.g., a *ball* is intended for the action *throwing*.

The encoding has the form of a set of triples for each artifact:

<artifact> *action* <verb1>
<artifact> *action* <verbN>
<artifact1> <relation1> <object1>
<artifact> <relationN> <objectN>

This structure of the triples denotes that for a triple

<artifact> *action* <verb >

<verb> is a typical intended activity or purpose for which the artifact was designed.

And for a triple:

<artifact> <relation> <object>

<object> describes a typical object involved in an activity for which the artifact was designed, and <relation> describes the semantic relation that holds between the activity and the object.

For a typical intended activity, there are 11 possible semantic relations. Note that these relations hold between the intended activity for the artifact (not the actual artifact) and the object of the activity. The semantic relations used in the teleological database are as follows:

RELATION	DESCRIPTION
agent	a rester is a (typical) AGENT of sleeping on a bed
beneficiary	an audience is a (typical) BENEFICIARY of

⁷ The term *teleology* derives from Greek and consists of the two roots *telos* which means “purpose or end” and *logos* which means “word” or “study”. Thus, *teleological* is concerned with the study of the purpose or end of things.

	showing a movie
cause	tiredness is a (typical) CAUSE of sleeping on a bed
destination	a shore is a (typical) DESTINATION of sailing a boat
experiencer	a child is a (typical) EXPERIENCER of swinging on a swing
instrument	a gun is a (typical) INSTRUMENT of shooting a bullet
location	a bedroom is a (typical) LOCATION of sleeping on a bed
result	rest is a (typical) RESULT of sleeping on a bed
source	a shore is a (typical) SOURCE of sailing a boat
theme	a passenger is a (typical) THEME of transporting by boat
undergoer	a target is a (typical) UNDERGOER of shooting an arrow

As an example, for the artifact *stick*, the database contains the following information:

stick	action	hit
stick	theme	ball
stick	agent	hockey_player
stick	location	skating_rink

Morphosemantic links

WordNet 3.0 contains derivational links connecting morphologically related nouns and verbs, where e.g. the noun *employer* is linked to the appropriate senses of the verb *employ*. In addition, the morphosemantic links give the semantic type of the relationship, as for example for the relation between *employ* and *employer* which is given as an agent relation. The database uses 14 morphosemantic relations, as listed below:

MORPHOSEMANTIC RELATIONS	
Relation	Example
agent	employer/employ

Ontologies

body-part	abduct/abductor
by-means-of	dilate/dilator
destination	tee/tee
event	employ/employment
instrument	poke/poker
location	bath/bath
material	insulate/insulator
property	cool/cool
result	liquify/liquid
state	transcend/transcendence
undergoer	employee/employ
uses	harness/harness
vehicle	kayak/kayak

3.3.2 EuroWordNet

EuroWordNet (EWN) is a multilingual database consisting of wordnets for a number of European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian), which has been designed with the primary perspective of information retrieval. The EuroWordNet project was funded by the EU, and was completed in 1999. After the completion of the project, the design of the database, the defined relations, the top-ontology and the Inter-Lingual-Index is stationary. However, other groups have since developed similar wordnets in other languages using the EuroWordNet specifications, including DanNet for Danish, as described below in section 3.3.3. Wordnets now exist for e.g. Swedish, Norwegian, Greek, Portuguese, Basque, Catalan, Romanian, Lithuanian, Russian, Bulgarian, Slovenian and Danish. The size of the individual wordnets in EWN is shown in Table 4.

Language	Synsets	Word Meanings	Language Internal Relations	Equivalence Relations
<i>Dutch</i>	44015	70201	111639	53448
<i>Spanish</i>	23370	50526	55163	21236
<i>Italian</i>	40428	48499	117068	71789
<i>German</i>	15132	20453	34818	16347
<i>French</i>	22745	32809	49494	22730
<i>Czech</i>	12824	19949	26259	12824
<i>Estonian</i>	7678	13839	16318	9004

Table 4 EuroWordNet statistics (Piek Vossen, 2001)

The individual wordnets in EWN are structured in the same way as Princeton WordNet, as described above, in that they contain synsets and basic semantic relations that hold between them. However, in addition, the wordnets are mapped to an Inter-Lingual-Index (ILI). This ILI is an unstructured list of concepts, primarily made up of concepts from Princeton WordNet. Via a mapping to this index, the various wordnets of EWN are interconnected, so that it is possible to go from a given word in one language to a word expressing an identical or similar meaning in any of the other languages represented in EWN. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language-specific properties are maintained in the individual wordnets. The structure of EuroWordNet is shown in Figure 8, and described in more detail below in section 3.3.2.1.

EuroWordNet is not available for download without the purchase of a license and, for this reason, it is not possible to provide authentic examples from the database in this dissertation. Examples shown in this chapter thus either derive from sources describing the database, or are constructed from general descriptions. Hence, the French wordnet example in Figure 8 is constructed by the author as an equivalent to the Spanish wordnet example in the same figure. Note, however, that the individual language-specific wordnets do not necessarily have the same structure. For the same reason, the individual language-specific wordnets are not described here.

3.3.2.1 Structure of the Language-independent Part of EWN

EuroWordNet has 1024 common base concepts. The base concepts are concepts that form the mutual core of EuroWordNet. A base concept can be described as a basic concept in a given language-specific wordnet in terms of which other word meanings can be defined (Piek Vossen, 2001). The set of base concepts was selected through an iterative process in which local base concept sets were produced and compared. The concepts were given priority based on a number of criteria, e.g. the ones that had the highest position in an ontology (WordNet1.5 or a local taxonomy), or had the largest number of relations to other concepts were preferred. The process resulted in a common set of base concepts that are part of all the language-specific wordnets.

Ontologies

In addition to the common base concepts, the project includes a reduced set of 164 core base concepts, consisting of concepts that are present in 3 or more individual wordnets. Finally, the set of core base concepts has been reduced to 71 base types. This set has been achieved by removing unbalanced hyponyms (when a given concept (hypernym) has only one hyponym, it is considered an unbalanced hyponym) and by combining closely related synsets (e.g. act and action) into a single synset. The set of base types can be regarded as a minimal set of fundamental concepts or semantic primitives (Piek Vossen, 2001).

The domain ontology shown in Figure 8 contains knowledge for grouping synsets with respect to domain, e.g. *Natural sciences*, *Traffic*, *Sports*, *Hospital*, etc. The project report (Piek Vossen et al., 1998) notes that the domain ontology is not to be implemented in full during the project, but only in parts for illustration. In Figure 8, all ILI-records in this example should in principle map to the same top-concepts as well as to the same domain-concept, but this has been omitted for clarity of exposition. The full lines represent language-internal relations, the dotted lines represent interlingual relations, and the stippled lines represent language-independent relations. The Inter-Lingual-Index (ILI) contains all synsets from Wordnet1.5, plus some synsets that have been added in order to link base concepts from the individual wordnets to the index. The ILI is not ordered by any relations, but is merely a list of concepts. This non-structure is chosen in order not to make any assumptions that could potentially be in conflict with language-specific relations. The individual language-specific wordnets are linked to the ILI via a set of interlingual or equivalence relations. The most important equivalence relations are (Piek Vossen, Díez-Orzas, & Peters, 1997):

- EQ_SYNONYM

This relation holds if there is a 1-to-1 mapping between a synset and an ILI-record.

- EQ_NEAR_SYNONYM

This relation holds when a synset matches multiple ILI-records (a 1-to-many mapping), when multiple synsets match the same ILI-record

(a many-to-1 mapping), or when there is doubt about the precise mapping.

- EQ_HAS_HYPERONYM

This relation holds when a synset is more specific than any available ILI-record.

- EQ_HAS_HYPONYM

This relation holds when a meaning is more general than any available ILI-record.

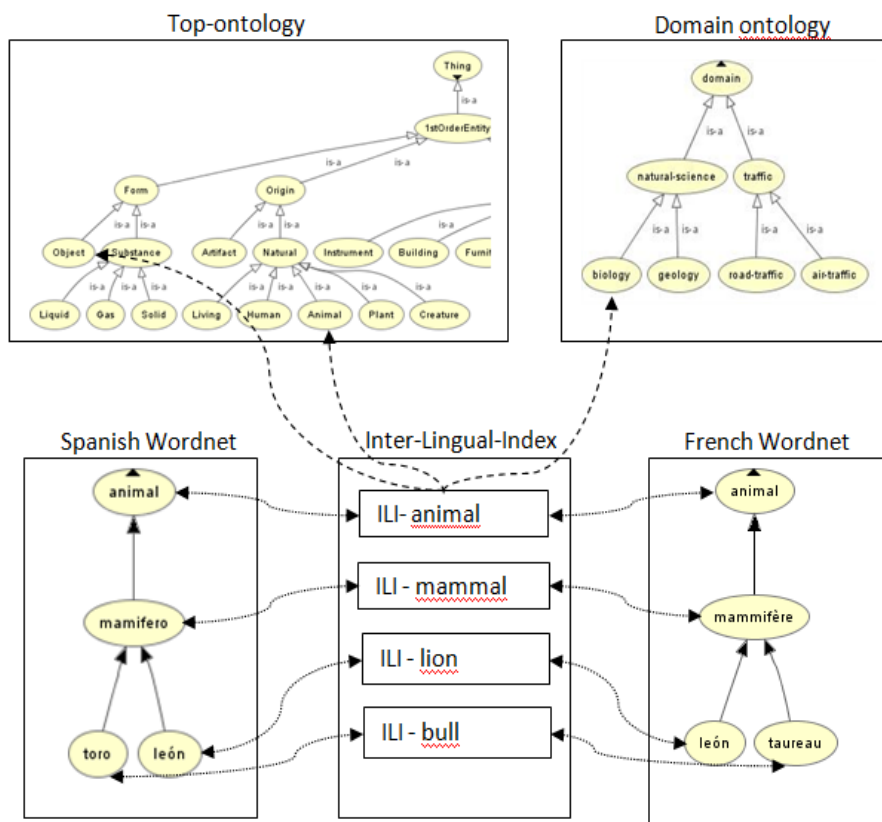


Figure 8 Structure of EuroWordNet. The figure is based on (Piek Vossen et al., 1997) and (Piek Vossen et al., 1998).

Ontologies

The common base concepts are specified in the form of ILI-records, and linked to a shared top-ontology. The purpose of this top-ontology is to provide a common semantic framework for all the included languages, while any language-specific characteristics are represented in the individual wordnets. The EuroWordNet top-ontology is a lattice structure consisting of 64 concepts. It is based on existing linguistic classifications, and has been targeted towards representing the diversity of the base concepts.

Wordnets are linguistic structures and, as such, they provide valuable information about the expressiveness of the language they describe, while this is not necessarily the case for formal ontologies or conceptual structures. For this reason, the EuroWordNet top-ontology incorporates semantic distinctions that play a role in linguistic approaches rather than in purely cognitive or knowledge engineering approaches. The top ontology is therefore based on familiar semantic classification paradigms: Aktionsart models, cf. e.g. (Levin, 1993; Pustejovsky, 1991b; Vendler, 1967; Verkuyl, 1972, 1989), entity orders (Lyons, 1977), Qualia-structure (Pustejovsky, 1995), as well as ontological classifications developed in other EU-projects such as Acquilex and Sift (Piek Vossen et al., 1998).

The first level of the top-ontology is divided into three types, based on (Lyons, 1977); 1stOrderEntity, 2ndOrderEntity and 3rdOrderEntity, as shown in Figure 9, and described in more detail below.

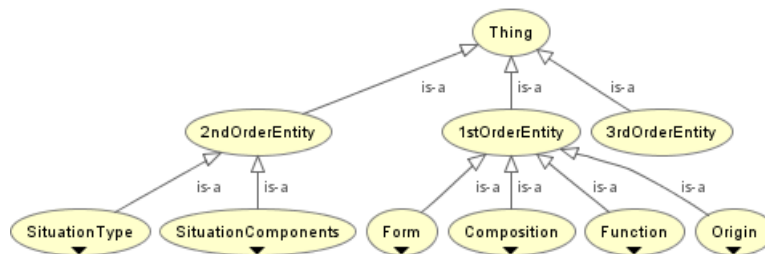


Figure 9 The first levels of the EuroWordNet top-ontology. The top node here labeled 'Thing', should in reality be labeled 'top'. The tool used to produce this figure (Protégé), however, did not allow for a renaming of the top node.

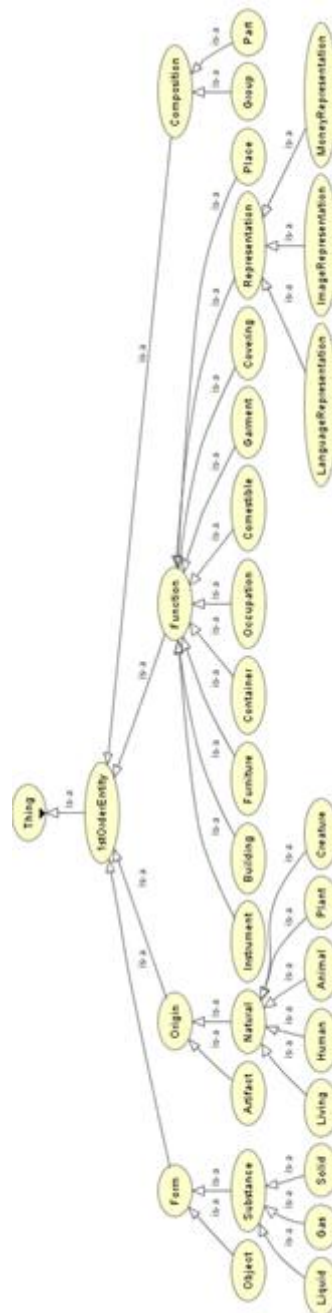


Figure 10 An unfolding of the concept *1stOrderEntity* in the EuroWordNet top-ontology

Ontologies

1stOrderEntity

1stOrderEntities roughly correspond to concrete, perceivable objects and substances, and are thus concrete entities (publicly) perceivable by the senses and located at a point in time in a three-dimensional space (Lyons, 1977).

The node *1stOrderEntity* specializes into the types *Origin*, *Form*, *Composition* and *Function* which are comparable with Qualia roles cf. (Pustejovsky, 1995). Base concepts may be classified by a combination of these four roles, and thus the top-concepts function more as features than as ontological classes (Piek Vossen et al., 1998).

2ndOrderEntity

States, situations and events.

Any Static Situation (property, relation) or Dynamic Situation which cannot be grasped, heard, seen, felt as an independent physical thing. They can be located in time and occur or take place rather than exist; e.g. continue, occur, apply.

3rdOrderEntity

3rdOrderEntities are mental entities such as ideas, concepts, knowledge.

An unobservable proposition which exists independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten. E.g. idea, thought, information, theory, plan.
3rdOrderEntities are not further specialized in the top-ontology.

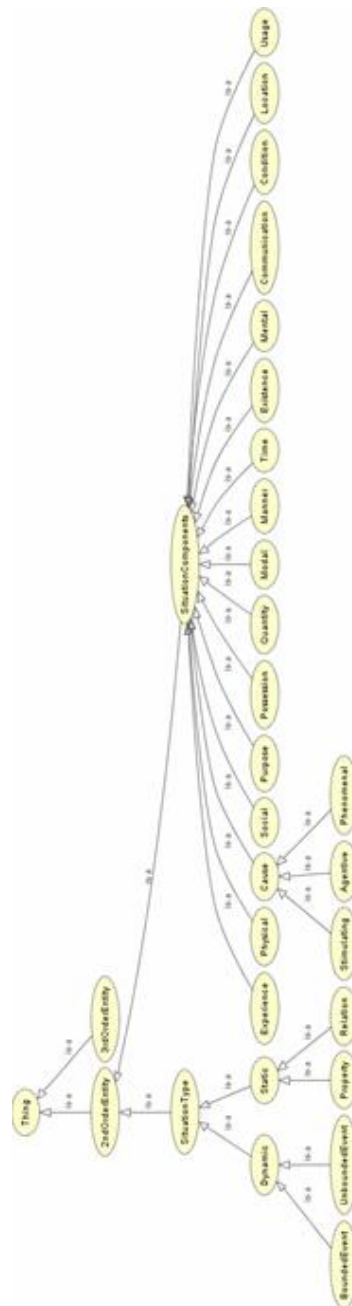


Figure 11 An unfolding of the concept *2ndOrderEntity* in the EuroWordNet top-ontology

Semantic features

As mentioned above, the top concepts should be seen as semantic features that can be applied disjunctively or conjunctively to synsets. For conjunctively added features, the result is a complex feature such as e.g. CONTAINER+PART+OBJECT+NATURAL.

Synsets consisting of words denoting concrete entities are classified as 1stOrderEntities, and are specified for the four Qualia Roles, cf. (Pustejovsky, 1995):

- Formal – “that which distinguishes the object within a larger domain”
- Constitutive – “the relation between an object and its constituents, or proper parts”
- Telic – “purpose and function of the object”
- Agentive – “factors involved in the origin or ‘bringing about’ of an object”

These roles correspond to the four top-ontology concepts FORM, COMPOSITION, FUNCTION and ORIGIN, and thus any specification of e.g. FORM would be in the form of a type below this concept. For *pod* (in the sense *a vessel that holds the seeds of a plant*), the specification with respect to these four roles would be CONTAINER+PART+OBJECT+NATURAL, where CONTAINER is a specification of FUNCTION, PART is a specification of COMPOSITION, OBJECT is a specification of FORM and NATURAL is a specification of ORIGIN.

Synsets consisting of words denoting abstract entities are classified as 2NDORDERENTITIES or 3RDORDERENTITIES. 2NDORDERENTITIES are further classified according to Aktions-Art by use of the top-ontology concepts SITUATIONTYPE and SITUATIONCOMPONENTS. Each 2ndOrder synset is classified according to one SITUATIONTYPE and a combination of SITUATIONCOMPONENTS, e.g.: CHANGE LOCATION would be classified as DYNAMIC+LOCATION+AGENTIVE(CAUSE)+PHYSICAL.

3.3.3 DanNet

DanNet, cf. e.g. (B. Pedersen et al., 2009; B. Pedersen & Sørensen, 2006; Bolette Sandford Pedersen, 2009; B. S. Pedersen et al., 2009), is a Danish language wordnet compiled by *Det Danske Sprog- og Litteraturselskab* (Society for Danish Language and Literature) under the Danish Ministry of Culture, and *Center for Sprogteknologi* (Centre for Language Technology), University of Copenhagen. The wordnet is structured according to the specifications for EuroWordNet, and the contents build on two existing data collections for Danish: *Den Danske Ordbog* (Hjort & Kristensen, 2003-5) and the Danish computational semantic lexicon *DK-SIMPLE* (B. S. Pedersen, 1999; Bolette Sanford Pedersen & Paggio, 2004). The DanNet project ran for four years from 2005-09 and was funded by the Danish Research Council. The wordnet is now being further developed under the *DK-CLARIN project*⁸.

The first version of DanNet, version 1.0, was released in March 2009 as an open source resource⁹, and the latest version so far, version 1.2, was released in March 2010. The following section describes the contents and structure up to version 1.1 (released July 2009), as this version is used for annotation of data in the experiments described in chapter 6.

3.3.3.1 Den Danske Ordbog

The primary source of words in DanNet is Den Danske Ordbog (The Danish Dictionary) (DDO) (Hjort & Kristensen, 2003-5). The dictionary is the most comprehensive work of contemporary Danish; it covers the period from 1955 to the present with its close to 100,000 lemmas. DDO is a corpus-based dictionary for which the lemma selection to a large extent is based on frequency in a corpus and, thus, the selection of words to be included in DanNet ‘inherits’ this frequency-based selection criterion. Apart from this obvious advantage, the dictionary has several features that make it suitable as a source of a wordnet. The typical structure of a DDO entry is shown in Figure 12. Without going deeper into the microstructure of this entry, we

⁸ ‘Common Language Resources and Technology Infrastructure’ (<http://dkclarin.ku.dk/>)

⁹ DanNet may be downloaded from <http://wordnet.dk>

Ontologies

see that it among other things contains an analytical definition (*billede malet med oliefarver, som regel på et stykke udspændt lærred, der bagefter indrammes* (picture painted in oil colour, usually on a stretched canvas, which is later framed)) as well as a synonym (*oliebillede*) and a sub-sense.

The screenshot shows the entry for 'oliemaleri' in the online version of the Danish Dictionary (DDO). The entry includes the following information:

- oliemaleri** substantiv, intetkøn
- BØJNING** -et, -er, -erne
- Betydninger**
- 1.** billede malet med oliefarver, som regel på et stykke udspændt lærred, der bagefter indrammes
 - SYNONYM** oliebillede
 - [han] maler dæmpede oliemalerier med ørkenmotiver eller heftige abstrakte billeder i krasse farver **BT90**
- 1.a** det at male billeder med oliefarver
 - rent oliemaleri kan ikke påvises før 1400-tallet **BoBedre92**

Figure 12 An example entry from the online version of DDO for the lemma *oliemaleri* (oil painting) (DSL, 2010)

From this information alone, firstly, it is possible to extract the information that a synset exists: {oliemaleri, oliebillede}. Further, if a semantic analysis is applied to the definition, it can be deduced that the concept that these words denote is a kind of BILLEDE, and that the concept distinguishes itself from other kinds of BILLEDE by a set of features: The material used in the production is *oil paint*, the medium is *stretched canvas* and it has a part which is a *frame*.

However, the DDO editors were forward-looking and included information in the entry files that was not intended for inclusion in the printed dictionary; e.g. information about the domain that the lemma pertains to, and, notably, *genus proximum* information was specified for all lemmas. While no ontology existed that could be consulted for this task, resulting in some inconsistency, the information was to a large degree useful for automatic hypernym extraction. Approx. 50 % of the material in the first DanNet version was semi-automatically extracted from DDO without further enrichment (DanNet, 2010).

3.3.3.2 The Semantic Lexicon SIMPLE

Another source of information in DanNet is the Danish SIMPLE lexicon DK-SIMPLE. The SIMPLE project (Semantic Information for Multifunctional Plurilingual Lexica) (Alessandro Lenci et al., 2000; A. Lenci et al., 1999; B. S. Pedersen, 1999; Bolette Sanford Pedersen & Paggio, 2004) was an EU-funded project which ran from 1998 to 2000. The objective of the project was to compile harmonized semantic dictionaries for natural language processing for 12 European languages; amongst these Danish. The project builds on the results of another dictionary project, namely LE-PAROLE, in which a basis consisting of morphological and syntactic description of 20.000 entries was developed. The aim of the Danish SIMPLE lexicon was to add a semantic layer to a subset consisting of 10,000 of these entries.

The structure of the SIMPLE lexicons builds on the fact that lexical items vary with respect to:

- 1) Number of meaning dimension
- 2) Number of senses

For 1), the different meaning dimensions are expressed by means of qualia roles cf. (Pustejovsky, 1995). Point 2) concerns polysemy, but is especially relevant with respect to regular polysemy. Regular polysemy exists when the same pattern of meaning change is found within a (semantically related) group of words. For example, all names of countries may denote a group of people (e.g. the government of the given country) or a geographical entity.

The SIMPLE lexicon constitutes three different types, which each receive different treatment. The types are:

- * Simple types
- * Unified types
- * Complex types

Simple types or basic categories correspond more or less to natural kinds, cf. (D. A. Cruse, 1986), and concepts with rigid properties, cf. (Guarino &

Ontologies

Welty, 2000). Such types are treated monodimensionally, and are thus only defined via the formal role (i.e. the hypomy relation). Examples of entries that are treated as simple types are *himmel* (*sky*), *søster* (*sister*) and *blomst* (*flower*). Unified types are treated multidimensionally. They are grounded on a simple type but with multiple coordinates (or qualia roles) added in the definition. Examples of entries that are treated as unified types are *biksemad* ((the dish) *hash*) and *lærer* (teacher).

Complex types are entries which exhibit regular polysemy. Thus, complex types allow for several semantic items to be included in a single lexical item by the addition of a feature *complex*. Examples of entries that are treated as complex types are *bog* (*book*) (SEMIOTIC ARTIFACT/INFORMATION), *kanin* (rabbit) (ANIMAL/DISH/FUR) and *universitet* (university) (INSTITUTION/HUMAN GROUP).

In other words, some word senses can be described by means of simple types where information is inherited from just one node in an ontology, and others are more complex, and inherit information from multiple nodes in an ontology. In SIMPLE, for each quale, a set of possible semantic relations were established, and further, templates were produced for each ontological type with suggestions to which semantic relations this type should include. In addition, the templates included slots for information about selectional restrictions, argument structure, derivational information, etc. The example in

Table 5 shows a template for the ontological type PHYSICAL CREATION which includes the verbs e.g. *build*, *construct*, *fabricate*, *manufacture*, as well as the noun *construction* (A. Lenci et al., 1999).

UseM:	1
BC Number:	182, 67
Template_Type:	[Physical_creation]
Template_Supertype:	[Creation]
Domain:	General
Semantic Class:	Creation
Gloss:	//free//
Event type:	transition
Pred_Rep.:	Lex_Pred (<arg0>,<arg1>,<arg2>)
Selectional Restr.:	arg0 = [human] arg1 = [concrete_entity] arg2 : default = [material] OR [substance]
Derivation:	<Derivational relation>
Formal:	isa (1,<UseM>:[Creation])
Agentive:	agentive_cause (1,<UseM>:[Cause])
Constitutive:	resulting_state (1,<UseM>:[Entity])
Telic:	<Nil>
Synonymy:	<Nil>
Collocates:	Collocates (<UseM1>,...<UseMn>)
Complex:	[Physical_creation] [Artifact]

Table 5 A SIMPLE encoding template for the ontological type PHYSICAL CREATION (A. Lenci et al., 1999)

The SIMPLE top ontology, or semantic type system, is based on the generative lexicon framework (Pustejovsky, 1995), where the complexity of semantic types is captured by means of qualia roles. Thus, the top-most layer of the type system is structured according to the four qualia roles *formal*, *constitutive*, *telic* and *agentive*.

Each entry in the SIMPLE lexicon is described minimally by a hyponymy relation to the closest language-specific node, as well as to a semantic type in the top-ontology and, thus, the lexicon in itself constitutes an ISA-hierarchy. This aspect has been exploited in the OntoQuery project, where the SIMPLE ontology was expanded with a set of concepts from the domain of nutrition and used in connection with content-based querying, cf. (Bolette Sanford Pedersen & Paggio, 2004). This ontology is the basis for the experiments described in chapter 5.

3.3.3.3 Contents and Structure of DanNet

The first version of DanNet contained 41,000 synsets, and the aim of the project is to reach 70,000 synsets. A large portion of the initial synsets, 27,000, describes nouns with a concrete sense, and while the wordnet thus

Ontologies

focuses on concrete nouns, it also contains synsets consisting of abstract nouns as well as of verbs and adjectives. For the first version, the distribution of the synsets between the included word classes was (DanNet, 2010):

- * 34,000 noun synsets (~26,460 lemmas)
- * 6,000 verb synsets (~3,100 lemmas)
- * 1,000 adjective synsets (~800 lemmas)

In version 1.1, ~10,000 synsets have been added. Most of these synsets are added without any further relations besides the `has_hyperonym` relation (plus any relations they may have inherited from their hypernyms) and ontological types. Also, the `is_instance_of` relation has been added and ~250 proper names of geographical areas and geopolitical entities have been added via this relation. All synsets have been described with hyponymy relations and ontological types, and subset of the synsets is linked to Princeton WordNet (acting as the ILI). The aim is to have 8,000 synsets linked to Princeton WordNet by the end of 2010.

Further, concrete nouns are linked to other synsets by an average of four semantic relations. As in SIMPLE, for synsets of specific ontological types, there is a defined template according to which for example all entities of a given ontological type are treated. Figure 13 shows an example of such a template for the complex ontological type VEHICLE+ARTIFACT+OBJECT.

Ontologisk type: VEHICLE +ARTIFACT+OBJECT

Test/forklaring	<i>John tog til Roskilde i en varevogn</i>
Eksempler:	<i>bil, vogn, fy, slæde, varevogn</i>
Kommentarer:	

Template

Topontologi:	Vehicle+Artifact+Object
Lemma i synset:	
Definition:	
Formal:	<i>has_hyperonym</i>
Constitutive:	<i>has_holo_part //optional//</i> <i>has_mero_madeof //optional//</i> <i>has_mero_part //optional//</i>
Agentive:	<i>made_by //optional//</i>
Telic:	<i>used_for</i>
Synonymi:	<i>near_synonym //optional//</i> <i>xpos_near_synonym //optional//</i>

Eksempel

Topontologi:	Vehicle+Artifact+Object
Lemma i synset:	<i>trillebør</i>
Definition:	<i>lille vogn med et enkelt hjul foran og to støtteben samt to håndtag bagved, brugt til at transportere mindre læs på fx en byggeplads el. ved havearbejde</i>
Formal:	<i>has_hyperonym vogn</i>
Constitutive:	<i>has_mero_part hjul</i> <i>has_mero_part støtteben</i>
Agentive:	<i>made_by fremstille</i>
Telic:	<i>used_for transportere</i>
Synonymi:	

Figure 13 A template for the ontological type VEHICLE+ARTIFACT+OBJECT in DanNet (B. S. Pedersen et al., 2009)

This means that for a given synset, if the ontological type is VEHICLE+ARTIFACT+OBJECT, in DanNet at least the relations *has_hyperonym* and *used_for* are specified, and further, the relations *has_holo_part*, *has_mero_madeof*, *has_mero_part*, *made_by*, *near_synonym* and *xpos_near_synonym* (the relation that links the synset to the ILI) may be specified.

The DanNet top-ontology is identical to the EuroWordNet top-ontology with very few differences; for example, the type BODYPART has been added as a hyponym to PART as shown in Figure 14 (B. S. Pedersen et al., 2009).

Ontologies

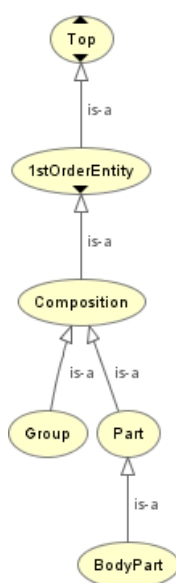


Figure 14 Excerpt of the DanNet top-ontology

Also, the relations that are used in DanNet are to a large degree identical to the relations used in EuroWordNet and Wordnet. Table 6 shows the DanNet relations where a bold type indicates that the relation has been added in DanNet (B. S. Pedersen et al., 2009).

<i>Relation</i>	<i>Example</i>
concerns	fodboldmål concerns sport
used_for	Hammer used_for hamre
used_for_object	clipse used_for_object clips
made_by	bagværk made_by bage
has_holo_madeof	mel has_holo_madeof brød
has_holo_member	partimedlem has_holo_member parti
has_holo_location	oase has_holo_location ørken
has_holo_part	øje has_holo_part ansigt
has_hyperonym	Birketræ has_hyperonym træ
has_hyperonym ortho	vejtræ has_hyperonym træ
has_mero_madeof	brød has_mero_madeof mel
has_mero_member	parti has_mero_member partimedlem
has_mero_part	hånd has_mero_part finger

has_mero_location	ørken has_mero_location oase
role_agent	passager role_agent rejse
role_patient	modtager role_patient modtage
involved_agent	violin involved_agent violinist
involved_instrument	violinist involved_instrument violin
near_synonym	si near_synonym dørslag
xpos_near_synonym	behandle xpos_near_synonym behandling
eq_has_synonym	bil eq_has_synonym car

Table 6 Semantic relations in DanNet (Pedersen, Braasch et al. 2009)

As shown in Table 6, DanNet has two types of hyponymy relations; a taxonomic and an orthogonal one. The taxonomic hyponymy relation holds between concepts that denote natural or functional kinds, and the orthogonal hyponymy relation relates concepts that denote nominal kinds to other concepts. The taxonomic hyponymy relation is identified by the test from (David Alan Cruse, 2002):

An X is a kind/type of Y

Similarly, the orthogonal hyponymy relation is identified by the test from (David Alan Cruse, 2002):

An X is a Y

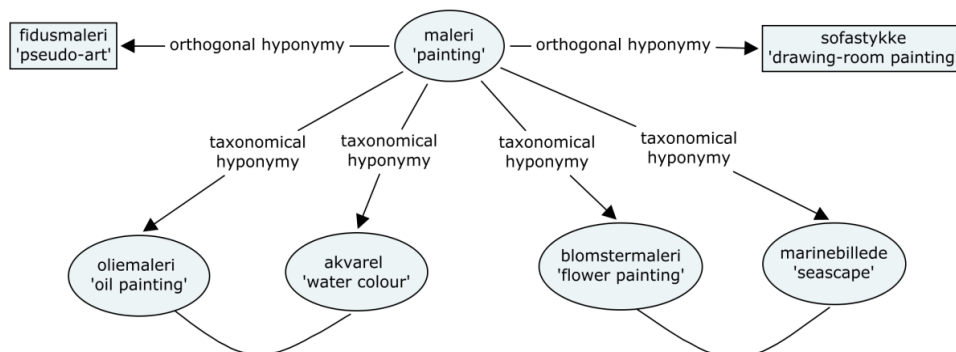


Figure 15 Orthogonal and taxonomical hyponyms of the concept MALERI (B. Pedersen et al., 2009)

Ontologies

Figure 15 shows an example of taxonomical and orthogonal hyponymic relations in DanNet. The orthogonal hyponymic relation is to be understood more as a level preserving relation than as a hierarchical relation. Any of the taxonomic hyponyms of MALERI could also be hyponyms of the orthogonally related types SOFASTYKKE or FIDUSMALERI, and thus the orthogonal hyponymy relation is perhaps more closely related to an equivalence/synonymy relation than it is to the taxonomic hyponymy relation. Figure 15 also illustrates the concept of paronymy (here shown as curved undirected arcs between concepts). Paronymy is a relation that holds between coordinate concepts in a conceptual cluster - they may correspond to a semantic field or a subdivision criterion (though unlabelled) as described above in section 3.2.2. A specific painting may at the same time be a WATER COLOUR and a FLOWER PAINTING, but it cannot at the same time be a WATER COLOUR and an OIL PAINTING since these two concepts are in the same conceptual cluster. The paronymic relation is not implemented in DanNet; however it is part of the modeling considerations insofar as one meaning dimension is chosen as a basis for the taxonomic structuring within a given subpart of the wordnet.

3.4 Summary

In this chapter, we have presented and explained the notion of an ontology as it is used in information theory as described as a formal representation of a set of concepts within a domain and of the relations that exist between them. We have referenced a number of definitions, and asserted that there is not yet complete consensus about a definition of the concept. Further we have presented two approaches to a categorization of ontologies; first the well known ontology spectrum by Lassila and McGuinness, for which we have presented a revised version, and second the ontology of ontologies by Madsen and Thomsen.

We have given a detailed description of some lexical ontologies, namely the wordnets Princeton WordNet, EuroWordNet and the Danish wordnet DanNet. We have described how wordnets have evolved from psycholinguistic models into proper ontologies. We have given lexical ontologies careful treatment in this chapter because we use such ontologies in our experiments described in chapters 5 and 6.

Chapter 4

Linguistic Expressions, Concepts and Semantic Relations

In this chapter, we will define linguistic expressions, concepts and semantic relations, such as those concepts are used in this dissertation. We will give an account of how we represent concepts and semantic relations in a generative ontology and discuss the notions of atomic and compound concepts. Further, we will discuss different aspects of problems that pertain to mapping from text to a conceptual representation in an ontology. Amongst these are the representation of relation denoting words such as verbs and prepositions as well as the representation of plurality denoting words.

4.1 Concepts and Relations as Signs

This section discusses how the notions of linguistic expressions, concepts and semantic relations are understood in this account. The relation between a linguistic expression and a concept or a relation is illustrated by means of the linguistic sign (cf. (Saussure, 1983)).

We will represent linguistic expressions in italics: *linguistic expression*, and concepts in small caps: CONCEPT. The combination of a linguistic expression and its associated concept, a sign, is represented in bold: **sign**. Since this presentation is concerned with the processing of written text and not speech, we will talk about linguistic expressions as sequences of letters when the cited works, e.g. (Saussure, 1983), also talk about sound patterns.

4.1.1 What is a Concept

Concepts exist in the minds of people, and are abstract ideas¹⁰ of entities in the world, cf. (Locke, 1690) as cited in (9). These ideas may be ideas of abstract or concrete entities, real or made up.

(9)

That men making abstract ideas, and settling them in their minds with names annexed to them, do thereby enable themselves to consider things, and discourse of them, as it were in bundles, for the easier and readier improvement and communication of their knowledge, which would advance but slowly were their words and thoughts confined only to particulars.
(Locke, 1690)

However, we have to assert that there is a large degree of consensus about the concept evoked by a given sound pattern or a sequence of letters in a group of language users, as communication otherwise would be impossible. This assertion is supported by experiments in the field of prototypicality theory (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). The experiments indicate that people to a large extent agree in a categorization task in which they are asked if different objects were good exemplars of a given category. These results support the idea that a common understanding exists of what a category or a general concept is.

In communication, we may refer to an idea by means of a label in the form of a sound pattern or a sequence of letters, but sometimes it is necessary to illustrate an idea without the use of language, e.g. in order to exemplify in a dictionary. Illustrating an idea is not always straightforward, but we can try to illustrate a given concept by means of a drawing or a photograph, as exemplified in Figure 16 and Figure 17.

However, such illustrations are in fact just visual representations of concepts, not very different from how sequences of letters are linguistic representations of concepts. Remember Magritte's painting 'Ceci n'est pas une pipe': The painting depicts a pipe, but yet it is NOT a pipe! It is merely a painting of a pipe. In the same way, Figure 16 is not a horse, but merely a

¹⁰ What Locke refers to through the word idea in (9) is somewhat broader than what is meant in today's use of the word and closer to today's use of the word thought. However, we will henceforth use the word idea to cover this meaning.

drawing of a horse representing one person's idea of what a horse is. It may also be seen as a prototypical representation of a horse.

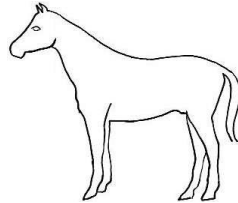


Figure 16 A drawing illustrating the concept horse



Figure 17 A photograph illustrating the concept horse

Figure 17, on the other hand, is not an illustration of an idea, but a photographic representation of a single referent of the concept HORSE. However, it may suffice in illustrating the concept by example.

To conclude, illustrations are not perfect means of representing concepts. However, for ease of exposition, and *faute de mieux*, we may choose to represent concepts as drawings in the following.

4.1.1.1 Concepts, Linguistic Expressions and Linguistic Signs

Concepts are diffuse and, in order to talk about them, we need to associate them with physical representations manifested as sounds or sequences of letters (cf. (9)).

Sequences of letters such as *horse*, *house*, *roof* or *loknap* have no meaning in isolation, but are simply sequences of letters. However, they may act as names of concepts that make it possible to talk about these concepts, and in that case they have a meaning. The chosen name for a

Linguistic Expressions, Concepts and Semantic Relations

given concept in a given language is mostly arbitrary¹¹, the key point being that it is used collectively as a name for the same concept within a group of language users (a geographical, social, etc. group).

Sequences of letters acting as concept names in natural language are termed linguistic signs by semioticians (cf. e.g. (Saussure, 1983)). The saussurean dyadic model of linguistic signs is an adequate model for explaining the relation between expression and meaning though it has its shortcomings e.g. in that it lacks compositionality. Thus, in the present account, we will use the notion of a linguistic sign as a means to illustrating the connection between expression and meaning, without going deeper into any shortcomings of the theory in relation to our treatment of text.

Linguistic signs consist of two parts: the signifier (the sequence of letters or sounds)¹² and the signified (the concept that exists in our brain, and is evoked when we read or hear the signifier). The sign is a unity, the structure of which is shown Figure 18. Note that linguistic signs are not connected to the referents in the world.

Horse is a linguistic sign, as illustrated in Figure 19: it has a linguistic form (or signifier): the letter sequence *h o r s e*, and a meaning (or signified): it denotes the concept HORSE. The sequence of letters *loknap*, in my idiolanguage at least¹³, is not associated with a concept, and thus it does not constitute part of a linguistic sign.

¹¹ An exception is perhaps onomatopoeias, which are not completely arbitrarily chosen, but are imitations of sounds made by or connected with the referents of a concept. However, these sounds may be perceived and at least reproduced differently from language to language, resulting in differences such as *croak* and *kvæk* for the sound made by a frog in English and Danish respectively.

¹² Saussure spoke solely about the sign as linking a concept and a sound pattern, and the link between a letter and a sound as a sign in itself. However, we modify the model here, and construe the link between a sequence of letters and a concept as a sign.

¹³ When discussing this with students, for almost any linguistic expression that I can come up with that is not associated with a concept in my mind, a student finds some example of the linguistic expression being associated with a concept – often in connection with fantasy games. Therefore, I find it necessary to hedge this assertion by explicitly stating that *loknap* is not associated with a concept in MY mind.

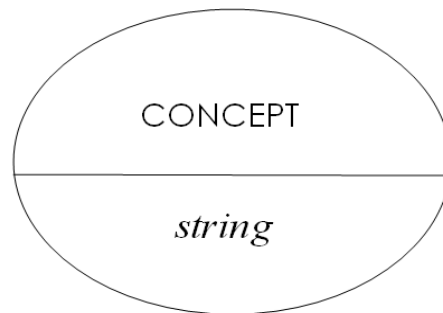


Figure 18 The general linguistic sign

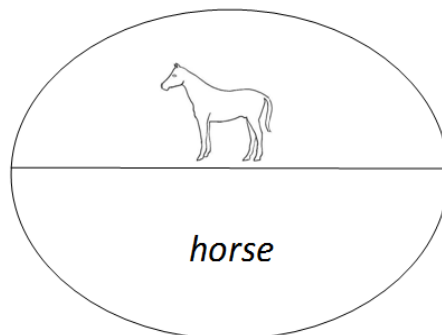


Figure 19 A specific linguistic sign 'horse'

As a curiosity, the Danish philosopher Carsten Graff has compiled a dictionary of words which denote concepts that most people allegedly are familiar with, but which are not (yet) part of the language (Graff, 2009). As an example, the dictionary contains the (Danish?) word *Jilousa* which is defined as *Den omvendte følelse af jalousi – dvs. glæden ved at opleve at den man elsker får kærlighed af andre* (The inverse feeling of jealousy – i.e. the joy of experiencing that the person you love is being loved by others). However, as thought-provoking as this dictionary may be, it is our bold claim that at a given point in time, a given language contains exactly the words that are necessary for communicating what needs to be communicated by the users of that language at that given point in time. All living languages have the ability to have new words or phrases added when

needed, and thus if a (creative) language user needs to communicate a thought that no known word expresses, he or she can coin a word or a phrase that describes it.



4.1.2 Semantic Relations

In Gulliver's Travels (Swift, 1726), Gulliver travels to an academy in Lagado where different projects are researched in order to improve the lives of the inhabitants. In one project concerned with shortening of discourse, a professor attempts to remove all the elements of language, except nouns, since all things imaginable are nouns.

While there is some grain of truth in the idea that all things imaginable are nouns, the whole scheme is, of course, ludicrous. If we were only able to express ourselves using nouns, we would not be able to express how the ideas denoted by these nouns relate to each other. For this, we need verbs, prepositions, etc.

Semantic relations¹⁴ can be understood as the conceptual glue that binds concepts together in discourse. Without the possibility of expressing such relations, we would only be able to state which concepts we are discoursing about, and not how these concepts interrelate.

It is important to note that semantic relations hold between concepts and not between words. In this account, we distinguish between semantic relations and lexical relations, which hold between words (e.g. synonymy). Figure 20 illustrates a lexical relation, namely the synonymy relation, holding between the words *hack* and *nag*. No relation holds between the signified levels of

the signs,  and , since, as is the case for true synonyms at any rate, the signified levels are congruent, and only the signifier is different.

Above, we have seen that linguistic expressions may be paired with concepts in order to form linguistic signs. These signs we will henceforth refer to as conceptual signs, and in the following, we claim that linguistic expressions may also be paired with semantic relations in order to form what we will call relational signs. This division of the notion of a linguistic sign into conceptual and relational signs is not referable to Saussure.

¹⁴ In the following, the term *semantic relation* is abbreviated to *relation*.

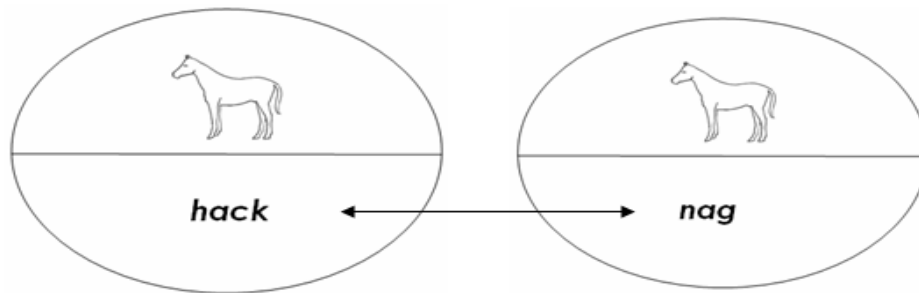


Figure 20 A synonymy relation between the words hack and nag

4.1.2.1 Relational Signs

Relations may have explicit linguistic realizations, but they may also be implicitly present without linguistic realizations. In the latter case, we say that they have a null-realization (marked ϵ). In the following, we will refer to the sign that consists of either a linguistic expression or a null-realization and a relation as a relational sign. Relational signs are thus analogous to conceptual signs, apart from the fact that they pair linguistic expressions (or null-realizations) with semantic relations instead of concepts.

As illustrated in Figure 21 and Figure 22, in the relational sign, relations are comparable to the signified level of the conceptual sign, and the linguistic form that denotes the relation is comparable to the signifier level. Figure 21 illustrates a relational sign consisting of a relation and a linguistic expression, and Figure 22 illustrates a relational sign consisting of a relation and a null-realization. A relation acts as a relator which relates concepts (its relata (sg. relatum)).

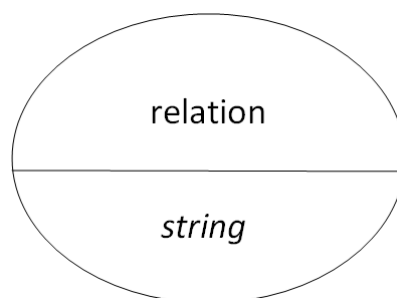


Figure 21 A linguistic expression-realization of the signifier in a relational sign

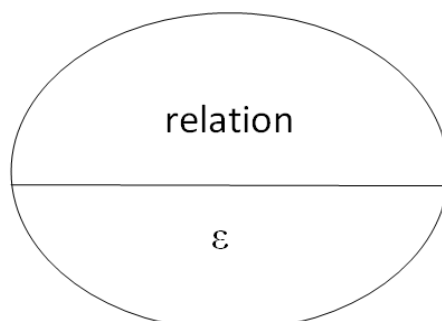


Figure 22 The null-realization as signifier in a relational sign

As mentioned in chapter 2, certain word classes (the relator classes), e.g. verbs and prepositions, consist of words that denote relations and words from these word classes thus constitute explicit linguistic realizations of relations.

(10) *The hair of the horse*

In example (10), we can identify two linguistic expressions that denote concepts (*hair* and *horse*) and one linguistic expression that denotes a relation (*of*). In this context, the linguistic expression *of* denotes a partitive relation. In other cases, the partitive relation may not be explicitly realized, e.g. for compounds as in (11).

(11) *Horsehair*

For the compound word in (11), we can identify two simplex words that denote concepts (*horse* and *hair*), but no explicit realization of the relation that holds between the two concepts denoted by the linguistic expressions. However, an implicit partitive relation exists between the two concepts HORSE and HAIR, and the meaning of the compound is the same as if the relation had been realized as in (10). Thus, in this case, only the relata are explicitly realized, the relator has a null-realization ε .

Relations between concepts may be realized at different syntactic levels; across sentence boundaries, as in (13), or within the boundaries of a sentence, a phrase or a word. The relations can be denoted by different parts of speech, such as a verb, a preposition, an adverb, an adjective or a possessive pronoun, or they can be inherently present in compounds, etc, as exemplified above in (11). Also clitica (e.g. genitive markers) can denote relations.

- (12) *A black horse.*
- (13) *Peter owns a horse. It is stubborn.*
- (14) *Peter gave the horse a carrot.*
- (15) *The horse in the field.*
- (16) *Peter's sister.*
- (17) *Peter's horse.*

Relations are n-ary, which is exemplified in examples (12)-(17) above, and glossed below: In (12), the relator is realized as the adjective *black* which denotes a unary relation where HORSE is the relatum. In (13), the relator is realized as the verb *owns* which denotes a binary relation where PETER and HORSE are the relata, and in (14), the relator is realized as the verb *give* which denotes a ternary relation where PETER, HORSE and CARROT are the relata. In (15), the relator is realized as the preposition *in* which denotes a binary relation where HORSE and FIELD are the relata. Examples (16) and (17) are both genitive constructions, however, we do not analyze them in the same manner (cf. (Vikner & Jensen, 2002)): In (16), the relator is inherently present in the relational noun *sister*. This relation that comes from the relational noun *sister*, which could be labeled HAS_SIBLING, denotes a binary relation where, in this case, PETER and his sister are the relata. Such nouns are difficult to characterize as either being part of a relational or a conceptual sign, but must be characterized as being both. In (17), the relator is realized as the clitical genitive marker *'s* which denotes a binary possession relation where PETER and HORSE are the relata.

4.2 Representation

We represent concepts and relations between concepts in the ontology description language ONTOLOG. As described in chapter 3, conventionally, ontologies are fixed ordering structures among concepts identified in the

domain of application; however, in this account¹⁵ we consider what we call generative ontologies. In the following section 4.2.1, we introduce the notion of a generative ontology and give a formal characterization of that notion.

4.2.1 Generative Ontologies and ONTOLOG

The term *generative ontology* is analogous to the terms *generative grammar* (e.g. (Chomsky, 1957)) and *generative lexicon* (Pustejovsky, 1991a, 1995) with good reason: A generative grammar, for a given language, offers a set of rules that will predict the set of possible combinations of words and phrases that form grammatical sentences in that given language. The generative lexicon, in turn, provides a framework for the composition of lexical meanings, thereby defining the well-formedness conditions for semantic expressions in a language (Pustejovsky, 1995). Thus, a generative grammar provides rules for combinations of words and phrases, and is only concerned with the surface level, and a generative lexicon provides rules for interpretation of combinations of lexical meanings, and is thus both concerned with the surface and the conceptual level, while, as will be described below, a generative ontology provides rules for combinations of concepts and relations, and is thus only concerned with the conceptual level (cf. Table 7).

	Surface level	Conceptual level
Generative grammar	X	
Generative lexicon	X	X
Generative ontology		X

Table 7 The levels of concern for various generative frameworks

The well-known example from (Chomsky, 1957), *Colorless green ideas sleep furiously* is an example of a grammatical sentence whose semantics is nonsensical, which would not receive any well-formed readings in the framework of a generative lexicon nor in a generative ontology framework

A generative ontology is to be understood as an ontology where it possible to introduce new compound concepts as they become known. Basically, a

¹⁵ The framework presented here takes place within the framework of the SIABO project, (Andreasen, Bulskov, Lassen et al., 2009).

generative ontology consists of a given finite ontology ordered by the ISA inclusion relation called the skeleton ontology, and a set of production rules (cf. generative grammars) that allows for production of compound concepts. The skeleton ontology consists of atomic concepts. Thus, a generative ontology is a non-finite set of concepts, and we therefore move from finite ontologies to infinite concept systems reflecting the recursive productivity of the phrase structures in natural language. Hereby, it becomes possible to map compound concepts denoted by complex linguistic structures into nodes in the ontology. This unlimited recursive productivity is achieved by applying a finite set of semantic relations that may act as attributes to a concept forming compound concepts. These semantic relations make it possible to produce concept feature structures which correspond to complex linguistic forms.

We define a generative ontology by generalizing the hierarchy to a lattice and by introducing a (lattice-algebraic) concept language. This language defines an extended set of well-formed concepts, including both atomic and compound term concepts.

We represent concepts and concept relations in an ontology using a lattice-algebraic concept language called ONTOLOG (NILSSON, 2001). The language introduces two closed operations, *sum* and *product* on concept expressions φ and ψ , where

- conceptual sum ($\varphi + \psi$) is interpreted as the concept being either φ or ψ
- conceptual product ($\varphi \times \psi$) is interpreted as the concept being φ and ψ

‘Conceptual sum’ and ‘conceptual product’ are also called ‘join’ and ‘meet’, respectively. The set of relations R is introduced algebraically by means of a binary operator ($:$) known as the Peirce product ($r : \varphi$), which combines a relation r with an expression φ . The Peirce product is used as a factor in conceptual products, as in $x \times (r : y)$, which can be rewritten to form the feature structure $x [r : y]$, where $[r : y]$ is an attribution of the concept x . Compound concepts may be formed by attribution.

Thus, a compound concept takes the form of a concept feature structure:

$c[r1:c1, r2:c2, \dots, rn:cn]$

where c is a concept from the skeleton ontology, and $r1, r2, \dots$ are semantic relations, and the $c1, c2, \dots$ are concepts or themselves recursively feature-structured concept terms. The recursive structure admits unlimited production of concept terms from a given finite supply of concepts and relations. This is reminiscent of the phrase structured production of sentential forms by means of rules in a generative grammar. Each concept term is associated with a node in the generative ontology.

The attributions $[r1:c1, r2:c2, \dots]$, which consist of pairs of relations and concept arguments, function as conceptual restrictions on the core concept c . This means that the term $c[r1:c1]$ is always situated below the node c in the ontology. This way, the concept terms open new paths stretching downwards towards increasingly specialized concepts in the given ontology.

However, the generative ontology should not admit arbitrary combinations of relations and concepts: We thus propose ontological affinities that may specify ontologically admissible ways of combining concepts. Ontological affinities are here specified as triples: $\langle c', r, c'' \rangle$.

Thus, given a minimal skeleton ontology:

saddle ISA device ISA artefact ISA phys_entity ISA entity
pony ISA horse ISA animate ISA phys_entity ISA entity
black ISA color ISA quality ISA entity

In addition to the ordering relation ISA, we have have a closed set of semantic relations:

{LOC, CHR, WRT}

And finally, a set of ontological affinity specifications:

$\langle \text{phys_entity}, \text{LOC}, \text{phys_entity} \rangle$
 $\langle \text{entity}, \text{CHR}, \text{quality} \rangle$
 $\langle \text{entity}, \text{WRT}, \text{entity} \rangle$

We are able to produce an infinite set of concept feature structures:

```
{ PONY[CHR:BLACK], PONY[CHR:COLOR]}, PONY[CHR:QUALITY],  
SADDLE[LOC:PONY], SADDLE[LOC:HORSE], SADDLE[LOC:ANIMATE],  
SADDLE[LOC:PHYS_ENTITY], SADDLE[LOC:PONY, CHR:BLACK],  
SADDLE[LOC:PONY[CHR:BLACK]],  
DEVICE[LOC:SADDLE[LOC:PONY[CHR:BLACK]]], DEVICE[CHR:BLACK,  
LOC:SADDLE[LOC:PONY[CHR:BLACK]]], PONY[CHR:BLACK, CHR:COLOR,  
CHR:QUALITY], PONY[WRT:PONY], PONY[WRT:PONY[WRT:PONY]], ... }
```

Thus, the skeleton ontology becomes generative by being supplemented with rules admitting the production of compound concepts represented as concept feature structures.

If we were to apply such rules in order to produce all conceivable concepts in the world, we would need more elaborate affinity specifications than the ones exemplified above, ruling out circular or pleonastic concept feature structures as e.g. PONY[WRT:PONY] and PONY[CHR:BLACK, CHR:COLOR, CHR:QUALITY], as well as plain nonsensical concepts. However, the perspective here is recognition of concepts in text and not concept generation, which makes the need for such elaborate specifications less important. The problem is comparable to linguistic grammars for language generation vs. analysis, where the former requires a stricter grammar than the latter. Our approach does not involve concept production, but simply provides a generative ontology framework which allows us to map compound concepts identified through analysis of text into appropriate positions in an ontology.

Our automatic text analysis is conducted chiefly by means of conventional linguistic grammars assisted by a generative ontology. Ideally, a sentence is turned into a term in the generative ontology which is supposed to represent the conceptual content of the sentence.

Here, the term *conceptual content* is to be viewed in distinction to the term *propositional content*. At the present state, our representation of the conceptual content of a given linguistic expression disregards such features as number, determination, negation and quantification, as we consider these features less central for search purposes. We argue that for a given search, texts expressing e.g. negations of the search phrase would be equally relevant as the opposite. Thus, in our framework, the linguistic expressions in (18) and (19) would have the same representation, namely (20).

(18) *Peter rides his horse*

(19) *Peter does not ride horses*

(20) RIDING[AGT:PETER, PNT:HORSE]¹⁶

4.2.2 The Relation between the Sign and the Ontology

The duality of a sign represents the tight relation between linguistic expressions in text and concepts and relations in the ontology. The following sections describe different aspects of this relation: Section 4.2.3 discusses different criteria we may choose to set up for deciding whether a given concept should be represented as an atomic or a compound concept in the ontology. Section 4.2.4 discusses how we may treat unknown words and the corresponding unknown concepts in an indexing process.

Within the framework of an ontology-based information retrieval system, our purpose is to map text chunks into appropriate nodes in a generative ontology, in order to make it possible to index texts according to their conceptual content. The purpose of the indexing is to facilitate retrieval of texts as answers to a query, where the texts match the query to some extent. In order to do this, we analyze documents sentence by sentence, and produce one or more corresponding expressions in the formal ontology language, ONTOLOG. An ONTOLOG expression represents a node in a generative ontology and acts as a definition of the concept.

4.2.3 Atomic and Compound Concepts

As described above, the formal ontology language ONTOLOG has two elements, concepts and relations: Concepts may be atomic or compound, and two types of relations exist, namely the subsumption relation 'ISA', and a given set of associative relations. Two concepts (atomic or themselves compound) may be related through associative relations, and thus form compound concepts. Such relations are formally represented as conceptual feature structures with attribute:value pairs, where the attribute is the relation, and the value is the related concept.

¹⁶ The conceptual content of the verb *ride* is here reified, resulting in the concept RIDING, cf. the account in section 4.3.

In a directed graphical model of an ontology, we label atomic concepts with their corresponding linguistic expressions, typically in the form of continuous strings. These linguistic expressions represent the natural language expression connected with the concept. Compound concepts are labeled with ONTOLOG expressions. Associative relations are shown as directed dotted arrows pointing towards the related concept, and labeled the name of the relation. The labels of associative relations are normally 3-letter abbreviations of the relation name, printed in capitals. The subsumption relation is shown as a directed solid arrow pointing towards the superconcept, and labeled 'ISA', as exemplified in Figure 23.

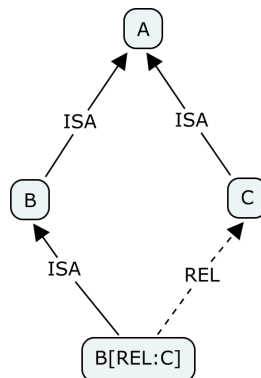


Figure 23 Ontology fragment showing the concepts A, B, C and B[REL:C]

Figure 23 shows a fragment of an generative ontology with the atomic concepts 'A', 'B' and 'C' and the compound concept 'B[REL:C]'. The sibling concepts 'B' and 'C' are both subsumed by the concept 'A', while the compound concept 'B[REL:C]' is subsumed by the concept 'B' and is related to the concept 'C' via an associative relation. The compound concept 'B[REL:C]' is defined as “a kind of ‘B’ that has the relation REL to ‘C’”.

There are no hard and fast rules for the types of linguistic expressions that form signs in combination with atomic or compound concepts, respectively. However, continuous strings (i.e. linguistic expressions that do not contain blanks) are often paired with atomic concepts in order to form signs, such as the linguistic expression *horse* and the concept HORSE. On the other hand, compound concepts may also be paired with continuous strings, such as

stallion. A stallion may be defined as HORSE[CHR: MALE]¹⁷, which states that this concept is an entity that is a HORSE which has the characteristic MALE. Especially for a language such as Danish that allows for a dynamic formation of compound words as continuous strings, such linguistic expressions may often be paired with compound concepts.

Atomic concepts may also be paired with discontinuous strings (i.e. linguistic expressions that do contain blanks), this is often the case for stable compounds such as *wild horse*, for which the corresponding concept is the atomic concept WILD_HORSE, that can be defined alone by stating that it is a type of HORSE.

Finally, discontinuous strings may be paired with compound concepts, e.g. *black horse*, for which the corresponding concept is HORSE[CHR: BLACK], as illustrated below in Figure 24.

Concepts may be paired with linguistic expressions of arbitrary length: from single continuous strings to sentences or entire texts.

It is important to keep in mind that the generative ontology allows us to dynamically add concepts to our ontology when we acknowledge their existence, as long as the basic building blocks as well as the relevant ‘glue’ in the form of relations are available. The basic building blocks are the atomic concepts that any given compound concept decomposes into. From an economy point of view, this means that we should not add more atomic concepts in the skeleton ontology than we need in order to model a given domain. However, it makes sense to add concepts denoted by lexicalized expressions in the domain, that will frequently occur as building blocks of concepts denoted by ad hoc expression formations.

For example, we may choose to define the concept WILD_HORSE alone by stating that it is a type of HORSE. We know that wild horses distinguish themselves from other members of the family Eqidae by at least one feature, namely the feature(s) that justifies the division of the family Eqidae into subspecies such as zebras, donkeys, wild horses, etc. This type of knowledge could be added to the ontology in the form of feature:value pairs. However, we may decide not to include this knowledge in our model, and

¹⁷ A different modeling could represent a stallion as an entity that has a hypernymy relation to both HORSE and MALE, and could thus be represented as HORSEXMALE

simply state that a WILD_HORSE is a type of HORSE. In doing so, we decide to represent the concept WILD_HORSE as an atomic concept. This type of model depicts a simplified view of the world, but we may choose to model the world in this manner if we find it suitable.

A possible approach, which we will apply here, is that concepts denoted by lexicalized linguistic expressions should be represented as atomic concepts, and that concepts denoted by non-lexicalized expressions should be represented as compound concepts. In order to decide whether a multi-word expression is lexicalized or not within a given domain, we can choose different approaches: We may apply a statistic measure (e.g. mutual information, chi-squared, log-likelihood or t-score) that can give us an indication based on the frequency of a given collocation compared to the frequency of its parts, or we may consult domain-specific glossaries and lexicons.

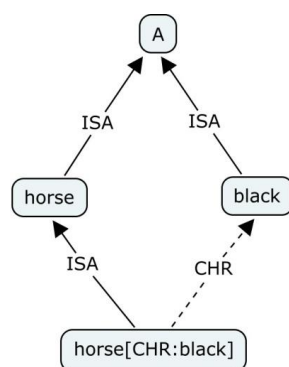


Figure 24 Ontology fragment showing the concepts horse, black and HORSE[CHR:BLACK]

For example (12) above, the concept denoted by the linguistic expression *A black horse* would be represented as a compound concept if the building blocks in order to form a compound concept are available, as illustrated graphically in Figure 24.

4.2.4 Treatment of Unknown Words and Concepts

In principle, only compound concepts may be added to our generative ontology. However, when analyzing natural language texts, we sometimes run into unknown words. For every given atomic concept in our ontology, there is information about the signifier level. This information may be in the form of a set of synonymous words that all denote the given concept (cf.

synsets in a wordnet, as described in chapter 3). Thus, if we do not know a given word, i.e. we do not have information about a corresponding concept, it may be the case that the concept it denotes is in not in the ontology, or it may be that the expression at hand is missing from a synonym set while the relevant concept is in fact in the ontology. For a given unknown word, we have no way of deciding whether the corresponding concept is or is not in the ontology, and we thus treat all unknown words as if the concept is absent. There are two possible treatments of unknown words/concepts: We may either disregard them in the conceptual indexing process, or we may map them to a provisional position in the skeleton ontology. The first solution can turn out to be problematic when we wish to map the conceptual content of larger text chunks in which a word is unknown, to the ontology in the form of compound concepts. This is especially true if the given unknown word is the head of the phrase in question, because heads of phrases become core concepts and thus cannot easily be ignored. Thus, we propose a heuristics for adding atomic concepts denoted by unknown words to the skeleton ontology: We propose to map the conceptual content of a given unknown word to a new concept subsumed by a dummy-concept positioned immediately below the top concept. We provisionally label this dummy concept `EVERYTHING_ELSE`, and the added concept is labeled the expression that gives rise to the concept.

Figure 25 shows a mapping of the unknown word *pedometer* to a new concept `PEDOMETER` which is subsumed by the dummy concept `EVERYTHING_ELSE`.

This new `PEDOMETER` concept allows us to represent the conceptual content of e.g. the linguistic expression *a pedometer for the estimation of walking distance* as a compound concept `PEDOMETER[PRP:ESTIMATION[WRT:DISTANCE[WRT:WALKING]]]` with the position in the ontology as shown in Figure 25.

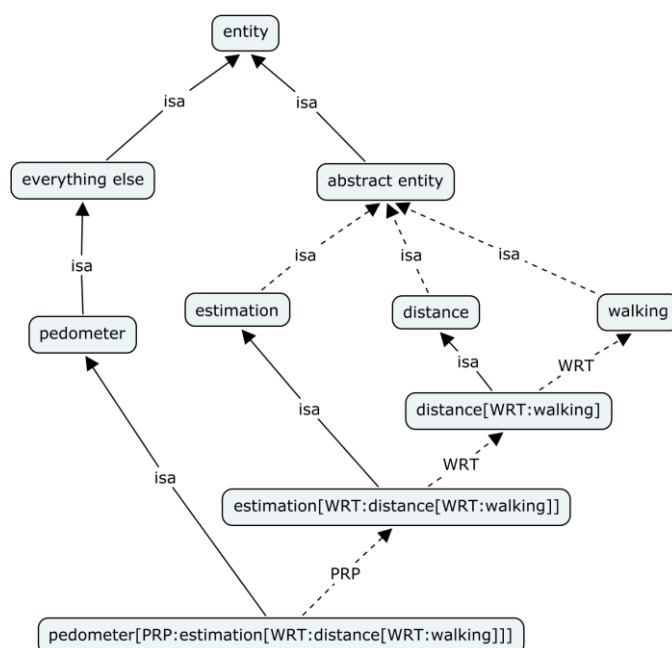


Figure 25 Position of the compound concept
 PEDOMETER[PRP:ESTIMATION[WRT:DISTANCE[WRT:WALKING]]
]

4.3 Relation Denoting Words

This section and the following section 4.4 describe considerations on two aspects of mapping text to an ontology: How do we treat relation-denoting words such as verbs and prepositions, and how do we treat plurality-denoting words.

This section describes how and why we treat the relational word classes verbs and prepositions differently. Relations that verbs express are reified, i.e. turned into concepts, while relations expressed by prepositions are treated as associative relations.

Section 4.3.1 gives an account of the arguments for treating verbs the way we do, and section 4.3.2 gives an account of our treatment of prepositions.

4.3.1 Verbs

Verbs are commonly viewed as the heads of sentences, where they assert some relation between the subject and possible complements and/or adjuncts (i.e. the arguments of the verb relation). The relations expressed by verbs typically denote events, actions, or states.

In a traditional grammarian's view (cf. e.g. (Greenbaum & Quirk, 1990)), verbs may take zero, one or two complements (in a variety of complementation patterns), and thus the arity of a relation denoted by a verb may be unary, binary or ternary. However, in our view, we include adjuncts as possible arguments of the verb relation, and the arity of such relations thus becomes n-ary.

For several important reasons, explained below in sections 4.3.1.1 to 4.3.1.5, relations expressed by verbs are reified, i.e. turned into concepts, in our framework. For a given verb¹⁸, the thematic roles of the arguments become associative relations, and the concepts that the arguments map to become related to the verb-denoted concept via these relations. For a detailed description of the application of lexical resources to this task, see (Andreasen, Bulskov, Jensen, & Lassen, 2009).

The following provides an example by way of the verb *build*. According to VerbNet¹⁹, the following thematic roles may play a part in a sentence headed by the verb *build*:

Agent
Material
Product
Beneficiary
Asset

Thus, in our approach, the reified 'build relation', the concept BUILDING, may be related to other concepts via the associative relations AGENT

¹⁸ Here, the term 'verb' covers only 'main verbs', and excludes the verb 'be' which receives a separate treatment, as well as modal and auxiliary verbs.

¹⁹ VerbNet class build-26.1

(AGT), MATERIAL (MAT), PRODUCT (PRD), BENEFICIARY (BEN) and ASSET (ASS).

According to VerbNet, for the frame syntactic construction (frame) 'NP-V-NP', where V is a form of a verb in the build-26-1 class, the subject and object NPs denote the thematic roles AGENT and PRODUCT respectively.

(21) *Peter built a house*

Thus, for example (21), the subject NP *Peter* plays the role of AGENT in the relation denoted by the verb *built*, and the object NP *a house* plays the role of PRODUCT.

An ONTOLOG representation of the conceptual content of example (21) becomes:

(22) BUILDING[AGT:PETER, PRD:HOUSE]

where the 'build relation' becomes reified such that a formal concept BUILDING²⁰ is attributed with attribute:value pairs for any thematic role that is identified in the text in example (21), namely AGT:PETER and PRD:HOUSE.

The VerbNet build-26-1 class has 12 different frames with defined roles, and we will not go through all of them here. However, later we will look at examples that match the frame 'NP V NP PP', where the preposition heading the PP must be in the set {from out_of}. For this frame, the roles are distributed in the order AGENT, PRODUCT and MATERIAL.

The following sections 4.3.1.1 to 4.3.1.6 discuss various reasons for reifying relations denoted by verbs.

²⁰ The label on the concept resulting from the reification of the verb relation is typically a nominalized form of the verb.

4.3.1.1 Limited Set of Relations

We wish to keep the set of relations small and closed. A small and closed set of relations is important, e.g. in a search setting where computation of semantic similarity between a query and texts would be very complex with a large or open set of relations.

The class of English verbs is an open class and furthermore dynamic, meaning that new verbs are frequently added to the language and old ones disappear.

The approximate size of the set of verbs in the English language is not easy to assess, but as an indication, the British National Corpus (BNC) has 1281 verbs among the lemmas with more than 800 occurrences in the whole 100M-word corpus²¹, and VerbNet currently describes 5751 verbs arranged in 270 VerbNet main classes and 200 subclasses.

Thus, if corresponding to the set of verbs, the set of relations would be a large and ever changing set which contrasts with our wish to keep the set of relations small and closed. There are two possible solutions to this problem: Abstraction and reification.

If all relations denoted by verbs were abstracted to general relations, by way of e.g. Levin's verb classification, we could achieve a smaller and closed set of relations. The drawback of this approach, however, is that we would either have to accept a substantial loss of information or a hierarchy of relations, as discussed in more detail below in section 4.3.1.2.

A reification of verb relations, however, would allow us to work with a closed set of relations with a limited number of members, more or less equivalent to the set of thematic roles (cf. e.g. Fillmore).

4.3.1.2 Multifaceted Conceptual Content

The conceptual content of a given verb²² is multifaceted, and often closely related to that of other verbs, as described in e.g. (Levin, 1993). That the conceptual content of a given verb is multifaceted means that it cannot be abstracted to a general relation without significant loss of information.

²¹ Source: Adam Kilgarriff's BNC frequency lists (<http://www.kilgarriff.co.uk/bnc-readme.html>)

²² The examples given in this account apply to English or Danish. The semantics of verbs may be more or less explicitly expressed in other languages.

According to Levin's classification of English verbs (Levin, 1993), the verb *build* belongs to the class 'verbs of creation and transformation', and so do the verbs *cook*, *knit*, *cut*, *hatch* etc. Thus, we could choose to abstract all verbs belonging to this class to the same relation, e.g. labeled 'create and transform'. However, if we look at definitions (here, from WordNet) of some of the verbs in the class, we observe information about different aspects of the verb meanings which would be lost in such an abstraction:

Cook: prepare a hot meal.

Knit: make (textiles) by knitting.

Cut: separate with or as if with an instrument.

Hatch: emerge from the eggs.

Build: make by combining materials and parts.

Based solely on the wording of the brief definitions above, we can identify the following types of information that would be lost in abstraction to a common superrelation: For *cook*, the information concerning the nature of the result of the process, namely that it is a hot meal, would be lost. For *knit*, information about the specific process leading to the result, knitting, as well as the nature of the result, that it is a textile, would be lost. For *cut*, the information that the process is a separation process as well as information that an instrument is involved in the process would be lost. For *hatch*, the information that the nature of the process is an emergence the source of which is an egg, would be lost.

For *build*, information about the nature of the process, it is a combination process, as well as information that the objects involved in this combination process are materials and parts would be lost.

However, we may assert that the conceptual content of the verb *build* is related to that of the verbs *cook*, *knit*, *cut* and *hatch*, while it is not particularly close to that of the verbs *run* or *administer*, which belong to different classes altogether. In order to capture such relatedness, or lack thereof, if we did not reify, we would have to apply a hierarchy of relations. For an ontology, we may choose to provide definitions of concepts, e.g. through axioms, or state ontological affinities that specify admissible combinations of ontotypes forming compound concepts (cf. (Nilsson & Jensen, 2003)). Also, in a given ontology framework, we may specify

different features of a relation, either formal properties such as transitive²³, symmetric²⁴, non-symmetric²⁵, etc. or restrictions on the ontological types of the relates. Thus, it would be possible to assert (some of) the differentiating features between concepts as well as between sibling relations in a relation hierarchy. Consequently, the choice between reification and a relation hierarchy cannot be made on the basis of expressivity.

However, applying a hierarchy of relations would increase the size of the index significantly and complicate computation of semantic similarity (see below in 4.3.1.3).

The fact that the concepts that the relations may be reified into, obviously also form a hierarchical structure does not present a problem for the search efficiency as this is an inherent feature of an ontology.

4.3.1.3 Implications of Adding a Relation Hierarchy

For ontology-based search purposes, when we wish to be able to return close, but not necessarily exact, matches to a query, a concept must be expanded. This expansion may either target the index or be performed at query time.

For index expansion, a concept is expanded upwards, such that for any given concept to be indexed, all its superordinate concepts become index terms as well. For query expansion, the expansion is typically downwards, such that for any given concept to be indexed, all its subordinate concepts become index terms.

For the sake of this example, we will here assume an index expansion, without going into arguments for or against one approach. For various purposes, and with an application of different similarity measures, a concept may be expanded in different directions; upwards, downwards, sideways, or a combination of these, where some cut-off factor would be applied. The expansion is performed in order to facilitate that a search for a given concept returns not only exact matches, but also concepts that to some extent are similar to the search term.

²³ A relation is transitive if: C ISA B and B ISA A, then C ISA A (e.g. *is ancestor of*)

²⁴ A relation is symmetric if: A ISA B, then B ISA A (e.g. *is sibling of*)

²⁵ A relation is non-symmetric if: A ISA B, then NOT B ISA A (e.g. *is parent of*)

In the index expansion approach exemplified here, we pose an expansion of a given leaf concept as far upwards as possible, i.e. to the top node of the ontology, which will facilitate that a search for a given concept includes a search for all the concepts that this concept subsumes, e.g. a search for ANIMATE would include PERSON and PETER, cf.

Figure 26. In such a concept expansion, any given concept is abstracted, or generalized, to the concept that it is subsumed by. We may then use these concept abstractions as indexing terms.

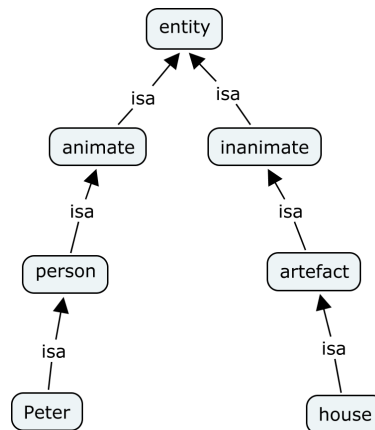


Figure 26 Ontology fragment containing the concepts peter and house

For compound concepts, abstraction includes decomposition. For a compound concept where any of the relates are themselves compound concepts, these will also be decomposed.

For any compound concept c' with the core concept c , where the direct attributions make up a set S , the set of concepts S' in the decomposition is equal to the power set of S attributed to c . Thus, the set grows exponentially for any attribution added to c . This is furthermore true for any compound concept in S .

(23)

$c' = \text{BUILDING}[\text{AGT:PETER}, \text{PRD:HOUSE}]$

$c = \text{BUILDING}$

$S = \{ \text{AGT:PETER}, \text{PRD:HOUSE} \}$

$P(S) = \{ \{ \text{AGT:PETER, PRD:HOUSE} \}, \{ \text{AGT:PETER} \}, \{ \text{PRD:HOUSE} \}, \{ \} \}$
 $S' = \{ \text{BUILDING}[\text{AGT:PETER, PRD:HOUSE}], \text{BUILDING}[\text{AGT:PETER}],$
 $\text{BUILDING}[\text{PRD:HOUSE}], \text{BUILDING} \}$

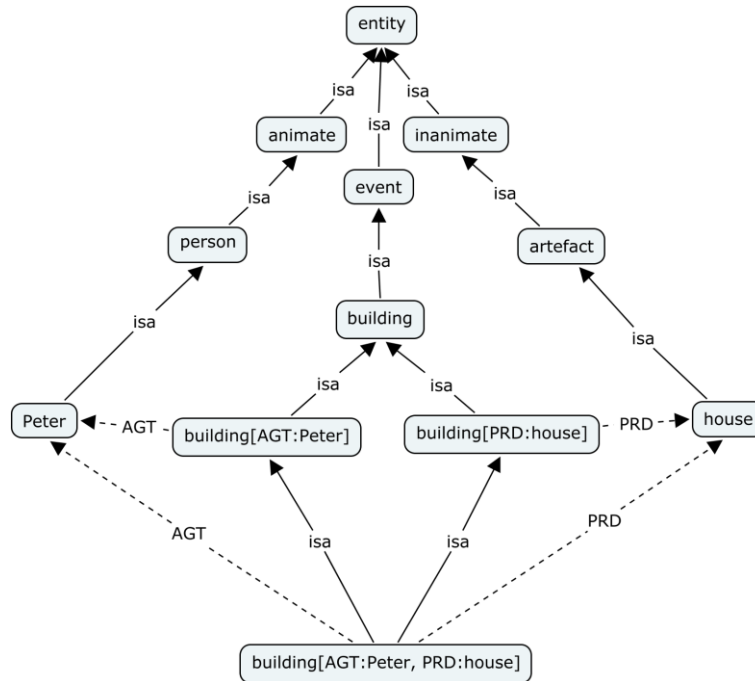


Figure 27 Decomposition of the compound concept BUILDING[AGT:PETER,PRD:HOUSE]

As exemplified in (23), given the compound concept ‘BUILDING[AGT:PETER, PRD:HOUSE]’, we decompose and abstract it to ‘BUILDING[AGT:PETER]’ and ‘BUILDING[PRD:HOUSE]’, both of which are abstracted to BUILDING, as illustrated graphically in Figure 27.

For a compound concept with only two attributions and atomic concepts as relates, such as for the example shown in (23) and Figure 27, this decomposition process results in 2^2 members of S' , and does thus not result in a significant increase in the number of indexing terms. However, if we add just one direct attribution, as in (24), we get 2^3 members of S' .

(24)

$c' = \text{BUILDING}[\text{AGT:PETER, PRD:HOUSE, MAT:WOOD}]$

$c = \text{BUILDING}$
 $S = \{ \text{AGT:PETER, PRD:HOUSE, MAT:WOOD} \}$
 $P(S) = \{ \{ \text{AGT:PETER, PRD:HOUSE, MAT:WOOD} \}, \{ \text{AGT:PETER, PRD:HOUSE} \}, \{ \text{AGT:PETER, MAT:WOOD} \}, \{ \text{PRD:HOUSE, MAT:WOOD} \}, \{ \text{AGT:PETER} \}, \{ \text{PRD:HOUSE} \}, \{ \text{MAT:WOOD} \}, \{ \} \}$
 $S' = \{ \text{BUILDING[AGT:PETER, PRD:HOUSE, MAT:WOOD]}, \text{BUILDING[AGT:PETER, PRD:HOUSE]}, \text{BUILDING[AGT:PETER, MAT:WOOD]}, \text{BUILDING[PRD:HOUSE, MAT:WOOD]}, \text{BUILDING[AGT:PETER]}, \text{BUILDING[PRD:HOUSE]}, \text{BUILDING[MAT:WOOD]}, \text{BUILDING} \}$

If just one of the relates is a compound concept, then additionally, S' grows with the possible decompositions of this compound concept attributed c .

Further, we can infer that a given relation combining two given concepts and thus forming a compound concept, exists at all levels of abstraction. Thus, to produce the set of abstractions for a compound concept, we compute the cartesian product of the sets of abstractions for each relate.

This would, given the skeleton ontology in Figure 26, for the compound concept $\text{PETER[BUILD:HOUSE]}$, where the sets of abstractions of the relates are $\{ \text{PETER, PERSON, ANIMATE, ENTITY} \}$ and $\{ \text{HOUSE, ARTEFACT, INANIMATE, ENTITY} \}$, give $4*4$ abstractions, namely: $\{ \text{PETER[BUILD:HOUSE]}, \text{PETER[BUILD:ARTEFACT]}, \text{PETER[BUILD:INANIMATE]}, \text{PETER[BUILD:ENTITY]}, \text{PERSON[BUILD:HOUSE]}, \text{PERSON[BUILD:ARTEFACT]}, \text{PERSON[BUILD:INANIMATE]}, \text{PERSON[BUILD:ENTITY]}, \text{ANIMATE[BUILD:HOUSE]}, \text{ANIMATE[BUILD:ARTEFACT]}, \text{ANIMATE[BUILD:INANIMATE]}, \text{ANIMATE[BUILD:ENTITY]}, \text{ENTITY[BUILD:HOUSE]}, \text{ENTITY[BUILD:ARTEFACT]}, \text{ENTITY[BUILD:INANIMATE]}, \text{ENTITY[BUILD:ENTITY]} \}$.

If we were to introduce a hierarchy of relations, which implications would that have on the size of this set of abstractions?

For the sake of this example, we propose a hierarchy of relations based on Levin's classification of English verbs (Levin, 1993). Levin asserts 57 general verb classes, among these 'verbs of creation and transformation' which has 7 subclasses. The verb *build* belongs to the subclass 'build' along with the verbs *cook*, *knit*, *cut*, *hatch*, etc. A fragment of the hierarchy of relations, restricted to verbs in this class only, is shown in Figure 28.

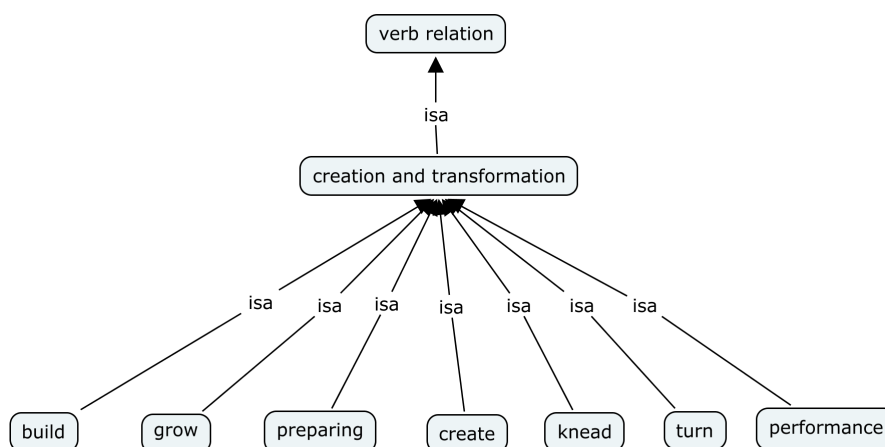


Figure 28 A fragment of a relation hierarchy based on Levin’s verb classes

The implications of adding such a relation hierarchy would be that we would have to not only abstract concepts, but also relations. For the minimal compound concept PETER[BUILD:HOUSE], we would thus have to produce the Cartesian product of three sets, namely both the sets of abstractions of the relates and of the relations. This would, given the skeleton ontology in Figure 26 and the relation hierarchy in Figure 28, for the compound concept PETER[BUILD:HOUSE], where the sets of abstractions of the relates and relation are, {PETER, PERSON, ANIMATE, ENTITY}, {HOUSE, ARTEFACT, INANIMATE, ENTITY} and {BUILD, CREATION_AND_TRANSFORMATION, VERB_RELATION}, give $4 \times 4 \times 3$ abstractions as indexing terms, namely: {PETER[BUILD:HOUSE], PETER[CREATION_AND_TRANSFORMATION:HOUSE], PETER[VERB_RELATION:HOUSE], PERSON[BUILD:HOUSE],...}.

Note that for any compound concept, its possible decompositions are included in the set of abstractions.

The addition of a hierarchy of relations thus dramatically increases the size of the set of abstractions for a compound concept. This makes the indexing process considerably more time consuming, it requires more space for the index and, in consequence, makes search slower.

4.3.1.4 Only Binary Relations

We restrict ourselves to only representing binary relations between concepts. The feature:value format of the ONTOLOG attributions restrict us to represent binary relations only.

Thus, if we were to represent *build* as a relation BUILD instead of a as the concept BUILDING, observing this restriction, we would not be able to represent the conceptual content of example (25), which expresses a ternary relation involving the ‘AGENT’, the ‘PRODUCT’ and the ‘MATERIAL’ roles of the ‘build relation’ (according to the information in VerbNet presented above in section 4.3.1).

(25) *Peter built a house from wood*

ONTOLOG allows us, in principle, to represent an infinite number of attributions in the form of attribute:value pairs. Thus, in the reification approach, we have no trouble representing the conceptual content of (25), as the roles become attributions to the concept BUILDING:

(26) BUILDING[AGT:PETER, PRD:HOUSE, MAT:WOOD].

In this way, the arity of the underlying verb relation becomes unimportant.

4.3.1.5 Modeling the Conceptual Content of Sentences.

We aim at constructing compound concepts reflecting the conceptual content not just of individual words but rather of text chunks. Ideally, we represent the conceptual content of sentences as compound concepts, but if we are not able to provide an analysis that results in a representation of the conceptual content of an entire sentence, we represent the content of the largest chunks that we are able to analyze.

Since verbs are commonly viewed as the heads of sentences, it makes perfect sense to view the conceptual content of the verb as the core concept of the compound concept denoted by a sentence. This approach also provides an adequate position for the compound concept in the ontology.

Reviewing example (21):

(21) *Peter built a house*

If we chose to represent 'build' as a relation, we could choose to represent this as two concepts, namely 'Peter' and 'house', related via the 'build relation'. A graphical representation of the position of the concepts in the ontology is shown in Figure 29.

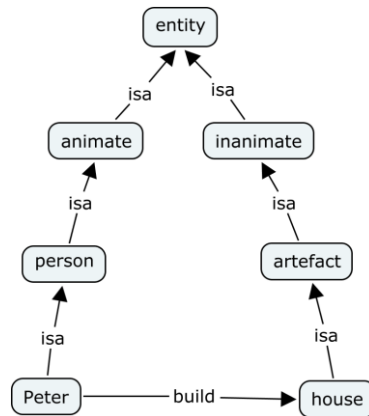


Figure 29 The atomic concepts Peter and house related via the 'build relation'

However, what we want is *one* compound concept reflecting the conceptual content of the whole sentence, that we can use as an indexing term. Thus, another solution would be to construct a compound concept while still representing 'build' as a relation, namely the concept PETER[BUILD:HOUSE], as illustrated in Figure 30.

While, at a first glance, this may look like an applicable solution, it is not. The compound concept PETER[BUILD:HOUSE] is subsumed by the concept PETER, which means that the compound concept is a kind of PETER, which in turn means that the compound concept would have the interpretation 'a kind of Peter that builds a house'. This is not quite the preferred interpretation of the text in example (21).

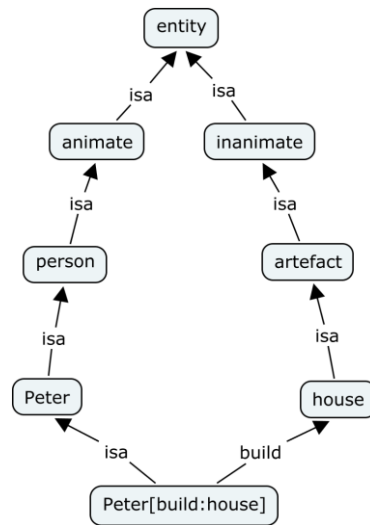


Figure 30 The compound concept PETER[BUILD:HOUSE]

The final solution includes the reified relation expressed by the verb as the core concept of the compound concept. This solution is illustrated in Figure 31.

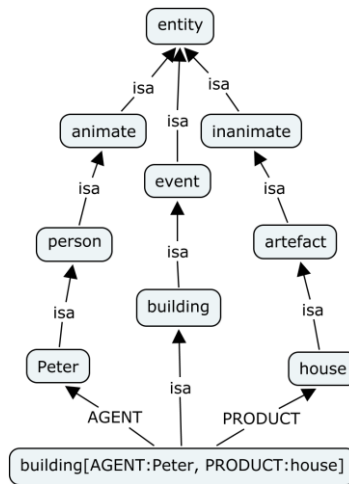


Figure 31 The compound concept BUILDING[AGT:PETER, PRD:HOUSE]

In Figure 31, the compound concept BUILDING[AGT:PETER, PRD:HOUSE] is subsumed by the concept building (which is subsumed by the concept

EVENT). What this means is that there exists a concept BUILDING which is a kind of EVENT. As the compound concept is subsumed by BUILDING, it would have the interpretation ‘a kind of building event that has an AGENT that is Peter and a PRODUCT that is a house’. This is the interpretation we are after.

4.3.1.6 Congruous Meaning of Nouns and Verbs

Some verb-noun pairs have identical or overlapping meaning. This is notably true for pairs of verbs and their nominalized forms, or vice versa, nouns and their verbalized forms. Examples of this phenomenon are the verb *build* and the corresponding deverbal noun *building* as exemplified in (21) and (27) below. Other verb/nominalization pairs are *act/action*, *treat/treatment*, etc., and examples of noun/verbalization pairs are *saddle/saddle*, *house/house*, etc.

(21) *Peter built a house*

(27) *The building of a house by Peter*

It is our claim that the conceptual content of the two texts in (21) and (27) is identical and thus should be mapped to the same node in an ontology, as described in more detail in (Andreasen, Bulskov, Jensen et al., 2009). For this to be achieved, we have to somehow treat a given verb and its nominalized form or a given noun and its verbalized form in the same manner. There are two ways of going about this: either we can revert the deverbal noun to its verb root and treat both words as relation denoting, or we can reify the relation denoted by the verb, and treat both words as concept denoting. Given the arguments for the reification approach in the present account, we choose the latter.

4.3.2 Prepositions

As described in chapter 2, prepositions resemble verbs syntactically (they take complements) as well as semantically (they denote relations). However, we do not treat the two word classes in the same manner. In our framework, relations denoted by prepositions are represented as associative relations in the ontology.

For prepositions, since they denote binary relations, we need neither to reify relations, nor introduce relations. However, we abstract the relations

denoted by prepositions to more general relations, such as LOCATIVE, TEMPORAL, CAUSATIVE, etc.

There are several reasons for this approach, as discussed below in sections 4.3.2.1 to 4.3.2.4.

4.3.2.1 Finite Set of Relations Denoted by Prepositions

It is possible to define a finite set of relations that prepositions denote, as will be shown in chapter 6 below. The class of prepositions is a closed class; this means that new prepositions are never, or very rarely, added to the class and, further, their semantics is stable. Thus, the set of relations denoted by prepositions is a finite set.

Further, a reification of relations denoted by preposition would result in a larger set of relations – the contrary of our aim. An example showing this phenomenon is given below.

(28) *The vase on the table*

If we reified the relation denoted by ‘on’ in example (28), which is a locative relation that holds between the concepts VASE and TABLE, we would have to add a location concept to our ontology. The compound concept denoted by example (28) would be subsumed by this ‘location concept’ and be related to the two concepts VASE and TABLE, here labeled ‘LOCATEE’ and ‘LOCATION’ relation respectively, as shown in Figure 32. Thus, instead of one relation, namely the LOCATION relation, we would have two. This would be the effect on all relations denoted by prepositions in our relation set.

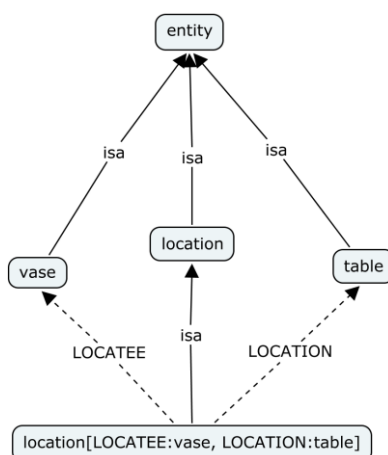


Figure 32 A reified LOCATION relation

4.3.2.2 Relations Denoted by Prepositions are Binary

In the third part our definition of the class of prepositions rendered in section 2.3, we asserted that *prepositions are pure relators that denote binary relations*. Thus, there is no need for a reification of prepositional relations on the grounds of their arity being variable, as is the case with verbs: Relations denoted by prepositions are binary. At least, that is the assumption we work under in this account. The preposition *between* (or *mellem* in Danish) is a special case, since it does not necessarily appear to denote a binary relation. The case of this preposition is described in more detail in section 4.4.3.

4.3.2.3 The Conceptual Content of a Preposition is not Multifaceted

In most cases, the conceptual content of a preposition may be abstracted to a general relation without the loss of much information, if any.

(25) *Peter built a house from wood*

Reviewing example (25) above, the preposition *from* may be abstracted to a MATERIAL relation, and not much else can be said about the semantics of this lexeme in the given context. Thus, the conceptual content of (25) is fully captured by a translation into the ONTOLOG expression BUILDING[AGT:PETER, PRD:HOUSE, MAT:WOOD].

Similarly, the prepositions *on*, *above* and *below* may all be abstracted to the LOCATION relation for examples (29), (30) and (31) below, yielding the ONTOLOG expression VASE[LOC:TABLE].

(29) *The vase on the table*

(30) *The vase above the table*

(31) *The vase below the table*

However, not everything is said about the semantics of the lexemes *on*, *above* and *below* by abstracting to a LOCATION relation. The difference between the position of the vase in examples (29), (30) and (31) is not fully captured by a common translation into the ONTOLOG expression VASE[LOC:TABLE]:

In (29), the position of the vase is conceived as being somewhere in a space above the table and contiguous with the upper side of the tabletop, as illustrated in Figure 33.

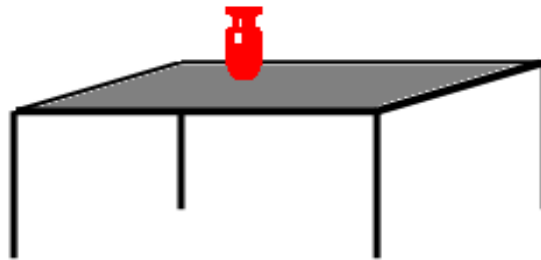


Figure 33 Gray area indicates possible position of the vase relative to the table given by *on*

In (30), the position of the vase is also conceived as being somewhere in a space above the table but *not* contiguous with the tabletop, as illustrated in Figure 34.

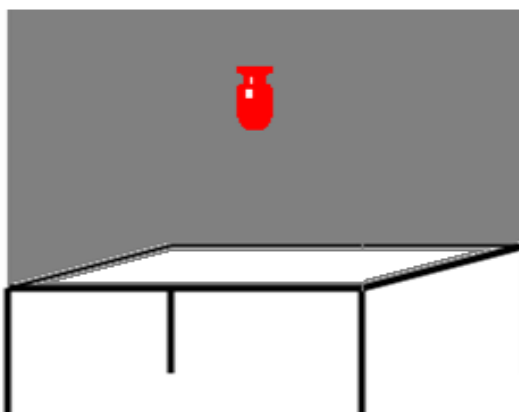


Figure 34 Gray area indicates possible position of the vase relative to the table given by *above*

And in (31), the position of the vase is conceived as being somewhere in a space below and probably not contiguous with the tabletop, as illustrated in Figure 35.

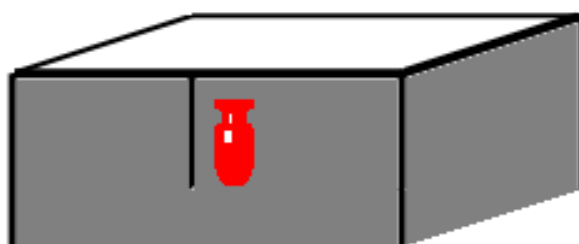


Figure 35 Gray area indicates possible position of the vase relative to the table given by *below*

Consequently, for prepositions represented as the locative relation, some loss of information exists. This is equally true for prepositions represented as the temporal relation.

Thus, while it is possible to come up with a finite set of general relations that prepositions denote, some relations may be broken further down into more fine-grained relations. Locative relations may for example be broken down with respect to features such as static/dynamic, penetrating/non-penetrating, contiguous/detached or defined in a vector space (cf. (Zwarts, 1997)), and temporal relations may for example be broken down with respect to features such as point/interval or event/state. However, such a fine

grained relation set is not considered in our approach as we do not need it and it complicates search unnecessarily.

4.3.2.4 Compound Concepts Reflecting the Conceptual Content of Sentences

As noted above, we aim at constructing compound concepts reflecting the conceptual content of sentences, not just individual words or phrases. With this aim in view, it does not make sense to reify relations denoted by prepositions.

Prepositions and their complements (in combination forming prepositional phrases) modify other parts of sentences; in our analyses, they may modify heads of noun phrases and thus be part of a noun phrase, or they may modify heads of verb phrases.

(32) *Peter broke the vase on the table*

In example (32), the prepositional phrase *on the table* is ambiguous as regards attachment. It may modify the noun *vase* or it may modify the verb *broke*. The different situations that the two attachment possibilities denote are illustrated in Figure 36 and Figure 37.

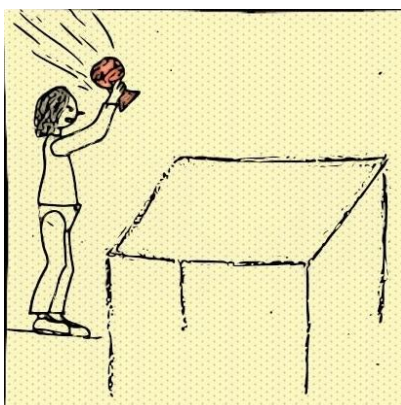


Figure 36 The situation described by (32) with *on the table* modifying the verb *broke*

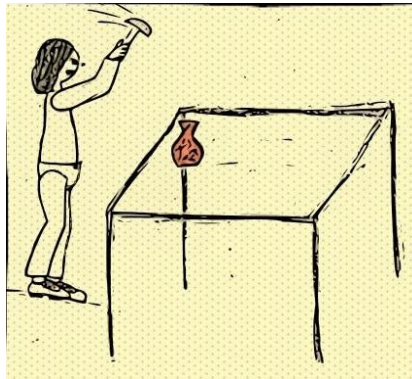


Figure 37 The situation described by (32) with *on the table* modifying the noun *vase*

These different readings are manifested at the surface level as well as at the conceptual level. The surface structures of the two analyses of example (32) are shown in Figure 38 (corresponding to the situation depicted in Figure 36) and Figure 39 (corresponding to the situation depicted in Figure 37), and various alternative conceptual representations are presented in (34) to (36).

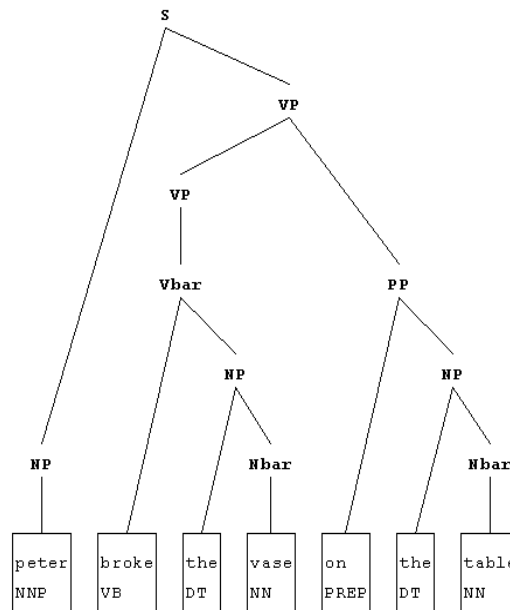


Figure 38 The PP *on the table* modifying the verb *broke*

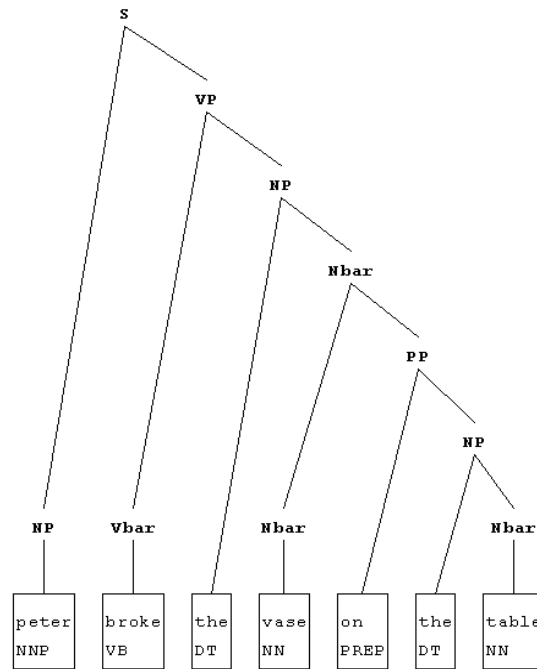


Figure 39 The PP *on the table* modifying the noun *vase*

For the two syntactic analyses illustrated in Figure 38 and Figure 39, we may produce the following corresponding ONTOLOG expressions:

- (33) BREAKING[AGT:PETER, PNT:VASE, LOC:TABLE]
 (34) BREAKING[AGT:PETER, PNT:VASE[LOC:TABLE]]

If we were to reify the LOCATION relation, constructing ONTOLOG expressions reflecting the conceptual content of the two readings of (32) is not a straightforward task. For the two readings, we could construct the compound location concepts ‘LOCATION[LOCATION:TABLE]’ (cf. Figure 38) and ‘LOCATION[LOCATEE:VASE, LOCATION:TABLE]’ (cf. Figure 39), respectively. An attempt at including these concepts in the representations of (32) could be:

- (35) BREAKING[AGT:PETER, CHR:LOCATION[LOCATEE:VASE, LOCATION:TABLE]]

- (36) BREAKING[AGT:PETER,
PNT:VASE,CHR:LOCATION[LOCATION:TABLE]]

Here, we relate the location via a characterization relation instead of relating the arguments directly via a location relation. This approach makes the representations more complex, makes use of a larger set of relations, and more importantly, it does not provide a better or clearer description. Thus, it is not an applicable solution.

4.4 Pluralities as Arguments

The following discusses the modeling problem that arises when arguments of verbs or prepositions denote pluralities. Such pluralities may be denoted by collective or mass nouns, coordinated NPs or by plural forms. A special case is arguments of verbs or prepositions that require either a coordinated or plural form argument (e.g. *intersect*, *between*).

In order to model such arguments in an ontology, we propose two alternative solutions: The first solution is to map them to a concept that is subsumed by a plurality-concept. If the concept denoted by the noun or NP is modeled as a compound concept with relations to the entities that make up the plurality, this relation is a MEMBER relation as shown in Figure 40.

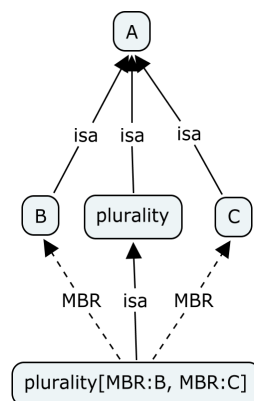


Figure 40 A plurality composed of the concepts A and B

The second solution proposed here is to only map inherently plurality-denoting nouns to a plurality-concept, to map plurals to the concept they denote and for coordinated NPs, list the attribute:value pairs where the relevant relation in the given context is repeated for each argument. The

relevant relation is for prepositional complements denoted by the preposition, and for arguments of verbs stem from the thematic role of the argument.

4.4.1 Concepts Denoted by Collective and Mass Nouns

Certain nouns inherently denote pluralities in some form; these include mass nouns (or uncountable nouns) and collective nouns. Examples of mass nouns are *cutlery* and *furniture*, and examples of collective nouns are *faculty*, *school* (e.g. *of fish*), *murder* (*of crows*) and *army*. In the ontology, the concepts that correspond to these nouns would all be subsumed by a group-denoting concept, in this account labeled PLURALITY. Thus, the graphical representation of the conceptual content of example (37) could be as illustrated in Figure 41.

(37) *Peter saw a school of herring*

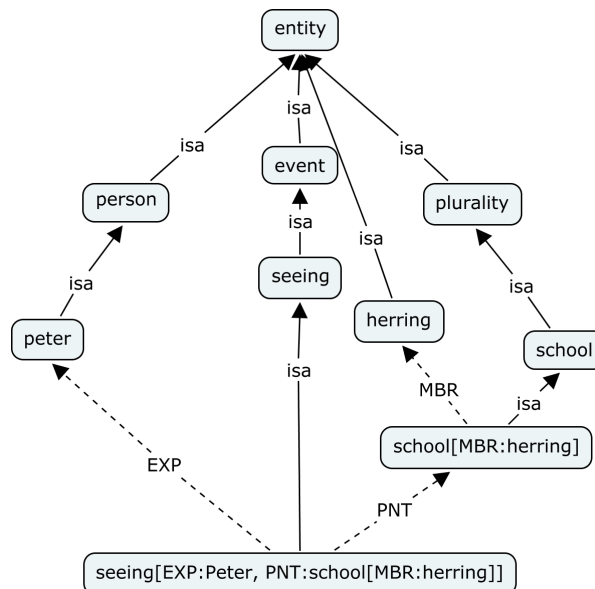


Figure 41 Mapping of the text *Peter saw a school of herring* into an ontology

However, the representation as illustrated in Figure 41 highlights a different problem: The group-denoting concept SCHOOL[MBR:HERRING] only has one

individual member, namely one individual of the concept HERRING, and a reasonable postulation would be that a group must have more than one member. We thus have to introduce an ontological assumption that any concept that is subsumed by PLURALITY has more than one member. If just one concept is in a member relation to this concept, the assumption would be satisfied if the related concept itself is subsumed by PLURALITY (e.g. a concept denoted by a mass noun), but otherwise it is inferred that more than one entity of the same type are related.

4.4.2 Concepts Denoted by Coordinated Phrases

For coordinated NPs that denote a group, it is also possible to apply a plurality-reading

(38) *Peter and Mary wrote a book*

For example (38), there are two possible readings: Either Peter and Mary wrote a book together, or they each wrote a book. A representation of the conceptual content of the collective reading could include a compound concept subsumed by the concept PLURALITY with member relations to the concepts PETER and MARY, as illustrated in Figure 42.

This representation would differentiate the conceptual content of the two readings, as the reading where Peter and Mary each wrote a book would be represented as illustrated in Figure 43 ('WRITING[AGT:PETER, RST:BOOK]' and 'WRITING[AGT:MARY, RST:BOOK]' possibly subsuming a compound concept 'WRITING[AGT:PETER, AGT:MARY, RST:BOOK]' representing the conceptual content of the whole text)

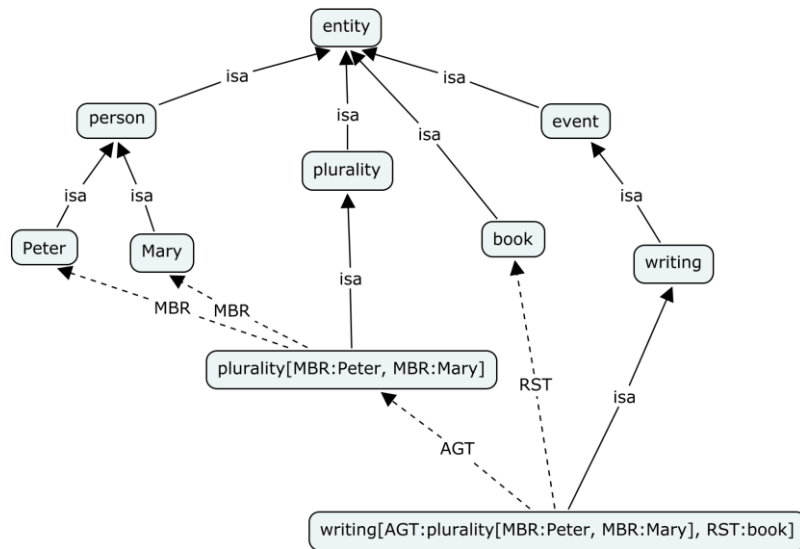


Figure 42 Plurality-reading of *Peter and Mary wrote a book*

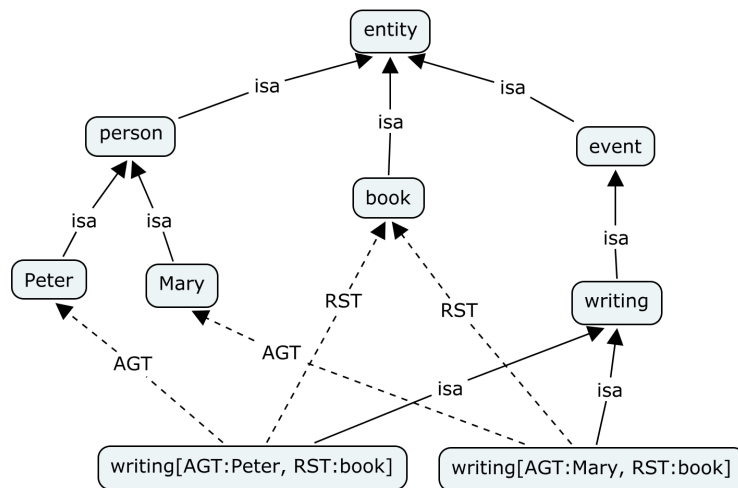


Figure 43 Individual-reading of *Peter and Mary wrote a book*

Thus, if we choose to model the conceptual content of coordinated NPs with a collective reading as a repetition of the relevant relation, we would get a representation that is identical to that illustrated in Figure 43.

4.4.3 Concepts Denoted by Preposition Complements

Relations denoted by prepositions are binary. At least, that is the assumption we work under in this account. The preposition *between* (or *mellem* in Danish) is a special case, since it does not necessarily appear to denote a binary relation, as illustrated by examples (39) and (40), where it could be analysed either as having multiple complements, or, as we choose to analyze it, to have one complement realized as a coordinated NP. In examples (41) and (42), the relation is undeniably binary as the preposition has only one complement. However in these cases, the second argument must be in the plural form or, at least for the Danish preposition *mellem*, itself denote a plurality as in (42). Thus, our claim is that the preposition *between* differs from most other prepositions in that it requires that its complement denotes a plurality.

- (39) *The house between the church and the pub*
(40) *The contract between the contractor, the state and the region*
(41) *The cream between the cookies*
(42) *Krummer mellem bestikket*
 Crumbs between the cutlery

For example (39), a plurality representation of the second argument of the relation could yield ‘HOUSE[LOC:PLURALITY[MBR:CHURCH,MBR:PUB]]’, as shown as a graphical representation in Figure 44.

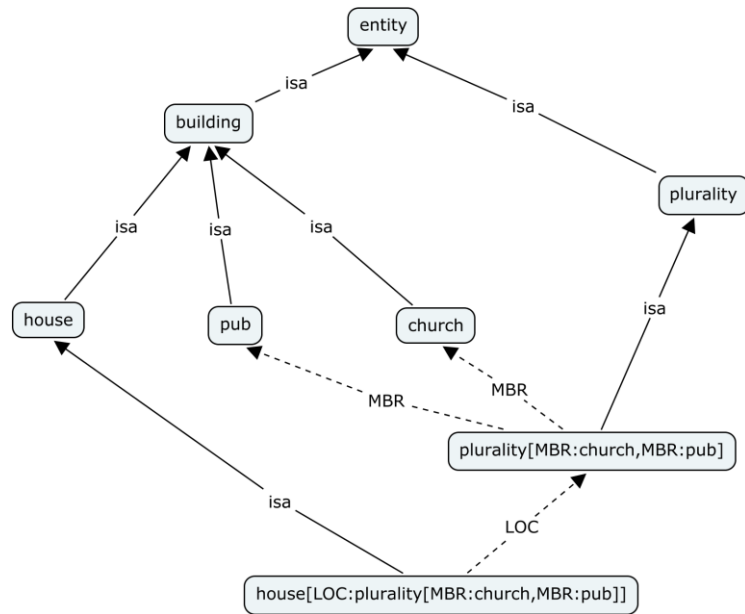


Figure 44 Plurality as a relatum

Another way of representing the conceptual content of (39) is, for the relation that the preposition denotes in the given context, to list the attribute:value pairs where the given relation as the attribute is repeated for each concept denoted by the individual NPs. For (39), this would yield: ‘HOUSE[LOC:CHURCH, LOC:PUB]’ as shown in a graphical representation in Figure 45.

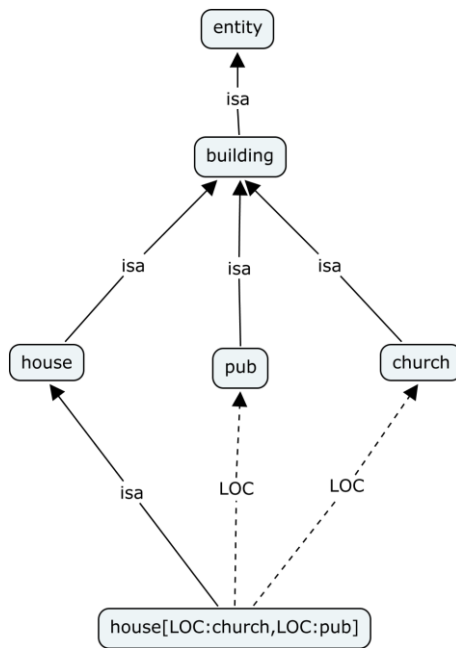


Figure 45 A plurality of relata

4.4.4 Concepts Denoted by Verb Arguments

Certain verbs require either a coordinated first argument as in (43), or an argument in the plural as in (44). Among these are *intersect*, *meet* and *be in love*.

(43) *38th Street and 35th Street intersect.*

(44) *The streets intersect*

We propose a treatment of these arguments similar to that of the prepositions *between* and *mellem*, yielding the two alternative representations in Figure 46 and Figure 47.

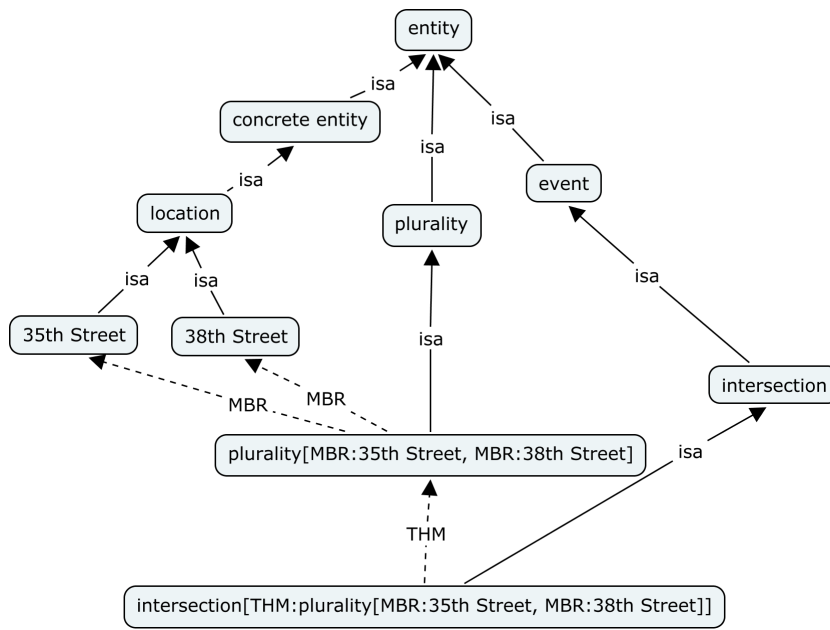


Figure 46 Plurality as a relatum for *intersection*

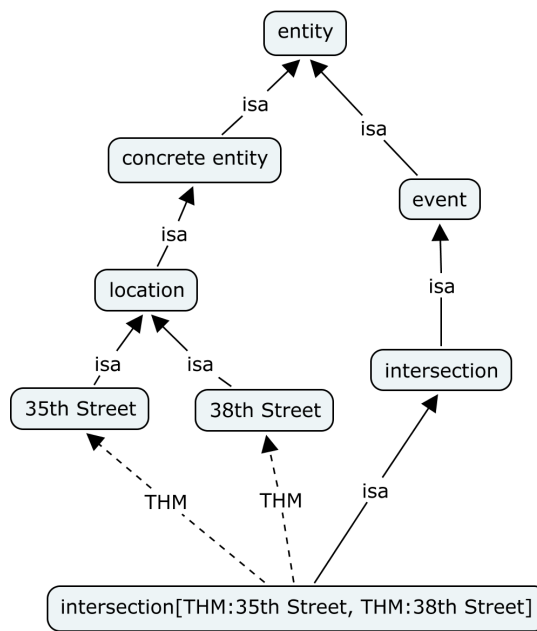


Figure 47 A plurality of relata for *intersection*

The choice of whether to represent collective readings of coordinated NPs as pluralities depends on the objective: is the purpose of the modeling to produce a representation that captures as much of the semantics of a given text as possible, or is it to represent adequate information in order to improve search? If the purpose were solely a question of capturing semantics, we would probably choose the plurality approach for all the examples presented above. However, as the aim of this project is to improve search through ontology-based indexing, it may not be a sound solution. Imagine a search for *books written by Peter*. The term by which the text in (38) '*Peter and Mary wrote a book*' is indexed with the plurality approach applied is 'WRITING[AGT:PLURALITY[MBR:PETER, MBR:MARY] RST:BOOK]', which does not directly put PETER in an AGENT relation to BOOK. As a result, depending on the similarity measure applied, books written by Peter alone would be ranked higher. An indexing term without applying the plurality approach would be 'WRITING[AGT:PETER, AGT:MARY, RST:BOOK]', which does put PETER in a direct AGENT relation to BOOK, resulting in an equal ranking of all books written by Peter without regard to possible co-writers. We would most likely prefer such a search to return all mentions of books, including ones of books written by Peter in collaboration with other people. Further, we would probably want them to be ranked equally high as the ones he may have written alone. For this reason, we choose to always represent coordinated NPs as individuals in a direct relation to the concept in question.

For complements of prepositions, neither of the proposed solutions manage to fully capture the semantics of *between*. For the individual representation of (39) '*The house between the church and the pub*', 'HOUSE[LOC:CHURCH, LOC:PUB]', an interpretation would be 'a house that is located in relation to a church and to a pub'. While this is not incorrect, there is no representation of the fact that the location is somewhere in a range **from** the church **to** the pub.

Similarly, the expression 'HOUSE[LOC:PLURALITY[MBR:CHURCH,MBR:PUB]]' does not indicate the location being in such a range. Since for most prepositions, we do lose some information in the abstraction of the relation to the limited set of relations that we use in our framework, this does not present a major problem. We thus choose the simplest solution, namely the individual representation.

For arguments of verbs such as *intersect*, *meet* and *be in love*, we choose the same solution, that is, to model arguments with the individual-representation. We choose this solution for the same reasons as presented above for complements of prepositions: Neither representation model fully captures the semantics of *intersect* and *meet*, namely that for two or more entities, there is one point in which their location overlaps. For e.g. *be in love*, the reciprocity of the relation is neither captured in the plurality approach or the individual approach.

Since neither of the proposed representation forms satisfy us with respect to capturing the full semantics of texts, we could choose a more elaborate representation language, or add expressive power e.g. in the form of axioms to the chosen language in order to remedy the shortcomings. However, for our search-oriented purpose, the ONTOLOG representation is sufficiently expressive, and we do not wish to add complexity to the language. In reality, we are satisfied that ONTOLOG representations are abstractions of the conceptual content of texts rather than complex expressions capturing all details of the semantics.

4.5 Summary

Above, we have discussed the notions of a concept, a relation and a linguistic expression as they are used in this dissertation.

We have asserted that concepts exist in the minds of people, and are abstract ideas of entities in the world, and a relation is the conceptual glue that binds concepts together in discourse. Linguistic expressions are sequences of characters, or linguistic forms that may or may not denote concepts or relations, i.e. they may or may not be part of signs. Linguistic expressions may furthermore be continuous or discontinuous strings. Signs are combinations of linguistic form and meaning, and in the context of this account, we differentiate between conceptual and relational signs. The signified level of a conceptual sign is a concept, and the signified level of the relational sign is a semantic relation.

The notion of a sign as a combination of the conceptual level and the expression level is crucial for our treatment of text and mapping into a generative ontology. We aim at constructing compound concepts reflecting the conceptual content not just of individual words but rather of text chunks. Ideally, we represent the conceptual content of sentences as compound

concepts, but if we are not able to provide an analysis that results in a representation of the conceptual content of an entire sentence, we represent the content of the largest chunks that we are able to analyze.

We have described different approaches to the treatment of relation denoting words, especially verbs and prepositions. We have asserted that we wish to keep the set of relations small and closed. This goal is problematic to achieve if we represent the conceptual content of verbs as relations, because the set of verbs in languages such as Danish and English is large and open. In order to achieve the goal, we reify relations denoted by verbs. Also, this approach makes the fact that verbs denote n-ary relations unimportant. We have argued that the reification approach is not viable for our treatment of prepositions, and as a result, we choose to represent the conceptual content of prepositions as associative relations.

Finally, we have provided an approach to the treatment of arguments where the conceptual content is a plurality. Such arguments may be realized as collective and mass nouns, coordinated phrases, certain prepositional complements and verbal arguments. Neither of the proposed representation forms fully capture the conceptual content of such textual forms, however, rather than choosing to use a more elaborate representation language, we choose to live with the limited expressive power of the ONTOLOG language. In reality, we are satisfied that ONTOLOG representations are abstractions of the conceptual content of texts rather than complex expressions capturing all details of the semantics.

Chapter 5

A Machine Learning Approach to Disambiguation of Semantic Relations

This chapter describes experiments in using machine learning for disambiguation of semantic relations denoted by prepositions.

The chapter reflects a body of work that has been carried out in collaboration with Thomas V. Terney (TVT). The bulk part of the results have previously been published in (Lassen, 2006; Lassen & Terney, 2006a, 2006b), as well as in (Terney, 2009). The work was carried out within the framework of the OntoQuery project²⁶.

The chapter is organized as follows: Section 5.2 situates relation disambiguation in relation to word sense disambiguation, and describes the aim of the experiments. Section 5.3 briefly describes a selection of available resources for semantic roles. Section 5.4 describes the task in further detail. Section 5.5 describes and exemplifies semantic relations and realization in linguistic expressions. Section 5.6 describes the corpus and the different levels of the annotation process, the sets of ontological types and semantic relations used in the corpus annotation. Section 5.7 describes how the experiments were carried out, the applied algorithms, and gives an analysis of some of the rules produced by the JRip algorithm. Section 5.8 summarizes.

²⁶ <http://www.ontoquery.dk>

5.1 Content-based Information Retrieval

What we in the following refer to as a content-based information retrieval system is to be understood as a framework in which texts are indexed and retrieved on the basis of their conceptual content instead of string occurrences. In our understanding of such a system, some ontology is required for this task.

In order to facilitate concept-based indexing and retrieval, we need to be able to perform a mapping from the expression level to the conceptual level. For most simplex words or multi-word expressions that denote atomic concepts in an ontology, this can be done simply by mapping from the expression level to an existing node in the ontology. However, this framework uses so-called generative ontologies, where compound concepts may be added when identified in a text. A compound concept can be represented as a conceptual feature structure of the general form CONCEPT[REL:CONCEPT]. Thus, we can refer to the task of identifying compound concepts in text as concept extraction.

5.2 Word Sense Disambiguation vs. Relation Disambiguation

The notion of *a word* is here to be understood as a linguistic expression that represents a unit of meaning, and is thus comparable to the notion of a linguistic sign (cf. section 4.1). A word may consist of a single unbound morpheme or of a combination of (unbound and bound) morphemes. In the context of automatic text processing, a word is typically defined as a continuous string. A given word may have several senses; these senses are to be understood as the possible meanings of the word as listed in a dictionary for a given lemma. For example, the dictionary entry for the lemma *hair* below²⁷ lists six senses:

Hair

–noun

1. any of the numerous fine, usually cylindrical, keratinous filaments growing from the skin of humans and animals; a pilus.
2. an aggregate of such filaments, as that covering the human head or forming the coat of most mammals.

²⁷ From: <http://dictionary.reference.com>

A Machine Learning Approach to Disambiguation of Semantic Relations

3. a similar fine, filamentous outgrowth from the body of insects, spiders, etc.
4. Botany. a filamentous outgrowth of the epidermis.
5. cloth made of hair from animals, as camel and alpaca.
6. a very small amount, degree, measure, magnitude, etc.; a fraction, as of time or space: He lost the race by a hair.

Word sense disambiguation (WSD) is concerned with associating a given word form with the appropriate sense in the given context of words, cf. e.g. (Ide & Véronis, 1998). Normally, the given word will have been assigned a part of speech by a tagger prior to the WSD, and the task is then to decide on a given sense from a set of possible senses for the word given the part of speech.

Thus, for examples (45) and (46) below, the challenge for the WSD-algorithm is, amongst others, to decide which of the senses 1-6 for the lemma hair to associate with the word form *hair*. In these cases, example (45) would be tagged with sense no. 1, and example (46) would be tagged with sense no. 2.

(45) *This results in enhanced ability of follicles to regenerate and grow hair.*

(46) *He had dark brown hair, blue eyes and a fair complexion.*

Word sense disambiguation in its traditional sense is important for any information retrieval system that does not rely solely on string occurrences, and in our approach, it is particularly important when we perform a mapping from the expression level to the conceptual level for individual words.

It is not a novel idea to use machine learning in connection with word sense disambiguation, and by itself, it is not a novel idea to include some kind of classification or abstraction of the concept that a given linguistic expression denotes in the learning task, cf. e.g. (Yarowsky, 1992). Other projects have used light-weight ontologies such as WordNet in this type of learning task, e.g. (Agirre & Martinez, 2001; Voorhees, 1993)

The challenge here is not identical to WSD; however it is a related problem. We attempt to learn compound concepts, and as part of this, we disambiguate the relation that holds between a core concept and a related

concept. We presuppose a traditional word sense disambiguation for context words, and perform a conceptual context-based relation disambiguation of prepositions. Thus, our disambiguation concerns more complex linguistic structures than just individual words, and the output of our analysis is a compound concept in the form of a conceptual feature structure: CONCEPT[REL:CONCEPT].

Our results indicate an unexploited opportunity for including prepositions and the relations they denote in content-based information retrieval. In our framework, a compound concept is, in principle, an unbounded recursive structure, however; as a result of the particular syntactic form of our input, the compound concepts that we identify in these experiments are restricted to the general form CONCEPT[REL:CONCEPT].

5.3 Semantic Role Information - Available Resources

Semantic role labeling is also closely related to the task of relation disambiguation. Where semantic roles apply to the arguments of a relator, semantic relations are the relations that exist between a relator and a relatum; the semantic relation that exists between a verb and an agent is an agent relation, the semantic relation that exists between a verb and a patient is a patient relation, etc. Thus, the notion of semantic roles and semantic role labeling is very relevant to our task.

Semantic roles are typically described as relating to verbs and their arguments. A large body of theoretical work has been published on this subject (e.g. (Dowty, 1991; Fillmore, 1968; Jackendoff, 1983, 1990)) , and several resources exist that describe the possible syntactic frames for English verbs and the roles associated with the corresponding arguments in these frames, notably VerbNet, FrameNet and PropBank.

In these resources, prepositions are included, but only as markers introducing arguments that may fill a given semantic role. For example, in the VerbNet frame *Cut-21.1*, a possible syntactic frame is ‘NP V NP PP.instrument’. For this frame, the preposition *with* is explicitly stated as a part of the syntactic pattern introducing the argument that fills the instrument role, but no general assertion is made about the semantics of this preposition.

However, a resource that provides specific information about possible semantic roles for arguments of prepositions also exists, namely The Preposition Project , TPP.

Below, we briefly describe the resources VerbNet, FrameNet, PropBank and TPP.

All of the resources described below have been used as a basis for training semantic role labelers. For an overview of techniques, see (Palmer, Gildea, & Xue, 2010). Also in a disambiguation task for prepositional senses, a combination of resources have been used. For experiments concerning disambiguation of prepositional senses, see eg. (Gildea & Jurafsky, 2000; Litkowski & Hargraves, 2007; O'Hara & Wiebe, 2009).

VerbNet²⁸

VerbNet (Kipper-Schuler, 2006) is a domain-independent, hierarchically formed verb lexicon for English with mappings to other lexical resources, e.g. WordNet and FrameNet. It is organized into verb classes (an extension of (Levin, 1993) verb classes with added subclasses), where each class contains a set of members (the set of verbs that are members of the class), a role set (the set of thematic roles that may be associated with verbs in the class), and a set of frame elements which specify the possible surface realizations of the argument structure, paired with a text example. Each argument in such a frame structure is associated with a thematic role, and in addition, selectional restrictions (such as +ANIMATE, +HUMAN, +CONCRETE) are supplied which constrain the semantic types of the arguments associated with a thematic role. A syntactic frame containing PP-arguments may also be constrained with respect to the allowed preposition heading this phrase. In addition, each frame element contains detailed semantic information expressed as an event structure.

Currently, the VerbNet database contains approx. 5,700 lexical units (senses) and 4,000 lemmas distributed into 471 classes. The thematic role set consists of 24 members.

FrameNet²⁹

FrameNet is a lexical resource for English, consisting of semantic frames, which builds on Frame Semantics (Fillmore, 1976), but with an extended

²⁸ <http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

²⁹ <http://framenet.icsi.berkeley.edu/>

A Machine Learning Approach to Disambiguation of Semantic Relations

role set compared to the original work by Fillmore, and supported by corpus evidence. Each frame in FrameNet is defined by a natural language definition, the set of lexical units belonging to the frame and a set of frame elements (semantic role labels). The lexical units belonging to a frame may belong to different word classes (verbs, nouns, adjectives). In addition, each frame provides up to 20 annotated example sentences containing the lexical units.

In a given frame, the frame elements are classified as being either core elements or non-core elements. Core elements are elements that are conceptually necessary for the frame (cf. complements or syntactically necessary arguments), and non-core elements which are not conceptually necessary but which provide additional information (cf. adjuncts). The frame elements are specific to the set of lexical units in a given frame (e.g. for the frame *apply_heat* which e.g. contains the lexical unit *bake*, the core roles are 'container', 'cook', 'food', 'heating_instrument' and 'temperature_setting'), and thus the overall number of frame elements used is very large. Lexical items are grouped together in a frame based solely on their semantic similarities, and not based on a combination of semantic and syntactic similarities as is the case in Levin. However, there is some overlap between verbs in Levin classes and FrameNet (cf. (Palmer et al., 2010)).

Currently, the FrameNet database contains approx. 11,000 senses of 6,000 lemmas, 900 frames and 150,000 annotated example sentences. The set of frame elements consists of 2500+ members.

PropBank³⁰

PropBank (Palmer, Gildea, & Kingsbury, 2005) differs from the two aforementioned resources, VerbNet and FrameNet, in that the goal of the PropBank project was not to build a lexical resource. PropBank is an annotated corpus intended for use in machine learning of semantic roles. It contains annotations of the Wall Street Journal part of Penn Treebank II, where the annotations consist of predicate-argument structures for verbs with semantic role labels attached to each argument. In PropBank, the semantic role labels are theory independent and of the form *Arg0*, *Arg1*, etc. However, the annotations are consistent across syntactic variations and, thus, a role labeled *Arg0* in connection with a given syntactic form denotes the same type of role as *Arg0* in another syntactic form for arguments to a

³⁰ <http://verbs.colorado.edu/~mpalmer/projects/ace.html>

given verb; e.g. ‘Joe[Arg0] ate the cake[Arg1]’ compared to ‘the cake[Arg1] was eaten by Joe[Arg0]’. The semantic roles are in general verb specific, but *Arg0* can generally be seen as a prototypical agent and *Arg1* can generally be seen as a prototypical patient, cf. (Dowty, 1991).

In addition to the treebank annotations, PropBank provides a lexicon of the verbs in the annotations consisting of, for each broad sense of a verb, the possible arguments and the corresponding specified roles with mappings to VerbNet classes and roles, as well as all possible syntactic realizations of the argument structure. For example, for the verb *bake*, the possible roles are specified as:

Arg0: baker (vnrole: 45.3-Agent, 26.3-Agent)

Arg1: creation (vnrole: 45.3-Patient, 26.3-Product)

Arg2: source

Arg3: benefactive (vnrole: 26.3-Beneficiary)

Currently, the PropBank lexical database contains approx. 6,000 senses of 5,000 lemmas and 150,000 annotated example sentences. The set of semantic roles consists of 16 generic argument types corresponding to 6,000+ verb specific roles.

The Preposition Project³¹

The preposition project (TPP) (Litkowski & Hargraves, 2005) includes a description of 334 prepositions distributed among 673 senses. TPP applies a broad definition of the class of prepositions and, as a consequence, a large part of the included prepositions are so-called phrasal prepositions; i.e. multi-word prepositions, as e.g.: *according to*, *because of*, *by courtesy of*, *by means of*, *depending on*, *give or take*, *in connection with*, *in front of*, *next door to*, *under cover of*, *with respect to*, *with the exception of*, etc.

To each prepositional sense, a variety of information details is attached, e.g. definition, example, semantic category, semantic role-type, other prepositions that may denote the same sense, information about the type of element that the preposition attaches to (attachment properties), information about the type of element that is the complement of the preposition (complement properties), information about any paragraph in (Quirk,

³¹ <http://www.cires.com/prepositions.html>

Greenbaum, Leech, & Svartvik, 1985) that provides a semantic description of the sense and information about FrameNet-elements that are identified during the annotation process.

5.4 The Task

In order to achieve extraction of compound concepts, we need some method for identifying semantic relations between concepts.

Our first experiments in this direction have been an analysis of syntactic structures in the form of NP-PREP-NP. Our aim is to justify that for such syntactic structures, ontological affinities exist between the ontological types of the NPs and the relation that the preposition denotes.

The task described in this chapter, which we refer to as relation disambiguation, is to some extent similar to the problems of word sense disambiguation and semantic role labeling, as described above. However, we believe our contribution lies in the fact that we attempt to learn compound concepts, and as part of this, we disambiguate the relation that holds between a core concept and a related concept.

The experiments described here have been carried out using a small annotated Danish language corpus consisting of sentences that contain prepositions surrounded by noun phrases (NPs). The corpus is a subset of the OntoQuery corpus, which is compiled of texts in the domain of nutrition, all stemming from the The Danish National Encyclopaedia.

For this corpus, we analyze all text chunks that have the form NP-PREP-NP, and annotate them with information about part of speech, lemma, phrase boundary, phrase head, associated concept for the NP heads and relation for the prepositions. The annotated text chunks are the input to the machine learning algorithm.

We set out with the knowledge that some relation holds between the two concepts denoted by the NPs, where the relation is expressed by the preposition. This knowledge could be expressed as a conceptual feature structure: CONCEPT1[REL:CONCEPT2], where REL is to be seen as a generic or uninstantiated relation, and CONCEPT1 is the NP in front of the preposition, and CONCEPT2 is the NP after the preposition. Our aim is to be able to exchange the generic relation REL with a more specific relation. For example, if we are able to determine that the given preposition in the given conceptual context of CONCEPT1 and CONCEPT2, denotes a partitive relation (POF), we can fill in the relation: CONCEPT1[POF:CONCEPT2].

The ability to identify such compound concepts in text facilitates content-based information retrieval, which is to be seen in opposition to a more traditional search approach where the information retrieval relies more or less exclusively on keyword recognition. In the content-based information retrieval framework of the OntoQuery project, the aim is to index texts according to the compound concepts that are identified through conceptual analysis of the NPs found in the texts. The conceptual content of each document is described as a set of arbitrarily complex conceptual feature structures that facilitate a detailed comparison of the conceptual content of documents. As a result, we can move from a linear structure, cf. keywords, to a graph structure that describes the concepts denoted by a given text in relation to each other. Thus, relevant documents can be retrieved based on a (partial) match between the conceptual content of a search term and of the documents (Andreasen et al., 2002; Andreasen et al., 2004). A given search term may result in retrieval of texts with different surface forms but with identical or similar conceptual content.

5.5 Semantic Relations

Semantic relations can exist at different syntactic levels; across sentence boundaries or within a sentence, a phrase or a word. They can be denoted by different parts of speech, such as a verb, a preposition or an adjective, or they can be implicitly present in compounds and genitive constructions. Semantic relations are n-ary. These properties are exemplified above in chapter 4.

In the framework of this experiment, we will only consider binary relations denoted by prepositions.

A given preposition can be ambiguous in regard to which relation it denotes. As an example, let us consider the Danish preposition *i* (Eng: in): The surface form *i* in ‘*A i B*’ can denote at least five different relations between the concepts denoted by *A* and *B*. These five relations are exemplified in examples (47) – (51):

(47)

<i>ændringer</i>	<i>i</i>	<i>stofskiftet</i>
changes	in	the metabolism

(48)

A Machine Learning Approach to Disambiguation of Semantic Relations

skader *i* *hjertemuskulaturen*
injuries in the heart musculature

(49)
mikrobiologien *i* *1800-tallet*
microbiology in the 19th century

(50)
antioxidanter *i* *ren* *form*
antioxidants in a pure form

(51)
forskelle *i* *saltindtagelsen*
differences in the salt intake

In (47), the preposition denotes a PATIENT relation (PNT) that holds between the concepts ÆNDRING and STOFSKIFTE. The PATIENT relation is a thematic role relation where the related concept is a patient of an event denoted by the core concept.

In (48), the preposition denotes a LOCATIVE relation (LOC) that holds between the concepts SKADE and HJERTEMUSKULATUR. The LOCATIVE relation denotes a location/position of one of the concepts compared to the other concept.

In (49), the preposition denotes a TEMPORAL relation (TMP) that holds between the concepts MIKROBIOLOGI and 1800-TAL. The TEMPORAL relation denotes a placement in time of one of the concepts compared to the other.

In (50), the preposition denotes a CHARACTERIZATION, or property ascription, relation (CHR) that holds between the concept ANTIOXIDANT and the compound concept FORM[CHR:REN]. The CHARACTERIZATION relation denotes a characterization of one of the concepts by a property.

In (51), the preposition denotes a WITH RESPECT TO relation (WRT) that holds between the concepts FORSKEL and the compound concept INDTAGELSE[WRT:SALT]. A WITH RESPECT TO relation is an

A Machine Learning Approach to Disambiguation of Semantic Relations

underspecified relation that denotes an 'aboutness' relation between the concepts.

The challenge is to deduce a set of rules that will predict the correct relation for a given preposition in a given conceptual context. Our working hypothesis, which is based on (Per Anker Jensen & Vikner, 2006), is that it is possible to predict the relation that a preposition denotes based on the ontological types of the surrounding NPs.

Our idea is to perform supervised machine learning that will take into account a variety of features including the surface form of the preposition and the conceptual level ontological type of the surrounding noun phrases, and on this basis be able to determine the relation that holds between noun phrases surrounding a preposition in text.

As input we need a training set that has been analyzed and annotated. Examples (52)-(56) provide examples of such analyses for the linguistic expressions in (47)-(51) above. The examples have been annotated with the ontological types of the NP-heads, and the relation denoted by the preposition. Here, the ontological types of the NP-heads are the most specific top ontology concept in the SIMPLE ontology.

(52)

Surface level ændringer i stofskiftet

Conceptual level cause change PNT change

(53)

Surface level skader i hjertemuskulaturen

Conceptual level state LOC body part

(54)

Surface level mikrobiologien i 1800-tallet

Conceptual level domain TMP time

(55)

A Machine Learning Approach to Disambiguation of Semantic Relations

Surface level antioxidanter i ren form

Conceptual level natural substance CHR physical property

(56)

Surface level forskelle i saltindtagelsen

Conceptual level quality WRT act

Given examples (52)-(56) as an input, we can imagine the following rules as output:

Tag the preposition *i* as a patient relation (PNT)

IF

the preceding NP head is of the type *cause change*

AND

the succeeding NP head is of the type *change*.

Tag the preposition *i* as a locative relation (LOC)

IF

the preceding NP head is of the type *state*

AND

the succeeding NP head is of the type *body part*.

Tag the preposition *i* as a temporal relation (TMP)

IF

the preceding NP head is of the type *domain*

AND

the succeeding NP head is of the type *time*.

Tag the preposition *i* as a characteristic relation (CHR)

IF

the preceding NP head is of the type *natural substance*

AND

the succeeding NP head is of the type *physical property*.

Tag the preposition *i* as a with respect to relation (WRT)

IF

the preceding NP head is of the type *quality*

AND

the succeeding NP head is of the type *act*.

5.6 The Corpus

The experiments have been carried out using a small annotated Danish language corpus consisting of sentences that contain prepositions surrounded by noun phrases (NPs). The corpus is a subset of the OntoQuery corpus, which is compiled of texts in the domain of nutrition, all stemming from *Den Store Danske Encyklopædi* (The Danish National Encyclopedia) (Lund, 1994-2002).

In order to establish a training set, a small corpus of approximately 950 sentences or 18,500 running words was extracted from the OntoQuery corpus. The sentences were selected on the grounds that they contained the phrase pattern we were interested in analyzing, namely NP-PREP-NP. Afterwards, all text chunks matching the pattern NP-PREP-NP were extracted from this corpus and annotated with information about part of speech, lemma, phrase boundary, phrase head, associated concept for the NP heads and relation for the prepositions³². We annotate lemmas with the most specific concept in the SIMPLE top ontology (cf. section 3.3.3.2), and thus the ontological type is very general. Any given NP will by definition receive the same ontological type annotation as the head of that given NP. For this reason, we have not invested any effort in producing (possibly compound) concepts reflecting the conceptual content of the whole NP, but simply mapped the head of the phrase to the SIMPLE top ontology.

All the text samples in our training corpus derive from nutrition-related articles in The Danish National Encyclopedia, and are thus not only limited domain-wise, but also of a very specific text type which can be classified as expert-to-non-expert. Thus, we cannot be certain that our results can be directly transferred to a broader or more general domain, or to a different text type. This aspect would have to be empirically determined.

5.6.1 Annotation

952 excerpts of the form NP-PREP-NP were extracted from the corpus and annotated with information about part of speech (POS), lemma, phrase

³² Extraction, POS-tagging and initial ontological and relation type annotation was done by Dorte Haltrup Hansen, CST, University of Copenhagen

A Machine Learning Approach to Disambiguation of Semantic Relations

boundaries and heads, ontological type and relation type for NP heads and prepositions, respectively. An example annotation for the text excerpt *To blodpropper i højre lunge* (En.:*Two blood clots in the right lung*) is given in Table 8.

Information type	Value					
Surface form	To	Blodpropper	i	højre	lunge	
POS	DN	NC	SP	AQ	NC	
Lemma	to	Blodprop	i	højre	lunge	
NP boundary	<		>	<		>
NP head		•			•	
Concept/relation		Disease	LOC		BodyPart	

Table 8 Example text annotated with all levels of information

Extraction, POS-tagging and initial ontological and relation type annotation was done by Dorte Haltrup Hansen, CST, University of Copenhagen. The lemmatization and head extraction was done automatically by use of a Danish word form lexicon and an NP grammar developed in the OntoQuery project. The ontological type assignment was done partly automatically by lookup in the SIMPLE ontology, and partly manually for linguistic expressions that are not connected to concepts in the ontology. The relation annotation has been done manually.

The tags used in the annotation on the various levels are:

POS-tags

The tagger uses a subset of the PAROLE tag set, consisting of 43 tags, which means that it is a low-level POS tagging with little morphosyntactic information. We only use the tags in order to identify NPs and prepositions, and thus the level of information in the tags is not important. At any rate, we do not need a more fine-grained information level than the one represented in the reduced PAROLE tag-set.

Ontological type-tags

The tags used for the ontological type annotation consist of abbreviations of the concept labels for the concepts in the SIMPLE top ontology. The top level concepts in the SIMPLE ontology comprise 151 concepts, and consequently the ontological type-tag set consists of 151 tags.

Relation tags

The tags used for the relation annotation derive from a minimal set of relations that have been used in earlier work. See e.g. (Per Anker Jensen & Fischer Nilsson, 2006; B. N. Madsen, Pedersen, & Thomsen, 2000, 2001; Nilsson, 2001). The tags used in annotation comprise a set of 12 tags (cf. Table 9).

The final manual relation annotation has been done by one annotator (this author). The ideal situation would probably have been to have several annotators annotate the whole corpus, so that the annotation would reflect a majority agreement and not the idiosyncrasies of one annotator.

5.6.2 The Ontological Type Annotation

As noted above, the ontological types used in the experiments derive from the SIMPLE top ontology (Alessandro Lenci et al., 2000; B. S. Pedersen, 1999). The heads of the phrases have been annotated with the most specific node, i.e. ontological type, in the top ontology. In the case of *blodprop* the annotation of ontological type is DISEASE, since DISEASE is the lowest node in the top ontology in the path from BLODPROP (blood_clot) to the top. This is illustrated in Figure 48, which shows the path from BLODPROP (blood_clot) to the top level of SIMPLE.

Thus, for the purpose of this project, we only consider one node for each concept: the most specific node in the top ontology. Another approach would be to consider the the full path to the top node, thus also including the path from the leaf node to the lowest node in the top ontology. For the example depicted in Figure 48, the full paths from BLODPROP to the top node would be (57) and (58). There are two possible paths to the top because the SIMPLE ontology allows for multiple inheritance: The concept DISEASE has an ISA relation to PHENOMENON as well as to AGENTIVE.

(57) blodprop→ hjerte_karsygdom→disease→phenomenon→
event→entity→top

(58) blodprop→hjerte_karsygdom→disease→agentive→top.

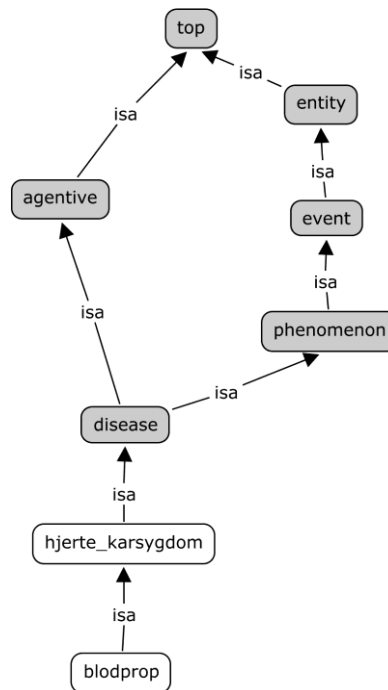


Figure 48 Ontology fragment showing the path from BLODPROP (blood clot) to the top level of the SIMPLE ontology. The grey nodes are top level nodes, and white nodes are domain level nodes.

5.6.3 The Set of Relations

For the purpose of the manual relation annotation, we need to decide on a finite set of possible relations that can be used in the annotation. This is a non-trivial task, as it is almost impossible, by introspection alone, to predict which relations prepositions *can* denote generally, and in the text type at hand specifically.

A Machine Learning Approach to Disambiguation of Semantic Relations

Relation	Description
TMP	temporal relations
LOC	location, position
PRP	purpose, function
WRT	with respect to
CHR	characteristic (property ascription)
CUM	cum (with, accompanying)
BMO	by means of, instrument, via
CBY CAU	caused by causes
CMP POF	comprising, has part part of
AGT	agent of act or process
PNT	patient of act or process
SRC	source of act or process
RST	result of of act or process
DST	destination of moving process

Table 9 The initial relation set consisting of 16 relations cf. (Nilsson, 2001)

The method that we decided to use was the following: An initial set of relations that have all been used in prior work, cf. e.g. (B. N. Madsen et al., 2000, 2001; Nilsson, 2001), were chosen as a point of departure. The final relation set was achieved by annotating the text segments using this set as the possible relation types, and the relations that were used in the annotation form the final subset that is used as input for a machine learning algorithm. The initial relation set is shown in Table 9, and the final subset is shown in Table10.

Relation	Description
AGT	Agent of act or process
BMO	By means of, instrument
CBY	Caused by
CAU	Causes
CHR	Characteristic (property ascription)
CMP	Comprising, has part
DST	Destination of moving process
LOC	Location/position
PNT	Patient of act or process
SRC	Source of act or process
TMP	Temporal aspects
WRT	With respect to

Table 10 The final relation set consisting of the 12 relations that were used in the annotation - a subset of the relation set proposed in (Nilsson, 2001).

5.7 Experiments

In order to discover any regularities that may exist in the data, we apply machine learning to our annotated data. The annotation process generates a feature space of six dimensions, namely:

- * The lemmatized form of the first NP
- * The ontological type of the first NP heads,
- * The preposition
- * The relation denoted by the preposition
- * The lemmatized form of the second NP
- * The ontological type of the second NP head

Since the corpus consists of only 952 text segments, data sparseness is a problem. In addition, the distribution of the data is highly skewed: More than half of the instances are annotated with the relation type WRT or PNT, and the rest of the instances are distributed among the remaining 10 relations with only 14 instances covering the three smallest classes, as illustrated in Figure 49.

The 952 preposition instances in the corpus are not evenly distributed among prepositions. Almost a third of the instances are of the preposition *af*, over half of the instances are of the prepositions *af* and *i*, and three

A Machine Learning Approach to Disambiguation of Semantic Relations

prepositions are unique or hapax legomena (*gennem*, *over* and *pr.*). The distribution of prepositions is shown in Figure 50.

There are 332 different combinations of ontological types, of which 197 are unique. Amongst the lemmatized NP heads, there are 681 different lemmas, of which 403 are unique.

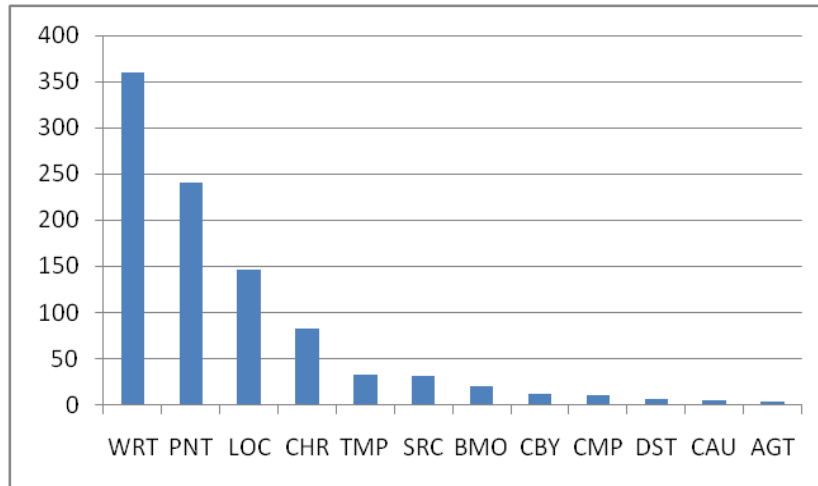


Figure 49 Frequency distribution for the 12 relations in the corpus

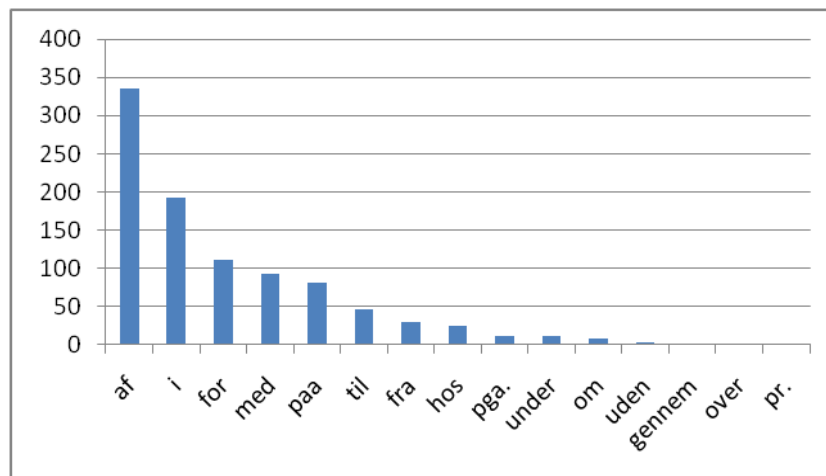


Figure 50 Frequency distribution for the 15 prepositions in the corpus

However, it is our assumption that the data show some regularity with respect to the relations that prepositions denote in particular contexts, and

hence the learning algorithms should be able to generalize well. We also hypothesize that the ontological type of the NP heads is the most vital information type in classifying the relation type, at least in this case where data is sparse.

We have performed the machine learning experiments using the Support Vector Machine algorithm SMO and the rule learning algorithm JRip. The former in order to get high accuracy in the results, and the latter in order to get easily interpretable rules for later analysis. The implementation of the algorithms that we used was the WEKA software package (Witten & Frank, 2005). The choice of SMO was arbitrary, and the choice of JRip was based on the fact that, for this dataset, it performed best (i.e. yielded the highest precision) of the rule learning algorithms implemented in WEKA.

The SMO algorithm³³ is an implementation of the Sequential Minimal Optimization support vector classifier using kernels (Keerthi, Shevade, Bhattacharyya, & Murthy). A support vector machine normally solves one large optimization problem, but the SMO algorithm essentially works by breaking down such a large problem into a series of the smallest possible problems which are then solved analytically.

The JRip algorithm³⁴ is an implementation of the rule learning RIPPER algorithm (Repeated Incremental Pruning to Produce Error reduction) (Cohen, 1995), which, in essence, works by first building rules that are 100% accurate by greedily adding all antecedents to the rules to maximize the information gain, and then pruning these rules to maximize the error reduction in a number of optimization runs.

The experiments were performed using 10-fold-cross-validation, with a further partition of the training set at each fold into a tuning and a training set. There are no hard and fast rules for setting up the size of these sets. However, the rule of thumb is to use 90% of the data set as a training set, and, when a tuning set is used, divide the remaining part in two: 9% of the data set serves as a test set and 1% as a tuning set (cf. Figure 51). We have

³³ Cf. <http://weka.sourceforge.net/doc/weka/classifiers/functions/SMO.html>

³⁴ Cf. <http://weka.sourceforge.net/doc/weka/classifiers/rules/JRip.html>

A Machine Learning Approach to Disambiguation of Semantic Relations

chosen to use a tuning set in these experiments, but not all machine learning experiments apply such a set.

What is the purpose of these subparts of the data set? A training set is used for building a model, a tuning set is used for tuning the learned rules, and a test set is used for testing the tuned model. Tuning is especially useful if the data is noisy (e.g. if the annotation is flawed, some instances have identical attribute-value pairs but belong to different classes, the data contains irrelevant dimensions, etc.) In the case of noisy data, a tuning set can be used for testing different parameter settings for learning rules on the training set, and thus optimizing the final rules.

Let us consider an example: We observe a group of 1000 people, of which 500 graduated from university, and 500 are former students who did not graduate. We have access to all their data concerning exams, group composition, professors, mailing list memberships, extracurricular activities, etc. Now we wish to find out if at an early stage we can predict whether a given student is going to graduate or not: How can we produce the most precise rules that will make this prediction? Since the students followed different educational directions, we worry that we may infer some rules that apply to linguists but not to biologists, or vice versa, or that we include parameters in our rules that are irrelevant, but seem to exhibit a regularity. The model that we arrive at must be universally applicable.

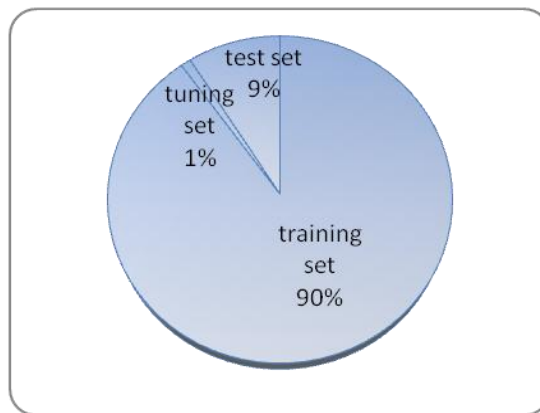


Figure 51 Distribution of the data set into a training, tuning and test set.

In order to build such a model, we start by mixing up the 1000 students, and

A Machine Learning Approach to Disambiguation of Semantic Relations

dividing them into 10 subsets of 100 students each (sets 1 to 10), and use sets 1-9 as a training set, and subdivide set 10 into a tuning set and a test set. We then build a model based on the training set, and initially test it on the tuning set. Based on the results of this test, we modify our model so that it fits better for the students in the tuning set. This process may, in principle, go on infinitely; however, repeating it excessively will result in overfitting the model to these 10 students. Finally, we test the model on the test set. Basically, the principle is to test and optimize the model on the basis of a tiny set of data that has neither been used for building the model, nor will be used for the final testing. This ensures that the model is not overfitted to the training set. It is important never to observe the remaining 90 students in the test set during the optimization process; if we consider them, we do not know how good our model is at predicting the unknown, and our investigation thus becomes without value.

For a 10-fold cross validation, this process is repeated 10 times. The second iteration with sets 1-8+10 as training set and set 9 as the tuning and test set, etc. The advantage of this method is that we have used our entire set of data for testing, tuning and training, and thereby avoided a potentially skewed distribution of the data from the onset. A skewed distribution of the data could result in incorrect results/estimates.

In our experiments, for the SMO algorithm, the tuning parameters were complexity, the applied kernel, and gamma for the RBF kernel. For the JRip algorithm, the tuning parameters were number of folds used for growing and pruning, minimum number of instances covered and number of optimization runs.

The experiments were performed on seven different combinations of the feature space, ranging from only using the lemmatized surface form of the heads to using the whole feature space (i.e. lemmatized NP heads, ontological types of the heads, preposition and relation). This was done in order to gain an insight into the importance of using ontological types in the learning process. The results of these experiments are shown in Table 11. The last column shows the accuracy for a projected classifier (PC) in the cases where it outperforms the trivial rejector. The projected classifier, in this case, assigns the relation that is most common for the corresponding input pair; e.g if the ontological type pair is DISEASE-HUMAN, the most common relation is PNT, which gets assigned. The trivial rejector assigns the most common relation in the whole data set to all the instances, in this

A Machine Learning Approach to Disambiguation of Semantic Relations

case WRT. Assignment by the trivial rejector here achieves an accuracy of 37.8%.

	Feature space	JRip	SMO	PC
1	Preposition	68.1	68.5	67.6
2	Lemma	61.7	73.3	-
3	Lemma and Preposition	71.8	83.4	-
4	Ontological types	73.1	77.0	61.8
5	Ontological types and Lemma	72.3	81.7	-
6	Ontological types and Preposition	81.4	86.6	-
7	Ontological types, Preposition and Lemma	81.4	88.3	-

Table 11 The percentage of correctly classified instances of SVM, JRip and a projected classifier (PC) on the seven different combinations of input features.

The following conclusions can be drawn from the results as shown in Table 11:

The support vector machine algorithm produces a result which in all cases is better than the baseline, i.e. we are able to produce a model that generalizes well over the training instances compared to the projected classifier as well as to the trivial rejector. This difference is not statistically significant at a confidence level of 0.95 when only training on the surface form of prepositions.

A comparison of lines 1-4 in Table 11 shows that training on ontological types seems to be superior to using lemmatized NP heads or prepositions or a combination of the two, although the superiority is not significant. When comparing lines 4-7, the differences between the results are not statistically significant. This fact may owe to data sparseness. However, comparing line

1 to line 6 or 7 shows that the improvement from adding the preposition and the lemmatized NP heads to the ontological types is statistically significant. These results indicate a good prospect for including ontological types of surrounding NPs in the identification of relations denoted by prepositions for information retrieval.

5.7.1 Analyzing the Rules

In this section we will take a more thorough look at the rules produced by the JRip algorithm. First, we examine the rules produced by the algorithm applied to the data set with only ontological types of the NP heads in the feature space that yields an accuracy of 73.1%. With only ontological types of the NP heads in the feature space, the JRip algorithm produced 22 rules. These rules are shown in Table 12. Next, we examine the rules produced by the algorithm applied to the data set with prepositions and ontological types of the NP heads in the feature space that yield an accuracy of 81.4%. These rules are shown in Table 14.

Table 12 shows the rules produced by the algorithm applied to the data set with only ontological types of the NP heads in the feature space. For this rule set, the most general rule is rule no. 18, which covers almost half of the instances. This rule is the default rule, that assigns all instances to the WRT relation if no other rules apply. At the other end of the spectrum, ten rules in combination (rules 1-10) cover no more than 34 instances, but with an accuracy of 100%. It is pointless to analyse all these rules in any depth, since they cover the most infrequent relations and hence they may be overfitting the data set. However, by looking at the rules, we can identify at least one of these rules that appears to be correct (rule 10): The rule says that if the ontology of the first NP head is DISEASE and the ontology of the second NP head is HUMAN, then the relation is PATIENT. This rule matches linguistic expressions as (59) and (60), and correctly associates both the preposition *hos* in (59) and *blandt* in (60) with the PATIENT relation.

(59)

jernmangel *hos* *kvinder*
iron deficiency in females

(60)

A Machine Learning Approach to Disambiguation of Semantic Relations

hudkraft *blandt* *afro-amerikanere*

skin cancer among African-Americans

	JRip Rule	Matches	Incorrect	Accuracy
1	(second_head_ontotype = QUA) => relation=CBY	2	0	1.00
2	(second_head_ontotype = STA) and (first_head_ontotype = NSU) => relation=TMP	2	0	1.00
3	(second_head_ontotype = EVE) => relation=TMP	2	0	1.00
4	(second_head_ontotype = AMO) and (first_head_ontotype = FOO) => relation=CHR	2	0	1.00
5	(first_head_ontotype = IST) => relation=CHR	2	0	1.00
6	(first_head_ontotype = BOD) and (second_head_ontotype = NSU) => relation=CMP	3	0	1.00
7	(first_head_ontotype = ATA) => relation=CHR	3	0	1.00
8	(first_head_ontotype = CCS) => relation=PNT	4	0	1.00
9	(first_head_ontotype = NSU) and (second_head_ontotype = MIC) => relation=LOC	6	0	1.00
10	(first_head_ontotype = DIS) and (second_head_ontotype = HUM) => relation=PNT	7	0	1.00
11	(second_head_ontotype = TIM) => relation=TMP	24	1	0.96
12	(first_head_ontotype = CRE) => relation=PNT	35	3	0.91
13	(first_head_ontotype = DIS) and (second_head_ontotype = CHA) => relation=CHR	4	1	0.75
14	(second_head_ontotype = UME) => relation=CHR	33	9	0.73
15	(first_head_ontotype = CHA) => relation=PNT	87	24	0.72
16	(first_head_ontotype = CAC) => relation=PNT	27	8	0.70
17	(first_head_ontotype = HUM) => relation=CHR	30	9	0.70
18	=> relation=WRT	452	134	0.70
19	(first_head_ontotype = ACT) => relation=PNT	103	33	0.68
20	(second_head_ontotype = STA) and (first_head_ontotype = CHA) => relation=TMP	3	1	0.67
21	(second_head_ontotype = ART) and (first_head_ontotype = ACT) => relation=BMO	15	5	0.67
22	(second_head_ontotype = BPA) => relation=LOC	106	35	0.67
	All rules	952	263	72.37

A Machine Learning Approach to Disambiguation of Semantic Relations

Table 12 The rules produced by the JRip algorithm for the data set with ontological types in the feature space, with number of matches, number of incorrect matches and accuracy score. The table is sorted by accuracy score.

One of the least surprising rules is rule no. 11; a rule with a high accuracy of 96%. This rule says that if the ontotype of the second NP head is TIME then the relation type is TEMPORAL. The rule covers linguistic expressions as (61) and (62).

(61)
diæt gennem mange måneder
a diet for many months

(62)
mikrobiologi i det 19. århundrede
microbiology in the 19th century

We hypothesize that this rule applies to a larger domain, including general language texts.

The rule with the second-highest coverage, but amongst the three rules with the lowest accuracy of 67%, is rule no. 22. This rule says that if the ontotype of the second NP head is BODY PART then the relation type is LOCATIVE. The rule matches linguistic expressions such as (63) and (64).

(63)
blodprop i hjertet
blood clot in the heart

(64)
jernoptagelse fra tarmen
iron absorbtion from the intestine

The rule correctly associates the preposition *i* in (63) with the LOCATIVE relation, but also incorrectly associates the preposition *fra* in (64) with the same relation. The annotated relation for (64) is SOURCE. However, this a

A Machine Learning Approach to Disambiguation of Semantic Relations

case where we may wish to reevaluate the annotation, and change the relation to LOC as the rule predicts.

With prepositions and ontological types of the NP heads in the feature space, the JRip algorithm also produced 22 rules, as shown in Table 14. The confusion matrix in Table 13 shows the classification distribution for the same rule set. For example, we can see that of the instances that denote the AGENT relation, one has been classified as the SOURCE relation and two have been classified as the PATIENT relation. None have been classified as the AGENT relation.

AGT	BMO	CAU	CBY	CHR	CMP	DST	LOC	PNT	SRC	TMP	WRT	classified as ←
0	0	0	0	0	0	0	0	2	1	0	0	AGT
0	18	0	0	0	0	0	0	2	0	0	0	BMO
0	0	0	0	5	0	0	0	0	0	0	0	CAU
0	0	0	11	0	0	0	1	0	0	0	0	CBY
0	2	0	0	57	1	0	5	1	0	3	15	CHR
0	0	0	0	7	0	0	0	0	0	0	3	CMP
0	0	0	0	0	0	0	0	0	0	0	6	DST
0	0	0	0	1	0	0	120	6	0	0	20	LOC
0	0	0	0	2	0	0	4	211	0	0	24	PNT
0	0	0	0	0	0	0	0	0	27	0	4	SRC
0	0	0	0	0	0	0	1	0	1	30	1	TMP
0	0	1	0	16	0	0	9	31	1	1	301	WRT

Table 13 Confusion matrix for the JRip algorithm applied with prepositions and ontological types in the feature space

A Machine Learning Approach to Disambiguation of Semantic Relations

	Jrip Rule	Matches	Incorrect	Accuracy
1	(preposition = af) and (first_head_ontotype = BOD) => relation=CMP	3	0	1.00
2	(preposition = pga.) => relation=CBY	11	0	1.00
3	(preposition = af) and (first_head_ontotype = SUB) => relation=SRC	2	0	1.00
4	(preposition = i) and (second_head_ontotype = BPA) => relation=LOC	65	0	1.00
5	(preposition = i) and (second_head_ontotype = GEO) => relation=LOC	16	0	1.00
6	(second_head_ontotype = TIM) => relation=TMP	23	1	0.96
7	(preposition = af) and (first_head_ontotype = CHA) => relation=PNT	65	3	0.95
8	(preposition = med) and (first_head_ontotype = ACT) => relation=BMO	20	2	0.90
9	(preposition = fra) => relation=SRC	30	3	0.90
10	(second_head_ontotype = UME) and (preposition = paa) => relation=CHR	19	2	0.89
11	(preposition = i) and (second_head_ontotype = MIC) => relation=LOC	9	1	0.89
12	(first_head_ontotype = CRE) => relation=PNT	36	4	0.89
13	(preposition = i) and (first_head_ontotype = NSU) => relation=LOC	25	3	0.88
14	(first_head_ontotype = CCS) => relation=PNT	7	1	0.86
15	(first_head_ontotype = CAC) => relation=PNT	32	6	0.81
16	=> relation=WRT	387	76	0.80
17	(preposition = af) and (first_head_ontotype = ACT) => relation=PNT	99	21	0.79
18	(first_head_ontotype = DIS) => relation=PNT	9	2	0.78
19	(preposition = under) => relation=TMP	11	3	0.73
20	(preposition = med) and (second_head_ontotype = SUF) => relation=CMP	3	1	0.67
21	(preposition = med) => relation=CHR	68	23	0.66
22	(preposition = i) and (second_head_ontotype = FOO) => relation=LOC	12	5	0.58
	All rules	952	157	0.84

Table 14 The rules produced by the JRip algorithm for the data set with prepositions and ontological types in the feature space, with number of matches, number of incorrect matches and accuracy score. The table is sorted by accuracy score.

A Machine Learning Approach to Disambiguation of Semantic Relations

The default rule, rule no. 16 that assigns all instances to the WRT relation if no other rules apply, here covers 40% of the instances. Five rules in combination (rules 1-5) cover 97 instances, or 10%, with an accuracy of 100%. All these five rules include the preposition in the rule.

Again, we see a rule saying that if the ontotype of the second NP head is TIME then the relation type is TEMPORAL without a specification of the preposition with an accuracy of 96%.

Comparing to rule no. 22 from Table 14, which says that if the ontotype of the second NP head is BODY PART then the relation type is LOCATIVE, we here have a rule that says that if the preposition is *i* and the ontotype of the second NP head is BODY PART then the relation type is LOCATIVE. The rule correctly assigns the LOCATIVE relation to *i* in (63) but not to *fra* in (64). Also, rule no. 9, which says that if the preposition is *fra* then the relation type is SOURCE, covers (64) and correctly assigns the SOURCE relation to *fra*. This rule is indeed very general, and it will incorrectly assign the SOURCE relation to *fra* in many cases where the preposition denotes other relations. Learning on a larger dataset would presumably result in more fine-grained rules including this preposition.

(63)

blodprop i hjertet

blood clot in the heart

(64)

jernoptagelse fra tarmen

iron absorbtion from the intestine

Generally, the rules produced by the algorithm with prepositions and ontological types in the feature space perform better than the ones with just ontological types. We hypothesize that the difference would be more significant with a larger corpus as input to the learning process.

5.8 Summary

In this chapter, we have described experiments with a machine learning approach to disambiguation of semantic relations denoted by prepositions. We have asserted that the task is similar to, but not identical to, word sense

A Machine Learning Approach to Disambiguation of Semantic Relations

disambiguation. The difference lies primarily in the purpose. Our purpose is to produce conceptual feature structures representing the conceptual content of complex linguistic expressions of the form NP-PREP-NP.

Even though the experiments have been performed on a limited test corpus, our experiments have shown that it is indeed possible to infer rules that predict the relation denoted by a preposition – at least within the domain of nutrition covered by the corpus and the domain ontology in these experiments. We have performed experiments with the inclusion of a variety of feature combinations, and compared the results. The results indicate that the inclusion of ontological types in the feature space does increase precision.

Future work includes annotation and investigation of a general language corpus. Also, a more thorough examination of the relations and prepositions could be also prove beneficial. Such work is described in the following chapter 6.

Chapter 6

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

The work described in this chapter is similar to, and builds on, the work described above in Chapter 5; however, we here use a larger general language corpus and a larger set of semantic relations that are the result of a thorough analysis of dictionary entries for a subset of Danish prepositions. Our aim is to test the hypothesis that ontological affinities exist between the ontological types of the NPs and the relation that prepositions denote in linguistic expressions of the form NP-PREP-NP in general language texts. What we aim at being able to do is to perform automatic relation disambiguation of prepositions, i.e. we attempt to uncover the semantic relation denoted by a given preposition in a given conceptual context by use of rules.

In order to meet this goal, we perform supervised machine learning on an annotated corpus of Danish general language texts. The bottleneck of any supervised machine learning task, including our experiments described above in chapter 5 and below in this chapter, is the availability of annotated data. For our experiments, we need a corpus that has been annotated with ontological types of NP heads, as well as with the relation denoted by preposition for NP-PREP-NP constructions. We do not have access to a corpus pre-annotated with these types of information for Danish, and thus we have to produce it ourselves. This is a highly labor intensive task, and since this author is the only person allocated to the task, the attainable size is limited.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

The experiments described in this chapter have been carried out using a small annotated Danish language corpus consisting of 2925 sentences that all contain one of a preselected set of prepositions, surrounded by noun phrases (NPs). The corpus is a subset of a citation version of the Danish language corpus Korpus 2000. For this corpus, we analyze all text chunks that have the form NP-PREP-NP, and annotate them with information about lemmas of the NP heads, associated concepts for the NP heads and semantic relation for the prepositions.

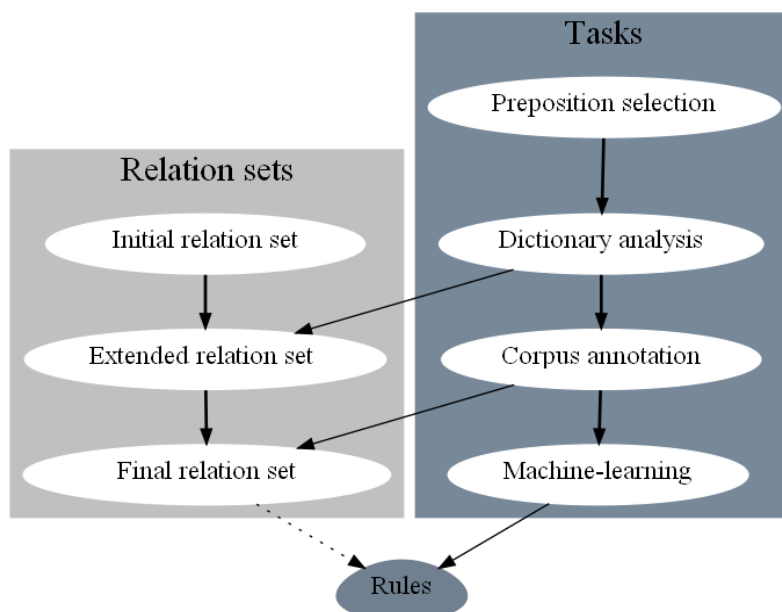


Figure 52 The tasks and relation sets involved in the experiments

In order to be able to annotate the prepositions with an appropriate semantic relation, we first select a subset of Danish prepositions, and analyze dictionary entries for these, as well as corpus examples containing the given preposition in the relevant syntactic form. As a result, we can list the possible semantic relations that this selected set of prepositions may denote. The annotated text chunks then become the input to a series of rule producing machine learning algorithms. The resulting rules are analyzed and refined before a subset of the inferred rules is 1) given in a form that may be used in applications, and 2) transformed into a *dictionary of prepositions* that expresses the obtained knowledge about relations denoted by the

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

selected set of prepositions. Figure 52 shows the various tasks and relation sets involved in the experiments.

The structure of this chapter is as follows: Section 6.1 describes the process of selecting a subset of Danish prepositions for further analysis. Next, in section 6.2, we describe the initial set of semantic relations that form the outset from which we produce the final relation set that the selected prepositions can denote. The next section, section 6.3, is a description of an analysis of dictionary entries for the selected prepositions. The result of this analysis is a pre-final relation set covering the relations that the set of selected prepositions denote according to descriptions in existing reference works. This set is described in more detail in section 6.4. Section 6.5 describes investigations in a Danish general language corpus, Korpus 2000, including the compilation of a subcorpus as well as the following analysis and annotation of this subcorpus. The subcorpus consists of corpus-evidences for the selected set of prepositions. Section 6.6 describes, for each preposition, the results of the analysis of the subcorpus. Section 6.7 describes the rules that can be deduced from the annotated corpus: First, we deduce purely frequency-based rules, next we apply machine learning in order to deduce more complex rules. For the results of the machine learning experiments, we analyze and evaluate selected rules in section 6.8; in section 6.8.1, we describe the 10 most precise rules, in section 6.8.2, we describe the 10 most covering rules and, finally, in section 6.8.3, we describe the 10 ‘best’ rules according to a rule quality score. In section 6.9, we describe a resulting dictionary of prepositions. Finally, in section 6.10, we summarize.

6.1 Selection of Prepositions for Further Analysis

This section describes the method used for selecting a subset of Danish prepositions for further analysis.

Defining the set of prepositions in the Danish language based on existing descriptions is not a trivial task and, thus, selecting a subset of Danish prepositions is not a trivial task either. The preposition inventories in reference works on the Danish language differ strongly; the differences may either be a result of different definitions of the class, or be the result of different criteria or methods for lemma selection for the given reference work. We cannot, based on an extraction of all prepositions in all Danish

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

reference works, expect to be able to produce a consensual list that we can pick from, and thus we have to use some heuristics for producing a subset of Danish prepositions for further analysis.

Our aim is to end up with a set of 10-20 entities that:

1. have a significant frequency count in a Danish general language corpus
2. are classified as prepositions in a selection of reference works

For the first criterion, we stipulate that the preposition has to be in the top 25 of frequency counts for all tokens that are tagged as prepositions in the Danish corpus, Korpus 2000.

Rank	Preposition	Frequency in Korpus 2000
1.	i	726 908
2.	til	362 984
3.	på	347 052
4.	af	339 848
5.	for	317 962
6.	med	278 265
7.	som	230 403
8.	om	170 646
9.	fra	114 634
10.	ved	76 659
11.	over	56 229
12.	efter	54 682
13.	end	32 821
14.	mod	30 076
15.	under	25 980
16.	siden	23 859
17.	mellem	21 772
18.	uden	21 772
19.	for	17 716
20.	blandt	14 940
21.	hos	14 824
22.	inden	14 641
23.	omkring	13 383
24.	gennem	12 060
25.	ifølge	9 886

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Table 15 frequency counts for the 25 most frequent prepositions in Korpus 2000.

For the second criterion, we stipulate that the preposition must be listed as a preposition in all of a selection of reference works. The reference works chosen for this check was the following seven sources:

1. Nudansk Ordbog (NO)
2. Den Danske Ordbog (DDO)
3. Vinterberg & Bodelsen Da-Eng (VB)
4. Ordbog over det Danske Sprog (ODS)
5. Retskrivningsordbogen (RO)
6. Ordbog Over Præpositioner (OOP)
7. Brøndal, Præpositionernes Theori (BR)

In combination, these sources described 222 different words classified as prepositions, ranging from the smallest number (16) in BR to the largest number (151) in ODS. A total number of 15 prepositions were present in all seven sources, as shown in Table 16.

Of the 25 prepositions with the highest frequency in Korpus 2000 listed in Table 15, 14 prepositions met the second criterion. Of the 15 prepositions that were present in all 7 sources, only the preposition *ad*, which ranked 32 in the frequency list, did not meet the criterion.

Thus, the prepositions selected for further analysis are the following 14: *af, efter, for, fra, gennem, hos, i, med, mellem, over, på, under, til* and *ved*.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Preposition	NO	DDO	VB	ODS	RO	OOP	BR
ad	•	•	•	•	•	•	•
af	•	•	•	•	•	•	•
efter	•	•	•	•	•	•	•
for	•	•	•	•	•	•	•
fra	•	•	•	•	•	•	•
gennem	•	•	•	•	•	•	•
hos	•	•	•	•	•	•	•
i	•	•	•	•	•	•	•
med	•	•	•	•	•	•	•
mellem	•	•	•	•	•	•	•
over	•	•	•	•	•	•	•
på	•	•	•	•	•	•	•
under	•	•	•	•	•	•	•
til	•	•	•	•	•	•	•
ved	•	•	•	•	•	•	•

Table 16 Prepositions present in all 7 selected sources.

6.2 Semantic Relations denoted by Prepositions

In this experiment, as in the experiments described above in Chapter 5, we take the relations proposed (Nilsson, 2001) as an outset. Here, we combine the relation inventory with an earlier suggestions in (Nilsson, 1999) and arrive at the initial relation set shown in Table 17. As mentioned in section 5.7, a more thorough examination of the relations and prepositions could be beneficial. We describe such a thorough examination in the following sections before we arrive at a final relation inventory.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Relation	Legend	Nils01	Jfn-workP
TMP	temporal relations	•	•
LOC	location, position	•	•
PRP	purpose, function	•	•
WRT	with respect to	•	•
CHR	characteristic (property ascription)	•	
CUM	cum (with, accompanying)	•	•
BMO	by means of, instrument, via	•	•
QUA	according to		•
DE	Possession		•
MNR	Manner		•
CBY	caused by	•	
CAU	causes	•	
CMP	comprising, has part	•	
POF	part of	•	
AGT	agent of act or process	•	•
PNT	patient of act or process	•	•
SRC	source of act or process	•	•
RST	result of of act or process	•	
DST	destination of moving process	•	•

Table 17 Relation inventory as a combination of suggestions in (Nilsson, 2001) and (Nilsson, 1999)

6.3 Investigating Preposition Senses

In the search for an inventory of semantic relations denoted by prepositions, we analyze preposition entries in a series of five reference works, to a great extent overlapping with the selection for extraction of preposition inventory above:

For each of the 14 prepositions that were selected for further analysis, a list of dictionary articles was prepared. The dictionaries from which the articles were extracted are:

1. Nudansk Ordbog
2. Ordbog over det Danske Sprog (online version)

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

3. Den Danske Ordbog³⁵
4. Vinterberg & Bodelsen Dansk-Engelsk
5. Dansk-Engelsk Ordbog Over Præpositioner

Below, each of the 5 dictionaries is briefly described:

Nudansk Ordbog (NDO) (Christian Becker-Christensen, 2005) is a dictionary of contemporary Danish, published by the publishing house Politikens Forlag. The dictionary was first published in 1953, and has since been published in 19 editions. The dictionary is known for its readiness to add new words, word senses and pronunciations to new editions.

Ordbog over det Danske Sprog (ODS) (Dahlerup, 1918-56) is a comprehensive reference work of the Danish language in 28 volumes. It was published by Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature) over a period of almost 40 years from 1918-56, and covers the Danish vocabulary in the period from 1700 to approx. 1950. During the years 2004-5, the complete work was digitalized and subsequently made freely available on the Internet (online version: <http://ordnet.dk/ods/>).

Den Danske Ordbog (DDO) (Hjorth & Kristensen, 2003-2005) is a comprehensive reference work of contemporary Danish in six volumes. It was published in the period from 2003-05 by Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature), and covers the Danish vocabulary from the 1950s to today. The dictionary is the successor of ODS. As of 2009, the work is freely available on the Internet (online version: <http://ordnet.dk/ddo/>).

Vinterberg & Bodelsen Dansk-Engelsk (VB) (V. H. Pedersen, 1999) is a bilingual dictionary for the word pair Danish-English, published by the publishing house Gyldendal. The dictionary is the most comprehensive Danish-English dictionary.

³⁵ An extract of preposition articles was kindly provided by Det Danske Sprog- og Litteraturselskab (Society for Danish Language and Literature)

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Dansk-Engelsk Ordbog Over Præpositioner (OOP) (Schwarz, 2007) is a bilingual dictionary of prepositions for the word pair Danish-English, published by the publishing house Handelshøjskolens Forlag.

The specific works were chosen for different reasons. NDO was chosen because it is probably the most widely used desk dictionary for Danish, and because of its readiness to adapt to the state of the language. ODS was chosen because it is the most comprehensive work for the Danish language, however in many cases archaic. Hence, its successor ODS is the most comprehensive work for contemporary Danish. This work further has the advantage that the lemma selection as well as the sense analysis is corpus-based. The last two dictionaries, VB and OOP, were chosen because of their bilinguality as well as their comprehensiveness. Furthermore, OOP is an obvious choice since it is a dedicated preposition dictionary. The considerations behind the choice of bilingual dictionaries as part of the basis for the analysis are partly based on the idea of semantic mirrors (Dyvik, 1998), where e.g. the idea is put forward that semantically closely related words are likely to have strongly overlapping sets of translations. For prepositions, this means that we can expect closely related senses to share a translation equivalent, and thus we can anticipate that bilingual dictionaries will group related preposition senses based on different criteria than those of a conventional monolingual sense distinction. This difference may result in useful differences in sense distinctions. We could equally have chosen other language pairs than Danish-English. Especially language pairs where the target language has case marking as e.g. German could turn out to be fruitful sources. As a matter of fact, a Danish-German dictionary was part of the list for an initial analysis of selected prepositions, however, it turned out that the sense distinction in this source was too fine grained for locative and temporal senses in particular, and thus it was not included in the final selection.

For the short list of the 14 selected prepositions, a file containing definitions from a selection of lexicographic works was prepared. An initial relation inventory that derives from (Nilsson, 2001) and (Nilsson, 1999), and is shown above in Table 17 was used as a starting point for marking up each sense in the lists with an appropriate relation.

The original proposal from Nilsson, suggested that the TMP relation could be divided into seven subrelations (START, END, DUR, TENSE (past,

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

present,future), PER, FREQ, DELAY). However, for various reasons, as discussed elsewhere, we do not allow for subrelations.

In ODS, the preposition entries are very fine grained in the sense that each preposition has many senses, and each sense typically has a number of subsenses, and furthermore, is extremely verbose. This makes the sense division in ODS far more fine grained than what we aim at. For this reason, only the most general sense levels in ODS were analyzed.

Prepositions as part of idiomatic or metaphoric expressions are not assigned a semantic role.

If none of the senses from the initial inventory fitted a given sense description in a dictionary entry, a new relation was added to the inventory. Correspondingly, if none of the relations from the initial relation inventory were used, they were removed from the final relation inventory. Our aim is to decide on a closed inventory of relations, for reasons discussed in chapter 4, and we want the inventory to be as small as possible. However, we also want the inventory to reflect the identified general sense divisions. Thus, we only allow for a new relation to be added to the inventory if we cannot justify examples to be annotated with an existing relation type.

When all senses from all five sources had been assigned a relation covering the described sense, a list of assigned relations were extracted from each preposition file, and further analyzed. Added relations were evaluated, and the added relation may either be accepted or assigned to an existing relation type. Some existing relations may in this process be made broader. In other cases, the relation is accepted as necessary, and added to the relation inventory. For each preposition, a list of possible relations that the preposition may express along with one or more examples from the dictionary entries is produced. This list serves as an aid in the relation annotation process.

As a result, we can produce a preposition/relation matrix where for each preposition, its possible relational level is marked. This matrix is shown in Table 18.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

	AF	EFTER	FOR	FRA	GENNEM	HOS	I	MED	MELLEM	OVER	PÅ	TIL	UNDER	VED
ADD		o	o							o	o	o		
AGT	o					o						o	o	o
BMO	o		o	o	o		o	o		o	o	o	o	o
CAU													o	
CBY	o		o	o			o	o		o	o		o	o
CHR	o		o				o	o				o	o	o
CMP	o						o	o	o		o			
COM		o	o				o	o		o	o	o	o	
CUM								o				o		
TAR		o					o							
INH	o		o			o	o			o	o	o		o
LOC	o	o	o	o	o	o	o	o	o	o	o	o	o	o
MEA	o		o	o			o	o		o	o	o	o	o
MNR								o			o		o	
MTH				o			o					o		
PNT	o	o	o	o	o	o	o	o	o	o	o	o		o
POF	o						o	o	o		o		o	
PRP		o	o				o					o		
QUA		o	o					o					o	
RCH	o					o				o		o	o	o
RLO								o						
RST							o					o	o	
SBT			o											
SRC	o	o		o	o	o								
HPR	o													
HPO													o	
SUP													o	
TMP	o	o	o	o	o		o	o	o	o	o	o	o	o
WRT	o		o	o		o	o	o		o	o	o		o

Table 18 Matrix of prepositions and relations based on the dictionary analysis

6.4 Final Relation Set – The Relations One by One

Below, each relation in the set is described in detail with paraphrases and examples containing prepositions that denote the given relation. For some relations, subrelations are given. Since we do not make use of a relation

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

hierarchy, these subrelations are merely informative, and serve as a specifications of the relation. In addition, the subrelations are useful in connection with the manual annotation task, since a subrelation typically gives rise to a corresponding paraphrase that can be useful for identifying the superrelation.

6.4.1 ADD – ADDITION

This relation relates an entity *a* with another entity *b* that is added to *a* sequentially: an addition of something, an accumulation. The relation exists between concrete entities as well as abstract entities.

Note that this relation does not relate numeric values that are added to each other mathematically; these are related via the MTH –relation.

Paraphrases:

a 'tilføjet' *b*

a 'added to' *b*

a 'lagt til/oven i' *b*

a 'on top of' *b*

Examples:

(Vi mister) det ene arbejde efter det andet

(We are losing) one job after another

Lag på lag

Layer upon layer

Én for én

One by one

6.4.2 AGT AGENT

Subrelations:

BENEFACTOR

EXPERIENCER

This relation is a generic agent relation that comprises all types of relations between an entity *a* that is an event, and a doer *b* (an agent, benefactor or

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

experiencer). Volition or intention is not required of *b* and the patient of *a* may or may not undergo change as a result of the act or process.

a '(som) er udført af' *b*
a '(that is) executed by' *b*

a '(som) er oplevet af' *b*
a '(that is) experienced by' *b*

En madonna af Raffael
A madonna by Raffael

Teksterne er udgivet ved en professor i latin
The texts are published by a Latin professor

Han er populær hos chefen
He is popular with the boss

6.4.3 BMO 'BY MEANS OF'

Subrelations:
INSTRUMENT
MEDIUM

Relates an event *a* with a means *b*.
b is intentionally applied in order to achieve/transmit *a*.

Paraphrases:
a 'ved hjælp af' *b*
a 'by means of' *b*

a 'ved anvendelse af (instrumentet)' *b*
a 'using (the instrument)' *b*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a 'via mediet' *b*

a 'via the medium' *b*

Examples:

Der kom en meddelelse over radioen

A message arrived over the radio

Kennedy faldt for en morderkugle i Dallas

Kennedy fell by a bullet in Dallas

Motorcyklen kører 50 km på 1 liter benzin

The motorcycle runs 50 km on 1 liter of gasoline

6.4.4 CHR HAS CHARACTERISTIC

The HAS CHARACTERISTIC relation is a property ascription relation, that relates an entity *a* with a property *b* that is ascribed to *a*.

This relation has the inverse relation RCH.

a 'er karakteriseret ved' *b*

a 'has the characteristic' *b*

Det er hende i den grønne kjole

It is her in the green dress

En mand af vælde og formue

A man of power and wealth

Han er ved godt helbred

He is in good health

6.4.5 RCH INVERSE CHARACTERISTIC

The INVERSE CHARACTERISTIC relation is an inverse property ascription relation that relates a property *a* with an entity *b*.

This relation has the inverse relation CHR.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Paraphrases:

a 'er et karakteristikum ved/karakteriserer' *b*

a 'is a characteristic of/characterizes' *b*

Examples:

Der er noget skummelt over ham

There is something sinister about him

En Djævelinde af en Kone

A she-devil of a wife

Jeg kan godt se styrken hos hende

I see the strength in her

6.4.6 CMP COMPRISING, has part

Subrelations (cf. (Winston, Chaffin, & Herrmann, 1987)):

COLLECTION-MEMBER

INTEGRAL OBJECT-COMPONENT

MASS-PORCION

OBJECT-STUFF

ACTIVITY-FEATURE

AREA-PLACE

This relation has the inverse relation POF.

The COMPRISING relation relates two entities *a* and *b* where *a* has *b* as a part(s). The individual subrelations each restrict the ontological types of the relates differently.

Note that the subrelation AREA-PLACE (e.g. Florida-Everglades) is closely related to RLO (INVERSE LOCATIVE). It is true for all entities that are related via AREA-PLACE that they are also related via a RLO relation, but not vice versa. In such cases, both RLO and CMP may be applied.

The material relation, which was an independent relation in the initial relation set, is now a subrelation of the CMP-relation (OBJECT-STUFF).

Paraphrases:

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a 'har (bla.) bestanddelen(e)' *b*
a 'has the constituent(s)' *b*

a 'består af' *b*
a 'consists of' *b*

a 'er lavet af' *b*
a 'is made of' *b*

a 'kan opdeles i' *b*
a 'may be subdivided into' *b*

a 'har delområdet' *b*
a 'has the subregion' *b*

as *b* (genitiv)
a 's' *b* (genitive)

Examples:

En bog med plasticomslag
A book with plastic binding

Grupper på fire personer
Groups of four people

Flokke af svaner
Flocks of swans

Hendes håndtaske af sort læder
Her purse of black leather

6.4.7 POF PART OF

Subrelations:

COMPONENT-INTEGRAL OBJECT

MEMBER-COLLECTION

PORTION-MASS

STUFF-OBJECT

FEATURE-ACTIVITY
PLACE-AREA

Relates two entities *a* and *b* where *a* is a part of *b*.
The individual subrelations each restrict the ontological types of the relates differently.

Note that the subrelation PLACE-AREA (e.g. Everglades – Florida) is closely related to the LOCATIVE relation. It is true for all entities that are related via PLACE- AREA that they are also related via a LOC relation, but not vice versa. In such cases, both LOC and POF may be applied. CMP

a 'udgør (en del af)' *b*
a 'constitutes (part of)' *b*

Bladene på træerne
The leaves on the trees

Hovedet på en knappenål
The head of a pin

Hun var mellem de få udvalgte
She was amongst the few chosen ones

Fremmed valuta hører under min kollegas område
Foreign currency falls under my colleague's responsibilities

6.4.8 COM COMPARISON

Subrelations:
MODEL
EQUIVALENCE
RANK

Relates two comparable entities *a* and *b*, between which there is som degree of (in)equivalence, (dis)resemblance, (dis)similarity.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Note that the relation is often expressed by a verb in combination with a preposition.

Paraphrases:

a 'er modelleret efter'/i stil med' *b*

a 'is modelled after/like' *b*

a 'tæller som' *b*

a 'counts as' *b*

a 'svarer til' *b*

a 'corresponds to' *b*

a 'er vigtigere end' *b*

a 'is more important than' *b*

a 'er mindre vigtig end' *b*

a 'is less important than' *b*

a 'er bedre end' *b*

a 'is better than' *b*

a 'er dårligere end' *b*

a 'is worse than' *b*

Examples:

Et menneske i vort billede

A man in our image

En kaptajn rangerer under en oberst

A captain ranks below a colonel

En sikker kopi efter Raphael

A definite copy after Raphael

Han trak på skuldrene til svar

He shrugged his shoulders in answer

6.4.9 CUM CUM (with, accompanying)

Relates two entities *a* and *b*, where *b* accompanies *a*.

Note that this relation should not be confused with ADD, where *b* is added to *a* sequentially. Here, two entities coexist/occur.

Paraphrases:

a 'med' *b*

a 'with/accompanying' *b*

a 'inklusive' *b*

a 'including' *b*

Examples:

En middag med vin

A dinner with wine

Jeg giver 887 kroner med varme

I pay 887 kroner with heating

Jeg går med Thomas hen til købmanden

I walk with Thomas to the grocer's

6.4.10 SRC SOURCE (of event)

Relates an event *a* with an entity *b* which is the source of *a*.

Note that for this relation *b* is not to be confused with a locative source relation where *b* is a location. However events may be related to metonymic readings of location names by a SRC relation (e.g. 'statsministeren modtog en forespørgsel fra Italien' 'The prime minister received an inquiry from Italy')

Paraphrase:

a 'fra' *b*

a 'from' *b*

Examples:

At rekruttere sit ordforråd hos dialekterne

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Recruiting one's vocabulary from the dialects

(Hus og gods) arves efter forældre
(House and goods) is inherited from the parents

Jeg hørte det gennem en bekendt
I heard it through an acquaintance

6.4.11 TAR TARGET (of event)

Relates an event *a* with an entity or a state *b* which is the potential destination of *a*.

a is directed towards *b*, but not *b* (yet).
Volition or intention to reach *b* is not required of the agent.

Note that for this relation, *b* is not to be confused with a locative destination relation where *b* is a location.

a 'rettet mod' *b*
a 'directed towards' *b*

Examples:
Higen efter kærligheden
Craving for love

(Han) stræber efter anerkendelse
(He) strives for recognition

(Hun er under) uddannelse til scenograf
Lit: (She is under) training to scenographer
(She is) training to be a scenographer

6.4.12 RST RESULT

Relates two entities *a* and *b*, where *b* is the (achieved) result of an act or process *a*.

Paraphrases:
a 'resulterer i' *b*
a 'results in' *b*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a 'er gjort til' *b*
a 'is turned into' *b*

Examples:

Slå til plukfisk

Lit: Punch into stewed codfish

Beat somebody to a pulp

Gøre vand til vin

Turn water into wine

(Hun blev) forfremmet til direktør

(She was) promoted to manager

6.4.13 CAU CAUSES

Relates a causing event *a* with a caused event *b*.

a causes an event *b* to occur, or causes *b* to undergo some form of change (of state). Volition or intention is not required of *a*.

This relation has the inverse relation CBY

Paraphrases:

a 'bevirker at' *b*

a 'triggers' *b*

a 'forårsager' *b*

a 'causes' *b*

Example:

(...) at ingen under vis Pengestraf maatte bruge Salvie

Lit: (...) that no one under certain monetary penalty could use sage

Using sage will result in a monetary penalty

6.4.14 CBY CAUSED BY

Relates a caused event *a* with a causing event *b*.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a undergoes some form of change (of state) or is an event caused by *b*. Volition or intention is not required of *b*.

Note that often, both a CBY and a TMP relation holds between entities that are related via a causal relation (e.g. ‘hun gjorde store fremskridt efter behandlingen’). In such cases, both CBY and TMP may be applied.

This relation has the inverse relation CAU

Paraphrase:

a '(er) forårsaget af' *b*

a '(is) caused by' *b*

Examples:

(Björn Borg) mistede mange millioner på fejlslagne forretninger

(Björn Borg) lost many millions in unsuccessful transactions

Han græder over sin fars død

He is crying over his father's death

Hun vågnede ved lyden af en spinkel fløjte

She woke up by the sound of a delicate whistle

6.4.15 INH INHERENT RELATION

This relation is a realization of the inherent relations in relational nouns.

a is related to *b* by the relation inherently present in *a*.

Note that this relation is primarily to be used if no other relation applies, that is, the inherent relation does not naturally fall under any of the relations described here. For the example ‘Kaptajn på skibet’ (LIT:captain on the ship) the relation between ‘kaptajn’ and ‘skib’ can be analysed as an inherent ‘captain of’ relation, but it may also be analysed as a LOC relation. Also, in such cases, both LOC and INH may be applied.

This relation does not have a general paraphrase, but may be identified through a couple of tests.

If we can ask the question e.g. ‘for whom/what is he an X?’, and the question cannot be answered with ‘nobody/nothing’, the noun is relational.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

If the answer ‘nobody/nothing’ is acceptable, the noun is not relational.

repræsentant for Miele (representative for Miele)

For the noun ‘representative’ in the above example, if we pose the question ‘for whom/what is he a representative?’ the answer is ‘Miele’. The answer ‘nobody/nothing’ would not be acceptable here – a person cannot be a representative for nothing!

Examples:

Professor i tysk
Professor of German

Verdensmester i skak
Lit: World champion in chess
World chess champion

Arving til tronen
Heir to the throne

Rektor for denne skole
Principal of this school

Et mindesmærke over slaget ved Odden
A monument over the battle at Odden

6.4.16 LOC LOCATIVE

Subrelations:

STATIC

DYNAMIC

VIA

DIRECTION

START

END

AREA

POINT

This general locative relation comprises all types of locative relations between a locatee *a* and a location *b*.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a may be a concrete or abstract entity that is situated in or in close proximity to *b*, or moving into, through, over, to or from *b* or a position near *b*. *a* may or may not be in physical contact with *b*. *b* is a concrete entity or a concrete reading of an abstract entity (e.g. company).

Note that when *b* is the location for storage of information (a semiotic artefact), and hereby also a medium by which communication can be transferred, this relation is closely related to the BMO - MEDIUM relation. In such cases, both LOC and BMO may be applied.

As mentioned above in section 6.4.7, it is also worth noting that the relations LOC and POF are very closely related. In many cases, both LOC and POF are appropriate analyses of a given preposition in a given context. For example, in an example such as '*udkanten af Paris*' (the outskirts of Paris), it is a valid analysis to say that *Paris* is a whole that has parts, and that *the outskirts* are such parts, and thus the POF relation applies. However, it is also possible to say that *the outskirts* are located relative to *Paris*, and thus the LOC relation applies. In an information retrieval setting, it is not always desirable to disambiguate such examples, since it minimizes the chance of retrieving appropriate matches to a query. For this reason, both LOC and POF may be applied in such cases.

This relation has the inverse relation RLO (Inverse LOcative).

Paraphrases:

a 'befinder sig i/over/under/i nærheden af lokaliteten' *b*

a 'is located in/over/under/near the location' *b*

a '(foregår) i (nærheden af) lokaliteten' *b*

a '(happens) near/at the location' *b*

a '(foregår) via lokaliteten' *b*

a '(happens) via the location' *b*

a '(foregår) i retning af' *b*

a '(happens) in the direction of' *b*

a '(foregår) med startpunkt i lokaliteten' *b*

a '(happens) with a starting point in' *b*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a '(foregår)med endepunkt i lokaliteten' *b*
a '(happens) with an end point in' *b*

Examples:

Den første station efter Aarhus
The first station after Aarhus

Bolden ramte ham i hovedet
The ball hit him on the head

De sejlede med vinden
They sailed with the wind

Han bor endnu hos sin mor
He still lives with his mother

Han trak dynen over hovedet
He pulled the duvet over his head

Jeg fik en vabel under foden
I got a blister under my foot

Miriam var blevet stående midt mellem bordet og døren
Miriam kept standing right between the table and the door

Øl på dåse
Beer in a can

6.4.17 RLO INVERSE LOCATIVE

An inverse locative relation between a location *a* and a locatee *b*.

This relation has the inverse relation LOC.

Paraphrase:

a 'er lokalitet for' *b*
a 'is the location of' *b*

Examples:

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

En kurv med blomster
A basket of flowers

En kasse med champagne
A box of champagne

En flaske med appelsinvand
A bottle of orange soda

6.4.18 MEA MEASURE

Subrelations:

ABSOLUTE

MAX

MIN

A relation between an entity a and some value b (on a scale). b may be an absolute value, or a maximum or minimum value.

a ' er målt/vurderet til ' b
 a 'is measured/estimated at' b

Examples:

Der var over 200 mennesker til stede
There were over 200 people present

Prisen faldt med 10%
The price dropped by 10%

Et kirketårn på tyve meter
A church tower of twenty metres

I en alder af seks år
At an age of six years

6.4.19 MNR MANNER

Subrelations:

SITUATION

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

MANNER

Relates two entities *a* and *b*, where *a* is an event that happens in the manner or in a situation as *b*.

MNR – SITUATION is closely related to the TMP relation.

E.g. 'Scarlett O'Haras liv under den amerikanske borgerkrig', where the relation can be characterized as MNR - SITUATION, or as TMP - DURATION, depending on whether we view the text as describing the situation under which she lives, namely a war, or the specific period of her life, namely during the war.

Paraphrases:

a '(foregår) på måden' *b*

En : *a* '(happens) in the manner' *b*

a '(foregår) under forholdet' *b*

a '(happens) under the circumstances' *b*

Examples:

Brug baldrian med varsomhed

Use valerian with caution

Han blev opereret under fuld bedøvelse

He underwent surgery under full anaesthesia

Julius drejede på uglevis hovedet

Lit: Julius turned on owl-like his head

Julius turned his head in an owl-like manner

6.4.20 MTH MATH

Subrelations:

FUNCTION

PLUS

MINUS

...

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

A mathematical relation between a number a and a number or a function b that is applied to a .

Paraphrases:

a 'som en funktion af' b

a 'as a function of' b

a 'plus' b

a 'plus' b

a 'minus' b

a 'minus' b

Examples:

en million i anden potens

A million to the power of 2

3 i trettende

3 to the thirteenth power

Tre til fire er syv

Lit: Three to four is seven

Three added to four is seven

syv fra ti er tre

Lit: Seven from ten is three

Seven subtracted from ten is three

6.4.21 PNT PATIENT

Subrelations:

PATIENT

THEME

RECIPIENT

BENEFICIARY

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

This general patient relation comprises all types of relations between an entity *a* that is an act or process, and an undergoer *b* (a patient, recipient, theme or beneficiary).

b may or may not undergo change as a result of the act or process.

Paraphrases:

a 'er rettet imod' *b*

a 'is directed at' *b*

a 'blandt' *b*

a 'amongst' *b*

a 'bliver modtaget af' *b*

a 'is recieved by' *b*

a 'er tiltænkt' *b*

a 'is intended for' *b*

Examples:

Det er skik hos indianerne

It is custom amongst the indians

En boltsaks skærer gennem blikket (som smør)

A bolt cutter cuts through the tin (like butter)

Han havde været vred på hende

He had been mad at her

Hvem stemmer for planen?

Who votes for the plan?

6.4.22 PRP PURPOSE

Relates two entities *a* and *b*, where *b* is the purpose of function of *a*.

Paraphrases:

a 'er beregnet til' *b*

a 'is intended for' *b*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a 'har et formål i forhold til' *b*
a 'has a purpose in relation to' *b*

a 'har en funktion i forhold til' *b*
a 'has a function in relation to' *b*

Examples:

Hjul til racerbiler
Wheels for racing cars

Værktøj til læderarbejde
Tools for leather work

Træer til skoven
Trees for the forest

Løbe efter hjælp
Run for help

6.4.23 QUA QUA

Relates two entities *a* and *b*, where *a* is an event that occurs by virtue of *b*.

Paraphrases:

a 'i henhold til' *b*
a 'in compliance with' *b*

a 'i kraft af' *b*
a 'by virtue of' *b*

a 'i egenskab af' *b*
a 'in the capacity of' *b*

Examples:

Han blev dømt efter loven
He was convicted under the law

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Hun blev sigtet som leder af foreningen
She was charged as head of the association

Han kunne have drevet det vidt med de talenter
He could have gone far with those talents

Alt er gået efter planen
Everything has gone according to the plan

6.4.24 SBT SUBSTITUTION

Relates two entities *a* and *b*, where *a* is a substitute for or executed on behalf of *b*

Paraphrases:

a 'i stedet for' *b*
a 'instead of' *b*

a 'på vegne af' *b*
a 'on behalf of' *b*

Examples:

Hun skrev under for direktøren.
She signed on behalf of the manager

Plante to træer for hvert der bliver fældet
Plant two trees for every one that is felled

6.4.25 HPR HYPERNYMY

This relation has the inverse relation HPO

Paraphrases:

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

a 'er hypernym for' *b*
a 'is a hypernym of' *b*

a 'er af typen' *b*
a 'is of the type' *b*

Examples:

Bæltedyret maa opføres under gumlerne
The armadillo must be entered under toothless mammals

6.4.26 HPO HYPONYMY

This relation has the inverse relation HPR

Paraphrases:

a 'er hyponym for' *b*
a 'is a hyponym of' *b*

a 'har undertypen' *b*
a 'has the subtype' *b*

Examples:

Følelse af sympati
A feeling of sympathy

6.4.27 SUP SUPERIORITY

Relates an event *a* with an entity *b* that is a superior rank or power and under whose superiority *a* takes place.

Paraphrases:

a '(sker) under styre af' *b*
a '(happens) under the rule of' *b*

Examples:

Han gjorde tjeneste under Eisenhower.

He served under Eisenhower

(...) *Jenny, der arbejder for direktør Bang.*
(...) Jenny, who worked for manager Bang

6.4.28 TMP TEMPORAL

Subrelations:

POINT (klokken x)

DUR

START (fra klokken x)

END (til klokken x)

PERIOD (I x timer)

DELAY (om x timer)

(DISTANCE) (x år efter y)

FREQ (hver x. time)

PER (i timen)

This general temporal relation comprises all types of temporal relations between an event *a* and an entity *b* that is a period or a point in time.
= (minutes, hours, days, years, ...)

Paraphrases:

a '(sker) på tidspunktet ' *b*

a '(happens) at the point in time' *b*

a '(sker) i tidsrummet ' *b*

a '(happens) during the period' *b*

a 'om ' *b*

a 'in' *b*

a 'hver' *b*

a 'every' *b*

a 'i' *b*

a 'a/an' *b*

Examples:

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

(Vi tales nok ved) engang efter Jul

(We will probably talk) some time after Christmas

Det har jeg vidst gennem mange år

That I have known for many years

Drabet skete mellem kl. 23.15 og 23.20

The murder took place between 11.15 pm and 11.20 pm

Hun er ansat fra torsdag

She is a member of staff as of Thursday

To gange i timen

Twice an hour

Han besøgte mig under min sygdom

He visited me during my illness

dit brev af 5. maj

Your letter of May 15

6.4.29 WRT WITH RESPECT TO

A weakly defined aboutness relation between an entity *a* and an entity *b*, where *a* is 'about' or 'relating to' *b*.

Paraphrases:

a 'med hensyn til' *b*

a 'with respect to' *b*

Examples:

Hun er god til fransk

She is good at French

Jeg er utilfreds med din opførsel

I am not satisfied with your behavior

Hvad betaler du i skat

How much do you pay in taxes

Bogen var lille af omfang

Lit: The book was small of size

The book was small

6.5 Investigations in Korpus 2000

The Danish language corpus Korpus 2000 is a freely available corpus which has been compiled by the Society for Danish Language and Literature (*Det Danske Sprog- og Litteraturselskab, DSL*) in order to document the use of the Danish language around the year 2000. Under the name KorpusDK, the corpus is searchable through a web interface³⁶. As illustrated in Figure 52, KorpusDK consists of two subcorpora, namely, Korpus 90 and Korpus 2000. DSL plans to add more recent texts in the future, so that the corpus will reflect the latest developments in the Danish language.

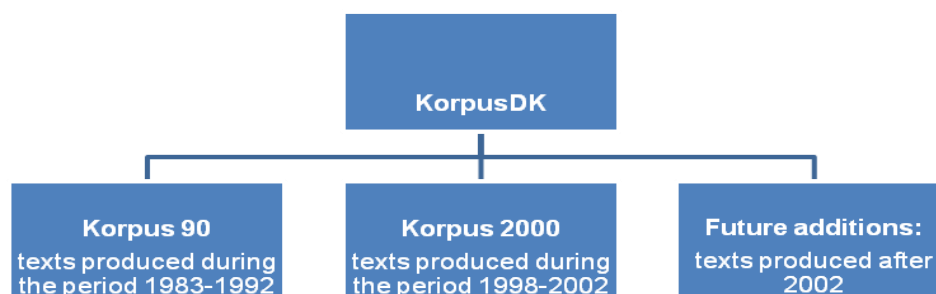


Figure 53 The contents of KorpusDK

In addition to providing search facilities through the web interface, DSL has made a selection of corpora available for download as so-called citation corpora - amongst others, Korpus 2000. A citation corpus consists of all the individual sentences from the original corpus, however, for copyright reasons, the order of the sentences has been scrambled so that reconstruction of the original texts is rendered impossible. The downloaded citation version of Korpus 2000 consists of 22.013.995 running words in 1.287.300

³⁶ <http://ordnet.dk/korpusdk>

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

sentences. The individual tokens in the citation corpus are annotated with information about part of speech, gender, number, definiteness, case etc. where applicable.

For our purposes, such a citation corpus meets our needs. We are not concerned with inter-sentential structures, but only intra-sentential structures.

6.5.1 Compiling a subcorpus

From the full citation corpus, we initially extract all sentences that match a grammar that has been constructed for these experiments. The grammar matches parts of a sentence that has the structure NP-PREP-NP, i.e. two noun phrases with a preposition in between. The preposition element matches only prepositions in the selected preposition set described above in section 6.6. The grammar is a shallow grammar, and it does not attempt to resolve PP attachment ambiguities.

As a result, we have a subcorpus where all sentences contain the lexical and syntactic structures we are looking for.

We aim at compiling a balanced data set of ~200 randomly chosen corpus-evidences per preposition. Since we have extracted the subcorpus from the citation version of Korpus 2000, we do not have to put any effort into scrambling of the corpus sentences - they are already in random order. However, we strive towards only extracting chunks where the prepositional phrase (PREP-NP) modify the first NP, and exclude chunks where this clearly not is the case. We extract the first 250 matching text chunks that seem to be NPs with a postmodifying PP per preposition. We extract an excess of 25% extra text chunks per preposition to accommodate for rejected text chunks due to incorrect matches or non-mappable NP heads to the ontology. The result is a balanced data set of ~3500 text chunks.

6.5.2 Annotation of Corpus-evidences

For our experiments, we need an annotated data set. The data set should consist of the features we wish to be able to include in the learning. In this case, we wish to include the following features:

- Preposition
- Relation

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

- Head of the first noun phrase
- Lemma for the head of first noun phrase
- Ontological type of the head of first noun phrase
- Head of the second noun phrase
- Lemma for the head of second noun phrase
- Ontological type of the head of second noun phrase

Part of this data set we can produce automatically. By use of the aforementioned grammar and the lemma information in the existing corpus annotation, we extract:

- Preposition
- Head of the first noun phrase
- Lemma for the head of first noun phrase
- Head of the second noun phrase
- Lemma for the head of second noun phrase

We thus lack information about ontological type for the heads of noun phrases and the relation denoted by the preposition.

6.5.3 Ontological Type Annotation

For the ontological type information, we use DanNet. In a local copy of DanNet in an extended form, we map all heads of noun phrases to a noun form in DanNet and, if found, we add information about the ontological type of the synset that the noun form belongs to the data set. If any of the two heads for a given chunk does not map to a noun form in DanNet, the chunk is discarded from the data set.

In our local copy of DanNet, we have added ~178,000 proper nouns (first names, middle names, last names, place names and names of companies) as instances of appropriate synsets. This allows us to annotate person names, cities, countries, etc. to appropriate ontological types to a far larger degree than otherwise possible³⁷.

³⁷ In DanNet version 1.1 of July 1, 2009, 250 proper names of countries and major geographical areas have been added as instances of nouns. (DanNet, 2010)

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Ontological types in DanNet are compound types, as described in section 3.3.3. This means, that for every lemma that is annotated with an ontological type, this type may be complex as e.g. in (65), where the lemma *hest* is annotated with the ontological types Animal and Object.

(65) *hest*,Animal+Object

In order to learn on the basis of such complex ontological types, there are a couple of approaches we can take: Either, we can add the complex ontological type as a single attribute, or we can split up the ontological types into simplex types. If we choose the former solution, we cannot identify subparts of the type as being identical. For example, we cannot learn that *hest* (horse) in (65) *bord* (table) in (66) share the ontological type Object.

(66) *bord*,Furniture;Artifact;Object

If we choose the latter solution, we can either add all possible ontological types as attributes, and specify whether the lemma in question is of this type, i.e. we have the possible values (0,1). The disadvantage of this approach is that we get a large number of attributes³⁸, which requires a much larger data set than the one we have in order to generalize well. We thus discard this solution. As an alternative, we can duplicate the line in the data set as many times as we need in order to get all possible combinations of the ontological types for the two lemmas for a given line in the data set, i.e. produce the cartesian product of the two sets of simplex ontological types for the two lemmas.

Example (67) below shows a version of the corpus evidence *En hest fra Amerika* (*A horse from America*) that has been annotated with lemma and complex ontological type. The semantic relation information is still to be added.

(67) *fra*,*hest*,Animal+Object,*amerika*,Place+Object

³⁸ The number of attributes is potentially equal to the number of distinct simplex ontological types in DanNet, namely 60. (191 distinct compound types)

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Given a data set line as in (67) with compound types each consisting of two ontological types, we get 2*2 lines with simplex types as shown in examples (68) – (71):

(68) fra,,hest,Animal,amerika,Place

(69) fra,,hest,Animal,amerika,Object

(70) fra,,hest,Object,amerika,Place

(71) fra,,hest,Object,amerika,Object

A possible drawback of this approach is that we may infer too many rules. As an example of this, given the ontological type pairs in examples (68) – (71) above, we may infer four rules instead of just one. In consequence, we choose to experiment with two data sets, one with complex ontological types, and one that has been split into simplex ontological types.

For word forms that are homonymous or polysemous and thus belong to more than one synset, we choose to annotate the word according to just one sense. The choice of sense is based on a heuristics based on empirical experience in a pilot annotation. The ranking prefers the most extensive ontological type to the less extensive. By this heuristics, if a given word form is part of two synsets, and one synset is e.g. BUILDING+OBJECT+PART, and the other is BUILDING+OBJECT, then the word form is annotated with the most extensive ontological type, namely BUILDING+OBJECT+PART.

6.5.4 Semantic Relation Annotation

The semantic relations denoted by the prepositions must be added manually. The chunks that have had ontological type information added to both heads of noun phrases are now subject to a manual assessment and following annotation with a relation from the relation set described above in section 6.4. The frequency distribution for the relations in the final data set is shown in Figure 53.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

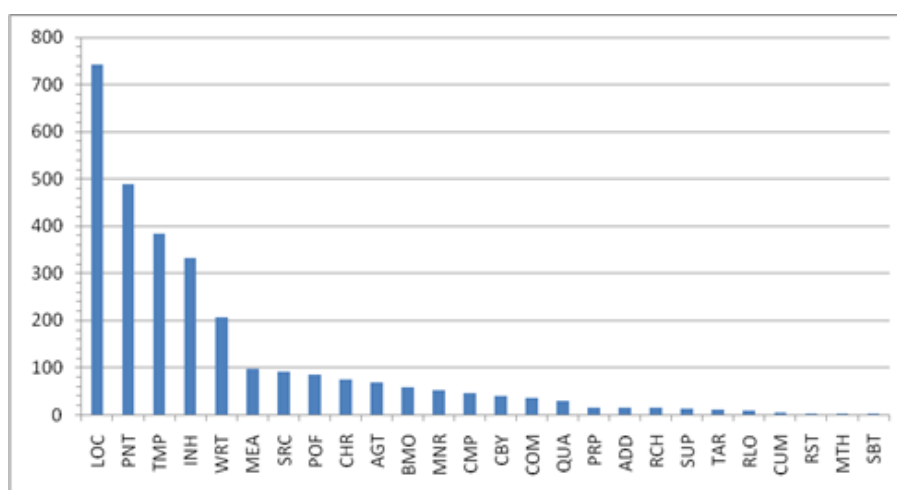


Figure 54 Frequency distribution for relations in the data set

Not all the relations that were the result of the analysis of dictionary entries, as described in section 6.4, were used in the annotation. 29 relations were identified in the dictionary analysis and, of these, 26 were used in the annotation process. The three relations that were not used in the annotation are CAU, HPR and HPO. The matrix in Table 19 shows:

- * Relations that are in the original relation set for a given preposition, but not used in the annotation (marked ○)
- * Relations that are in the relation set and are used in the markup for chunks containing the given preposition (marked ●)
- * Relations that were not part of the relation set for the given preposition, but were used in the annotation (marked ⊕).

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

	Af	efter	for	fra	gennem	hos	i	med	mellem	over	på	til	under	ved
ADD		●	●							○	○	○		
AGT	●			⊕		●	⊕	⊕				○	○	●
BMO	○		○	○	●		●	●		●	●	○	○	●
CAU													○	
CBY	●	⊕	○	●	⊕		○	●		●	○		○	●
CHR	●		○				●	●				○	○	○
CMP	●		⊕				○	●	○		●			
COM	⊕	●	○				○	○		●	○	○	●	⊕
CUM								●				○		
INH	●	⊕	●	⊕		●	●	⊕	⊕	●	●	●		●
LOC	●	○	●	●	●	●	●	○	●	●	●	●	●	●
MEA	●		●	●			○	●	⊕	●	●	●	●	●
MNR								●			●		●	⊕
MTH		⊕		○			○					○		
PNT	●	●	●	○	●	●	●	●	●	●	●	●	⊕	●
POF	●			⊕		⊕	●	○	○		●		●	⊕
PRP		○	○				○					●		
QUA		●	●					○					●	
RCH	●					●				●		○	○	●
RLO	⊕							●						
RST							○					●	○	
SBT			○											
SRC	●	●		●	●	●								
HPR	○													
HPO													○	
SUP													●	
TAR		●					○					⊕		
TMP	●	●	●	●	●		●	○	●	●	●	●	●	●
WRT	●		●	●		●	●	●	⊕	●	●	●		●

Table 19 ○ In relation set from the dictionary analysis, but not used in the annotation ● In relation set, and used in the annotation ⊕ Not in relation set, but used in the annotation

The fact that not all relations that are identified in the dictionary analysis for a given preposition are used in the annotation should not necessarily be regarded as an indication that they are not frequently denoted by the preposition in question. Possible reasons for this difference are to be found in the fact that we investigate prepositional senses in a restricted syntactic form, namely primarily nouns plus modifying PPs. Other senses may be

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

dominant for other syntactic forms. For example, it is probable that we would see a different set of applied relations if we had analyzed verbs plus modifying PPs. However, the fact that we identify relations that are not part of the initial set for a given preposition is interesting. We can now add new senses that we have not previously identified to a number of prepositions.

6.6 Prepositions and the Relations they Denote

Below, we describe each of the 14 prepositions by the relations they can denote. The description includes the relations that were identified in the dictionary analysis as well as in the annotation process.

6.6.1 Af

For the preposition *af*, we have identified the following relations in the corpus: AGT, CBY, CHR, CMP, COM, INH, LOC, MEA, PNT, POF, RCH, RLO, SRC, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 54. Below, we give an example for each relation denoted by *af*.

AGT	<i>De tidlige sange af Alban Berg</i> The early songs by Alban Berg
CBY	<i>syge af Salmonella</i> ill from Salmonella
CHR	<i>modeller af typen A</i> models of type A
CMP	<i>flokke af svaner</i> flocks of swans
COM	<i>i skikkelse af høje, unge kvinder</i> in the guise of tall, young women
INH	<i>æresmedlem af Dansk Brygmester Forening</i> honorary member of the Danish Master Brewer's Association
LOC	<i>udkanten af Paris</i>

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

the outskirts of Paris

MEA *den beskedne sum af 30.000 kroner*
the minor sum of 30,000 kroner

PNT *bearbejdning af nye, ukendte input*
processing of new, unknown input

POF *en filial af Danske Bank*
a branch of Danske Bank

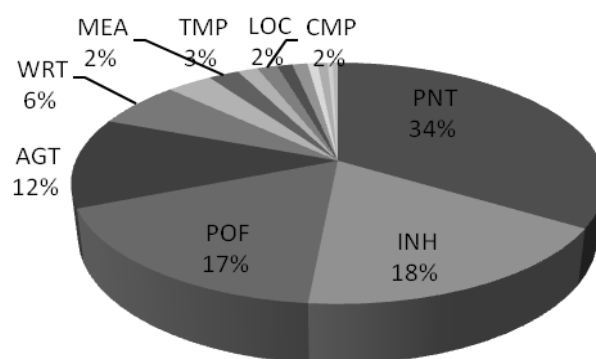
RCH *en djævelinde af en kone*
a she-devil of a wife

RLO *et bæger af syre*
a cup of acid

SRC *resultatet af forhandlingerne*
the result of the negotiations

TMP *udgangen af det 20. århundrede*
the end of the 20th century

WRT *lørdagens udgave af JyllandsPosten*
Saturday's edition of JyllandsPosten



Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Figure 55 Frequency distribution for the relations denoted by *af*

6.6.2 Efter

For the preposition *efter*, we have identified the following relations in the corpus: ADD, CBY, COM, INH, MTH, PNT, QUA, SRC, TAR and TMP. The frequency distribution for the relations is illustrated in Figure 55. Below, we give an example for each relation denoted by *efter*.

ADD	<i>den ene celle efter den anden</i> one cell after another
CBY	<i>graviditet efter et ubeskyttet samleje</i> pregnancy after unprotected intercourse
COM	<i>en kamp efter vestligt forbillede</i> a fight by western example
INH	<i>enke efter arkitekt Preben Hansen</i> widow of architect Preben Hansen
MTH	<i>boligudgiften efter fradrag</i> the housing expenses after deductions
PNT	<i>Den øgede interesse efter den ægte vare</i> The increased interest in the real McCoy
QUA	<i>erstatning efter de gældende regler</i> compensation under the current rules
SRC	<i>arven efter Nielsen</i> the inheritance from Nielsen
TAR	<i>det uendelige begær efter penge</i> the perpetual desire for money
TMP	<i>et stykke tid efter VM</i> some time after the world championships

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

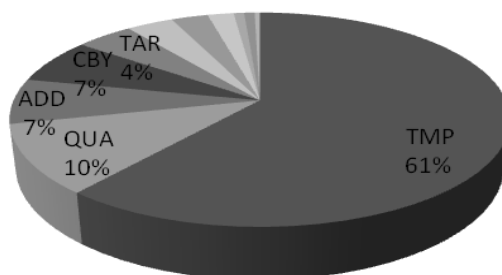


Figure 56 Frequency distribution for the relations denoted by *efter*

6.6.3 For

For the preposition *for*, we have identified the following relations in the corpus: ADD, CMP, INH, LOC, MEA, PNT, QUA, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 57. Below, we give an example for each relation denoted by *for*.

ADD *dag for dag*
day by day

CMP *en nystartet bogklub for sygeplejersker*
a newly founded book club for nurses

INH *formand for Aarhus sejlkub*
chairman of Aarhus yacht club

LOC *nord for Viborg*
north of Viborg

MEA *en 28 dages kur for 200 kr*
a 28-day cure for 200 kr

PNT *dansk lobbyisme for Baltikum*
Danish lobbying for the Baltic states

QUA *Belønning for hårdt arbejde*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Reward for hard labour

TMP *dronning for en dag*
queen for a day

WRT *Landsforeningen for Autisme*
The national association for autism

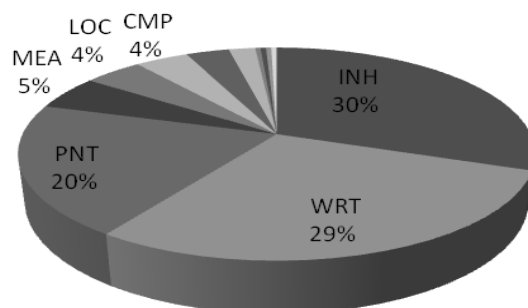


Figure 57 Frequency distribution for the relations denoted by *for*

6.6.4 Fra

For the preposition *fra*, we have identified the following relations in the corpus: AGT, CBY, INH, LOC, MEA, POF, SRC, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 58. Below, we give a corpus example of each relation denoted by *fra*.

AGT *et flot sololøb fra Mads*
a beautiful solo run by Mads

CBY *det dunkle skær fra bålet*
the dim glow of the bonfire

INH *en repræsentant fra Novo*
a representative from Novo

LOC *to forskere fra Aalborg*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

two researchers from Aalborg

- MEA *enhver distance fra 1500 m*
every distance from 1500 m
- POF *blade fra en busk*
leaves from a bush
- SRC *opbakningen fra den borgerlige gruppe*
the support from the right-wing parties
- TMP *dansk kunst fra det 18. århundrede*
Danish art from the 18th century
- WRT *forskellene fra Bush*
the differences from Bush

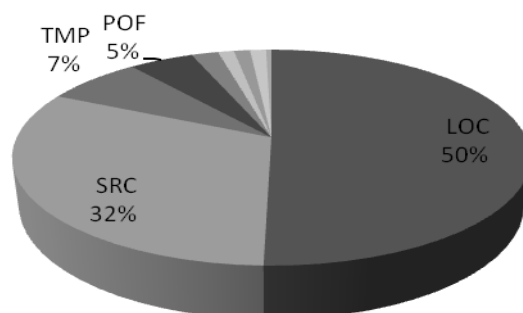


Figure 58 Frequency distribution for the relations denoted by *fra*

6.6.5 Gennem/igennem

The preposition *gennem* is treated alongside *igennem* in several of the consulted dictionaries. We agree with these sources, and consider the two forms as being synonymous. For this reason, the compiled subcorpus contains both forms. For the preposition *gennem/igennem*, we have identified the following relations in the corpus: BMO, CBY, LOC, PNT, SRC and TMP. The frequency distribution for the relations is illustrated in

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Figure 59. Below, we give a corpus example of each relation denoted by *gennem/igennem*.

- BMO *kvælstof gennem kunstgødning*
nitrogen through fertilizers
- CBY *leukæmi gennem stråling*
leukemia through radiation
- LOC *Boringer gennem Indlandsisen*
Drillings through the ice cap
- PNT *en vej gennem systemet*
a path through the system
- SRC *oplysninger gennem Interpol*
information through Interpol
- TMP *adskillige forhøjelser gennem året*
numerous increases throughout the year

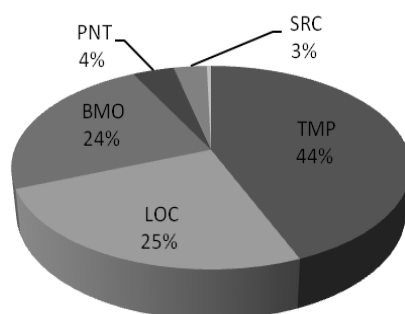


Figure 59 Frequency distribution for the relations denoted by *gennem*

6.6.6 Hos

For the preposition *hos*, we have identified the following relations in the corpus: AGT, INH, LOC, PNT, POF, RCH, SRC and WRT. The frequency distribution for the relations is illustrated in Figure 60. Below, we give a corpus example of each relation denoted by *hos*.

Uncovering of the Semantic Relations Denoted by a Selection of Danish
Prepositions

AGT	<i>en negativ reaktion hos tilhørerne</i> a negative reaction with the listeners
INH	<i>behandlinger hos en fysioterapeut</i> treatment at a physical therapist's
LOC	<i>En læreplads hos den internationalt berømte Georg Jensen</i> An apprenticeship with the internationally renowned Georg Jensen
PNT	<i>Forsinket sårheling hos de 3 førstnævnte patientgrupper</i> Delayed wound healing with the 3 first mentioned patient groups
POF	<i>en afdeling hos Warner</i> a division of Warner
RCH	<i>den øgede økologiske bevidsthed hos forbrugerne</i> the increased ecological awareness of the consumers
SRC	<i>trøst hos andre mænd</i> consolation in other men
WRT	<i>den ansattes fremtidige stilling hos arbejdsgiveren</i> the future position of the employee with the employer

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

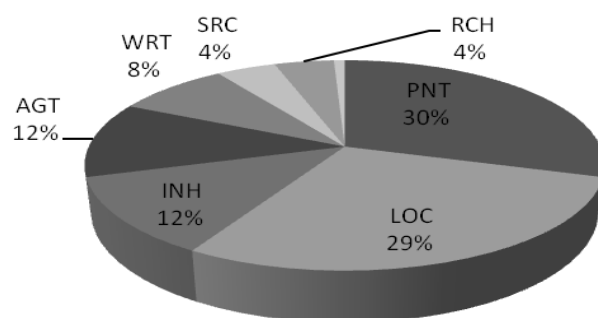


Figure 60 Frequency distribution for the relations denoted by *hos*

6.6.7 I

For the preposition *i*, we have identified the following relations in the corpus: AGT, BMO, CHR, INH, LOC, PNT, POF, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 61. Below, we give a corpus example of each relation denoted by *i*.

AGT *Forslagets første behandling i Folketinget*
the proposal's first reading in the Parliament

BMO *oplysningerne i Det Fri Aktuelt*
the information in Det Fri Aktuelt (newspaper)

CHR *hans sidste runde i 69 slag*
his last round in 69 strokes

INH *afdelingschef i juridisk afdeling*
head of department in the legal department

LOC *afdelingskontorerne i Skanderborg*
department offices in Skanderborg

PNT *en afdæmpet vækst i den amerikanske økonomi*
weak financial growth in the US economy

POF *medlemsstater i Den Europæiske Union*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

member states in the European Union

TMP *affæren i efteråret*
the affair in the fall

WRT *En nagende mistanke i det nye forhold*
A nagging suspicion in the new relationship

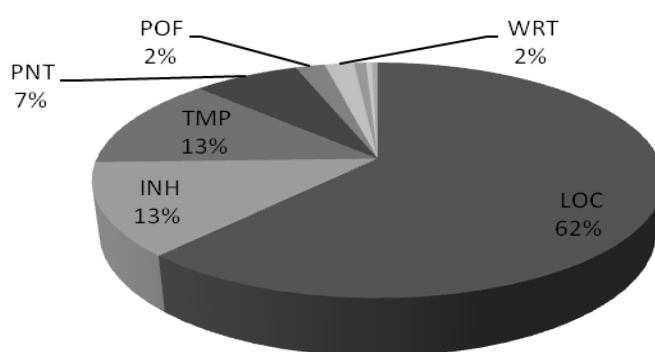


Figure 61 Frequency distribution for the relations denoted by *i*

6.6.8 Med

For the preposition *med*, we have identified the following relations in the corpus: AGT, BMO, CBY, CHR, CMP, CUM, INH, MEA, MNR, PNT, RLO and WRT. The frequency distribution for the relations is illustrated in Figure 62. Below, we give a corpus example of each relation denoted by *med*.

AGT *pigtrådsmusik med The Donkeys*
beat music with The Donkeys

BMO *en flot fejende bevægelse med hånden*
a grand sweeping gesture with the hand

CBY *i sengen med feber*
in bed with a fever

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

CHR	<i>personer med nedsat arbejdsevne</i> individuals with reduced work capacity
CMP	<i>De fleste biler med 15 tommers fælge</i> Most cars with 15" wheel rims
CUM	<i>En mor med en lille dreng</i> A mother with a small boy
INH	<i>kontrol med levende dyr</i> control with live stock
MEA	<i>største procentvise fald med knap 20%</i> largest percentage-wise decrease of barely 20%
MNR	<i>deres liv med andre børn</i> their lives with other children
PNT	<i>telefonsamtaler med Miguel</i> phone conversations with Miguel
RLO	<i>biler med tunesiske fans</i> cars with Tunesian fans
WRT	<i>sager med somaliske ansøgere</i> cases of Somali applicants

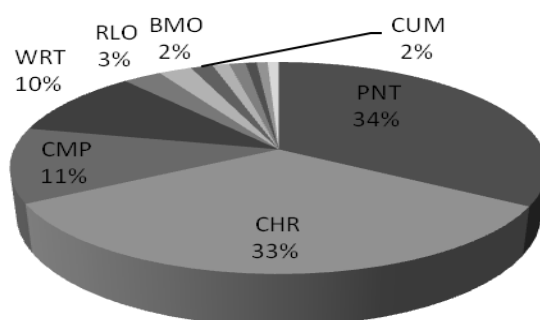


Figure 62 Frequency distribution for the relations denoted by *med*

6.6.9 Mellem/imellem

The preposition *mellem* is treated alongside *imellem* in several of the consulted dictionaries. We agree with these sources, and consider the two forms as being synonymous, as was also the case with the preposition *gennem/igennem*. For this reason, the compiled subcorpus contains both forms. For the preposition *mellem/imellem*, we have identified the following relations in the corpus: INH, LOC, MEA, PNT, SRC, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 63. Below, we give a corpus example of each relation denoted by *mellem/imellem*.

- INH *relationerne mellem de øvrige EU-lande*
the relationship between the other EU countries
- LOC *vejen mellem Holbæk og Sjællands Odde*
the road between Holbæk and Sjællands Odde
- MEA *alle børn mellem to uger og et år*
all children aged between two weeks and one year
- PNT *samværet mellem børn og forældre*
the contact between children and parents
- TMP *natten mellem lørdag og søndag*
the night between Saturday and Sunday
- WRT *de store forskelle mellem de forskellige parceller*
the large difference between the individual plots

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

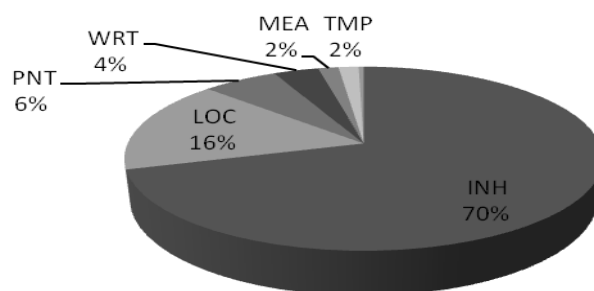


Figure 63 Frequency distribution for the relations denoted by *mellem*

6.6.10 Over

For the preposition *over*, we have identified the following relations in the corpus: BMO, CBY, COM, INH, LOC, MEA, PNT, RCH, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 64. Below, we give a corpus example of each relation denoted by *over*.

BMO *bedre kommunikation over Internettet*
better communication over the Internet

CBY *sorg over det store antal dræbte*
grief over the large number of dead

COM *et pænt stykke over den bogførte værdi*
well above the book value

INH *et monument over to brødres skæbnesvangre samlermani*
a monument to the two brothers' disastrous mania for collecting

LOC *udsigt over hele Kattegat*
a view over all of the Kattegat

MEA *danskere over 80 år*
Danish citizens above 80 years of age

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

- PNT *en detaljeret liste over projekter*
a detailed list of projects
- RCH *stil over moden*
Lit: style over the trend
'that trend has style'
- TMP *24 millioner kroner over en fireårig periode*
24 million kroner over a period of four years
- WRT *et kort over London*
a map of London

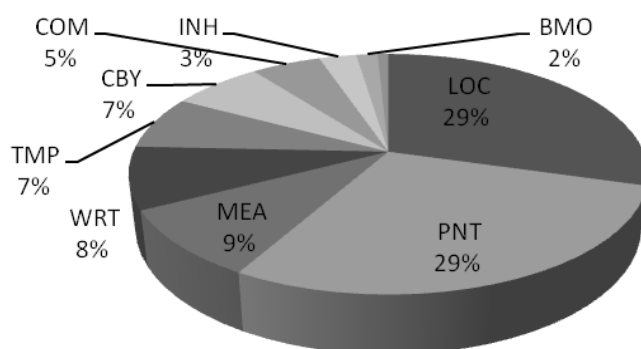


Figure 64 Frequency distribution for the relations denoted by *over*

6.6.11 På

For the preposition *på*, we have identified the following relations in the corpus: BMO, CMP, INH, LOC, MEA, MNR, PNT, POF, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 64. Below, we give a corpus example of each relation denoted by *på*.

- BMO *handel med varer på computeren*
trade in goods on the computer
- CMP *Front Nationals gruppe på 11 medlemmer*
Front National's group of 11 members

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

INH	<i>chefredaktør på Der Spiegel</i> chief editor of Der Spiegel
LOC	<i>fem medarbejdere på 13. sal</i> five employees on the 13 th floor
MEA	<i>en lønforhøjelse på 100 kr</i> a pay increase of 100 kroner
MNR	<i>blikkenslagere på akkord</i> plumbers on piecework
PNT	<i>stor indflydelse på blodets indhold af stoffet homocystein</i> a great influence on the homocysteine levels in the blood
POF	<i>stroppen på badedragten</i> the strap on the bathing suit
TMP	<i>den mest romantiske dag på hele året</i> the most romantic day of the whole year
WRT	<i>valgmulighederne på andre områder</i> the options in other areas

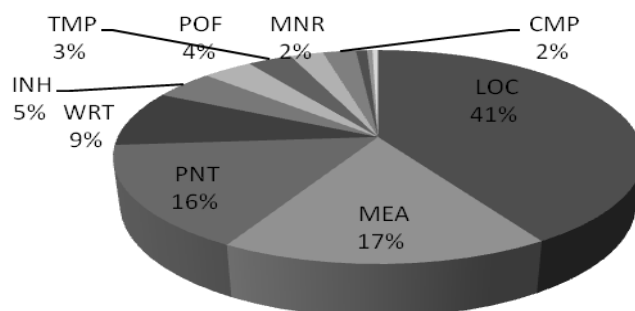


Figure 65 Frequency distribution for the relations denoted by på

6.6.12 Til

For the preposition *til*, we have identified the following relations in the corpus: INH, LOC, MEA, PNT, PRP, RST, TAR, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 66. Below, we give a corpus example of each relation denoted by *til*.

INH	<i>nedtællingen til et historisk vendepunkt</i> the countdown to a historic turning point
LOC	<i>dyre rejser til New York</i> expensive travels to New York
MEA	<i>verdens dyreste frimærke til 13 millioner kroner</i> the world's most expensive stamp at 13 million kroner
PNT	<i>Forfatteren til bogen</i> The author of the book
PRP	<i>midler til både olie og lønninger</i> means for oil as well as salaries
RST	<i>Isminurs forvandling til Lone</i> Isminur's transformation into Lone
TAR	<i>uddannelsen til speciallæge</i> education as specialist doctor
TMP	<i>11. marts til 10. April</i> March 11 th to April 10 th
WRT	<i>tid til en gåtur</i> time for a walk

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

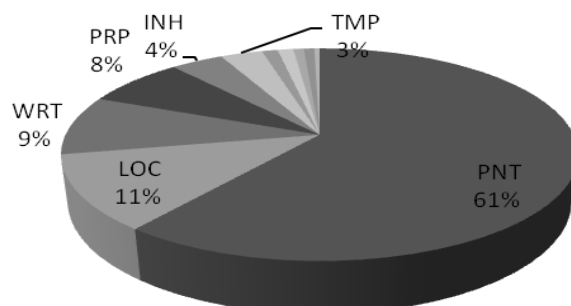


Figure 66 Frequency distribution for the relations denoted by *til*

6.6.13 Under

For the preposition *under*, we have identified the following relations in the corpus: COM, LOC, MEA, MNR, PNT, POF, QUA, SUP and TMP. The frequency distribution for the relations is illustrated in Figure 67. Below, we give a corpus example of each relation denoted by *under*.

COM *et enkelt slag under par*
a single stroke under par

LOC *det lille bord under vinduet*
the small table below the window

MEA *fonde med formuer under fem millioner*
foundations with less than five million in assets

MNR *vin under kontrollerede former*
wine under controlled circumstances

PNT *vel meget fut under økonomien*
'too much fire under the economy'

POF *Et udvalg under justitsministeriet*
A committee under the Ministry of Justice

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

QUA *sin pligt under sædvaneretten*
one's duty under common law

SUP *et Europa under tysk herredømme*
a Europe under German supremacy

TMP *løn under barsel*
salary during maternity leave

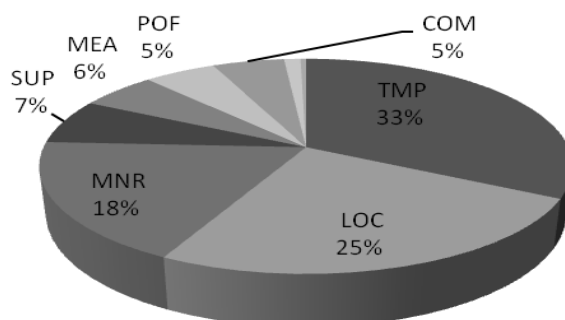


Figure 67 Frequency distribution for the relations denoted by *under*

6.6.14 Ved

For the preposition *ved*, we have identified the following relations in the corpus: AGT, BMO, CBY, COM, INH, LOC, MEA, MNR, PNT, POF, RCH, SRC, TMP and WRT. The frequency distribution for the relations is illustrated in Figure 68. Below, we give a corpus example of each relation denoted by *ved*.

AGT *en helt ny oversættelse ved mag. art. Ole Vesterholt*
a brand new translation by MA Ole Vesterholt

BMO *formering ved tilfældig knopskydning*
reproduction by spontaneous gemmation

CBY *skade ved branden*

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

damage from the fire

COM *en gammel græker ved navn Iannis*
an old **Greek** by the name of Iannis

INH *deltagelsen ved OL*
participation in the Olympics

LOC *en dame ved skranken*
a lady at the counter

MEA *en forvarmet ovn ved 225 grade*
a pre-heated oven at 225 degrees

MNR *pæne ord ved festlige lejligheder*
nice words at festive occasions

PNT *start ved et stævne*
start at a race

POF *kvartfinalerne ved VM*
the quarter finals at the world championships

RCH *den gode stemning ved Gaimars hof*
the good atmosphere at Gaimar's court

TMP *overraskelsen ved søndagens delstatsvalg*
the surprise at Sunday's federal state elections

WRT *flere ulemper ved en model*
several drawbacks of the model

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

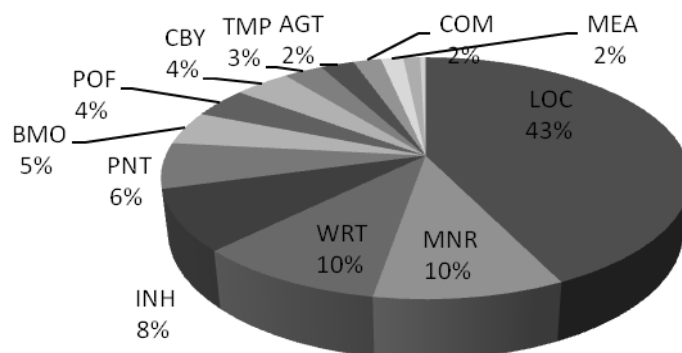


Figure 68 Frequency distribution for the relations denoted by *ved*

6.7 Rules

The annotated corpus can now serve as a gold standard to which we can compare the results of various rule sets that we can infer from the data set. The first part of this section is concerned with human evaluation of the data set, resulting in frequency-based rules. The second part is concerned with applying machine learning to the data set in order to produce rules that include different combinations of features from the annotation.

Our aim is to produce a rule-based dictionary of prepositions that can be applied to text in order to disambiguate preposition senses.

6.7.1 Frequency-based Rule Deduction

The probability for a correct annotation based purely on chance is $1/26 = 0.038$, since we have a relation set of 26 elements. This corresponds to an expected precision of 3.8%.

The simplest rule that we can deduce is a trivial rejector equal to the overall most frequent relation in the annotated corpus. In the annotated corpus, the relations are distributed as shown in Table 20. If we apply the overall most frequent relation, LOC, to all instances, we can achieve a precision of 25.40%. This improvement of the precision score from 3.8% to 25.40% is significant at a confidence level of .95.

We are able to improve this result by not simply applying the overall most frequent relation to all instances regardless of the preposition, but applying the most frequent relation for a given preposition to all instances of this

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

preposition. This approach is equal to applying a projected classifier for the simplest feature space consisting of only prepositions. Figure 68 shows the precision scores for the most frequent relation per preposition.

Relation	Frequency	Percentage of all instances
LOC	743	25,40%
PNT	489	16,72%
TMP	384	13,13%
INH	333	11,38%
WRT	207	7,08%
MEA	97	3,32%
SRC	91	3,11%
POF	86	2,94%
CHR	74	2,53%
AGT	69	2,36%
BMO	58	1,98%
MNR	53	1,81%
CMP	46	1,57%
CBY	40	1,37%
COM	36	1,23%
QUA	30	1,03%
ADD	15	0,51%
PRP	15	0,51%
RCH	15	0,51%
SUP	14	0,48%
TAR	12	0,41%
RLO	8	0,27%
CUM	4	0,14%
RST	3	0,10%
MTH	2	0,07%
SBT	1	0,03%

Table 20 Relations, frequencies and percentage of all relations

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

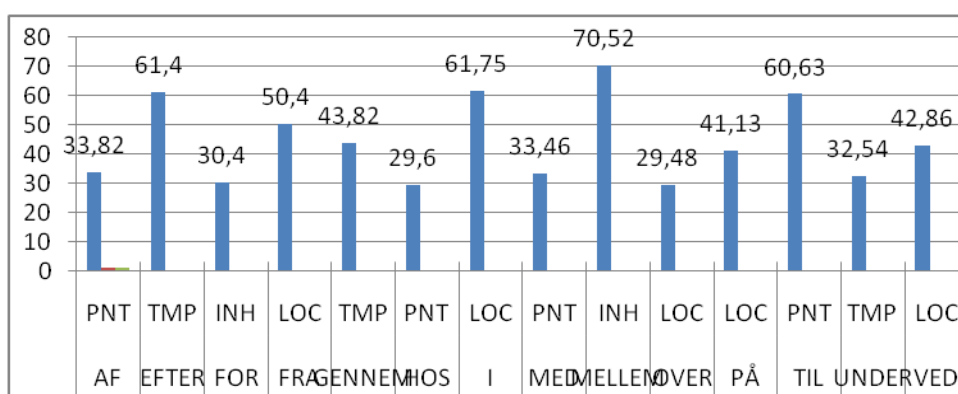


Figure 69 Precision scores for the most frequent relation per preposition

As shown in Figure 69, all prepositions occur with a near identical frequency in the annotated corpus, which means that the overall precision if we apply the most frequent relation per preposition is close to equal to the arithmetic average of all individual precision scores per preposition, namely ca. 45%, as shown in Table 21.

However, in free text, prepositions are not equally frequently occurring. Figure 70 shows the absolute frequencies in the citation version of Korpus 2000 for the 14 studied prepositions as well as an accumulated frequency for all other tokens in Korpus 2000 that are tagged as prepositions. In the citation version of Korpus 2000, 2,911,511 tokens are tagged as prepositions, and of these 2,386,444 or 82%, are instances of the 14 selected prepositions.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

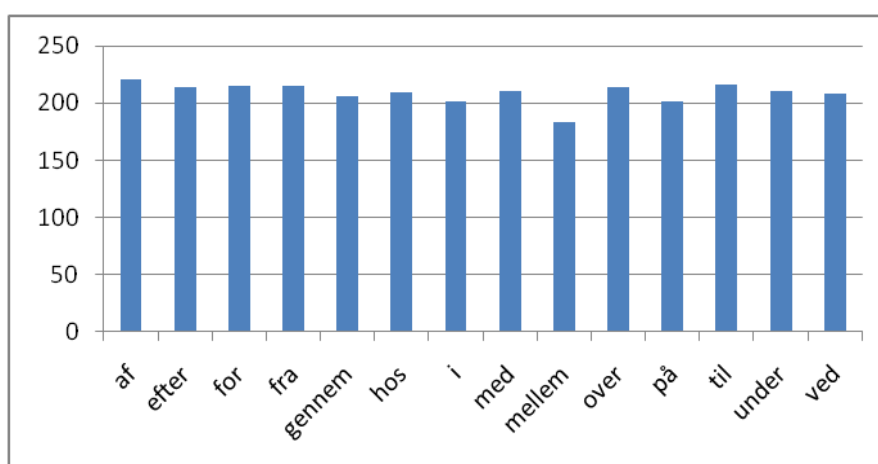


Figure 70 Frequency distribution for the 14 prepositions in the data set

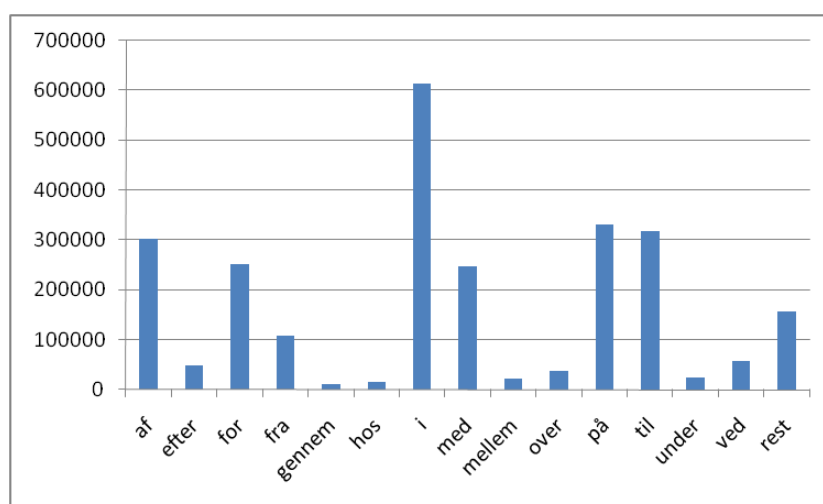


Figure 71 Frequency distribution for the 14 prepositions in the citation version of Korpus 2000. The column *rest* shows the accumulated frequency for all tokens in Korpus 2000 that are tagged as prepositions.

If we take these figures into account and produce a weighted average, we can predict an overall precision of 46.97% for a most-frequent-per-preposition rule when applied to free text.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Preposition	Most frequent relation	Precision for the most frequent relation	Percentage of instances in Korpus 2000
af	PNT	33,82%	12,70%
efter	TMP	61,40%	2,01%
for	INH	30,40%	10,55%
fra	LOC	50,40%	4,54%
gennem	TMP	43,82%	0,51%
hos	PNT	29,60%	0,62%
i	LOC	61,75%	25,67%
med	PNT	33,46%	10,35%
mellem	INH	70,52%	0,91%
over	LOC	29,48%	1,58%
på	LOC	41,13%	13,85%
til	PNT	60,63%	13,31%
under	TMP	32,54%	1,02%
ved	LOC	42,86%	2,40%
Arithmetic average		44,42%	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Weighted average		46,97%	$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$

Table 21 Precision scores for the most frequent relation per preposition with arithmetic and weighted average

6.7.2 Learning Rules with WEKA

In order to learn rules that are based not only on frequency, but also consider complex feature combinations, we now choose to apply machine learning algorithms to the data set. As for the experiments described in Chapter 5, the implementation of the algorithms that we used was the WEKA software package (Witten & Frank, 2005).

Since the aim of our experiments is to come up with a set of rules that can be implemented as well as analyzed by humans in order to gain knowledge about the preference of a range of prepositions and the relations denoted by these for the ontological types of their arguments, we choose to apply rule-learning algorithms to our data set.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Apart from the a priori decision to apply rule-producing algorithms, the choice of the two specific algorithms described below was based on a pilot application of all rule-producing algorithms implemented in WEKA to the data set with ontological types and prepositions in the feature space. The two chosen algorithms were the ones that, for this specific feature space, performed best (i.e. yielded the highest precision) of the rule learning algorithms implemented in WEKA. As a result, we will use two classifiers for these experiments; the JRip classifier and the PART classifier:

- JRip: The JRip algorithm implements the RIPPER propositional rule learner (Cohen, 1995), which induces an initial (large) rule set, and then increases the accuracy of the whole rule set by revising individual rules. (cf. section 5.7)
- PART: The PART algorithm (Eibe & Witten, 1998) learns one rule at a time by repeatedly creating partial decision trees, and generating rules from them.

The experiments were performed using 10-fold cross-validation, and performed on four different combinations of the feature space, ranging from only using the preposition to using the whole feature space (i.e. lemmatized NP heads, ontological types of the NP heads and preposition). The results of these experiments are shown in Table 22 and Table 23 below. Table 22 shows the percentages of correctly classified instances for the output of the JRip and PART algorithms with the split data set as input. Table 23 shows the percentage of correctly classified instances with the non-split data set as input. The non-split data set contains complex ontological types, and the split data set contains simple ontological types. The ontological type annotation is explained in more detail above in section 6.5.3.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Feature space	JRip-split	K-score	PART-split	K-score
1 Preposition	36.96	0.19	45.43	0.33
2 Ontological types (firstonto+secondonto)	27.84	0.05	38.00	0.24
3 Preposition and Ontological types (firstonto+secondonto)	49.57	0.37	59.50	0.52
4 All: Ontological types (firstonto+secondonto), Preposition and Lemma (firsthead+secondhead)	85.09	0.83	error	n/a

Table 22 The percentage of correctly classified instances and K-score for the output of the JRip and PART algorithms on four different combinations of input features with the split data set as input.

Feature space	JRip	K-score	PART	K-score
0 Chance 3.8%				
1 Preposition	36.14	0.17	43.45	0.31
2 Ontological types (firstonto+secondonto)	35.42	0.16	42.09	0.31
3 Preposition and Ontological types (firstonto+secondonto)	50.80	0.39	53.13	0.45
4 All: Ontological types (firstonto+secondonto), Preposition and Lemma (firsthead+secondhead)	52.85	0.41	51.32	0.40

Table 23 The percentage of correctly classified instances and K-score for the output of the JRip and PART algorithms on four different combinations of input features with the non-split data set as input.

The Kappa Statistics score (κ), or Kappa coefficient, is a measure for annotation agreement that is often used in connection with scoring of inter-annotator agreement in tagging tasks, as described in (Di Eugenio & Glass, 2004). The K-score is output by WEKA as part of the statistics information pertaining to a given experiment. The score gives an indication as to what

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

degree the agreement between the annotation by way of the rules produced by the learning algorithm and the annotation of the data set can be achieved by chance.

K is computed as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \cdot P(A)$$

Where $P(A)$ is the observed agreement among the annotations, and $P(E)$ is the probability that the annotators agree by chance. The values of K are constrained to the interval $[-1, 1]$, where $K = 1$ means that there is perfect agreement, $K = 0$ means that agreement is equal to chance, and $K = -1$ means 'perfect disagreement'. Thus, for a K-score that is greater than 0, we can assume that the result is not entirely achievable by chance. The closer to 1, the more reliable is the result. For the results of the experiments shown in Table 22 and Table 23, all have a K-score between 0 and 1.

Uncovering of the Semantic Relations Denoted by a Selection of Danish
Prepositions

Algorithm	Corpus	No of Instances	Feature spaces	Precision scores	Improvement (% points)	Significant?
JRip	Split	20046	Prep → prep+onto	36.96 - 49.57	12.61	yes
JRip	Split	20046	Prep → prep+onto+lemma	36.96 - 85.09	48.13	yes
JRip	Split	20046	Onto → prep+onto	27.84 - 49.57	21.73	yes
JRip	Split	20046	Onto → prep+onto+lemma	27.84 - 85.09	57.25	yes
JRip	Original	2925	Prep → prep+onto	36.14 - 50.80	14.66	yes
JRip	Original	2925	Prep → prep+onto+lemma	43.45 - 51.32	7.87	yes
JRip	Original	2925	Onto → prep+onto	35.42 - 50.80	15.38	yes
JRip	Original	2925	Onto → prep+onto+lemma	35.42 - 52.85	17.43	yes
PART	Split	20046	Prep → prep+onto	45.43 - 59.50	14.07	yes
PART	Split	20046	Onto → prep+onto	38.00 - 59.50	21.05	yes
PART	Original	2925	Prep → prep+onto	43.45 - 53.13	9.68	yes
PART	Original	2925	Prep → prep+onto+lemma	43.45 - 51.32	7.87	yes
PART	Original	2925	Onto → prep+onto	42.09 - 53.13	11.04	yes
PART	Original	2925	Onto → prep+onto+lemma	42.09 - 51.32	9.23	yes

Table 24 Improvement in precision for rules produced with increasingly large feature spaces

The probability for a correct annotation based purely on chance without any rules applied is $1/26 = 0.038$, corresponding to a predicted precision of

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

3.8%. Compared to this, all results shown in Table 22 and Table 23 above show an improvement that is statistically significant at a confidence level of .95.

The two simplest feature space settings are only prepositions and only ontological types. Compared to the results for these two settings, all additions of features in the feature space result in improvements that are statistically significant, as shown in Table 24.

The experiment setting with the largest precision (85.9) as well as the largest K-score (0.83) is the JRip algorithm with all input features in the feature space and the split version³⁹ of data set. However, our aim is to assess whether it is possible to determine the relation denoted by a preposition by looking at the surrounding ontological types, and thus settings 1 and 4 serve as control settings, and settings 2 (learning on the basis of surrounding ontological types only) and 3 (learning on the basis of the preposition and surrounding ontological types) are the settings we are most interested in analyzing.

The good result of the experiment with all features in the feature space does, however, suggest that we could benefit from adding lexical rules to our final rule set. An example of a rule type that we would consider including is shown in

(72) if (firsthead = 'brev' && prep = 'fra') → (rel eq SRC)

This rule covers 6 instances in the data set, all of which are correctly annotated with SRC. This rule covers examples such as example (73).

(73)

Det fremgår af et brev fra Nordisk Gentofte

It is apparent from a letter from Nordisk Gentofte

³⁹ We cannot compute the corresponding result applying the PART algorithm on a standard computer. The algorithm consumes a large amount of memory, and when applied to the split corpus with all features in the feature space, where the size of the sets of values for the first/second head feature is 1434/1441, the algorithm consumes all available memory (physical as well as max allowed virtual) on a 4Gb PC, and cannot finish.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

The best of the results that include only ontological types and/or preposition in the feature space, is applying the PART-algorithm to the split corpus with preposition and ontological types in the feature space (setting 3). This experiment yields a precision of 59.50% and a K-score of 0.52. We thus choose this rule set for evaluation outside of the machine learning package, as well as for further refinements.

Since the evaluation in WEKA was carried out as a 10-fold cross validation, we now choose to evaluate how the rules perform on the entire data set. Also, we conducted the experiments on the split data set, but now, we test the rules on the original non-split data set. We anticipate that learning on the split data set will result in redundant rules, and thus, this evaluation will uncover such redundant rules.

The rule set consists of 687 rules, which are all applied to the non-split data set. For all rules, we output the number of matches, plus the number of true positives. The result of this first evaluation step yielded the following results:

Instances in data set: 2925
Matched Instances: 2925

Correct: 1781
Non-correct: 1144

Precision: 60.89%
Recall: 100%
f-score: 75.69

As many as 392 of the rules either did not match any instances, or did not produce any true positives, and were thus discarded. This leaves us with 295 rules. These rules were then applied to the data set, yielding the following results:

Instances in data set: 2925
Matched Instances: 2925

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Correct: 1802

Non-correct: 1123

Precision: 61.61%

Recall: 100%

f-score: 76.24

We were thus able to improve the results by 0.72 percentage point by discarding the ineffective rules. This difference is not significant; however, the result is significant compared to the 59.50% precision on the split data set.

6.8 Evaluation and Analysis of Selected Rules

In the following, we will take a closer look at selected rules. We apply three different rankings of the rules, and analyse:

- the 10 most precise rules
- the 10 most covering rules
- the 10 'best' rules

In order to select the 'best' rules, we apply a rule quality measure that allows us to rank the rules.

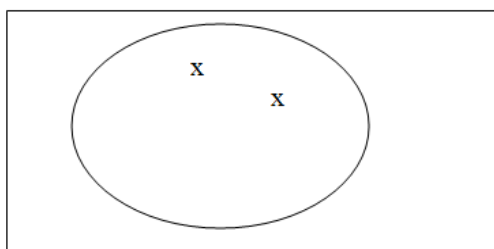


Figure 72 Rule with low coverage and high precision. Box indicates rule coverage, circle indicates correct classification

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

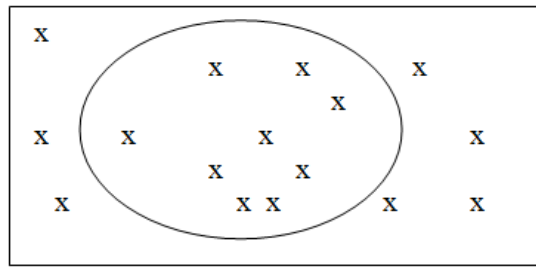


Figure 73 Rule with high coverage and low precision. Box indicates rule coverage, circle indicates correct classification

No absolute measure exists by which we can say that one rule is better than another rule; some rules cover very few instances, but do it with high precision (cf. Figure 71), and some rules cover many instances but with lower precision (cf. Figure 72). Which rule is the better?

Some classes are large (i.e. have many members) and some classes are smaller. If two distinct rules both cover a given number of instances, but the first rule correctly classifies instances to a small class (cf. Figure 73), and the other rule correctly classifies instances to a larger class (cf. Figure 74), is the former rule then better than the latter because it correctly classifies a larger percentage of the class members?

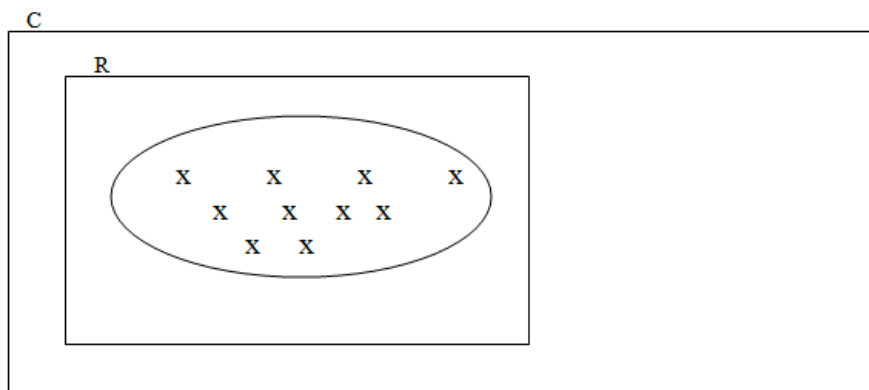


Figure 74 Rule R with high recall, which classifies 100% of the members of the class C correctly. Outer box indicates the class C, inner box indicates coverage of rule R, and circle indicates correct classification.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

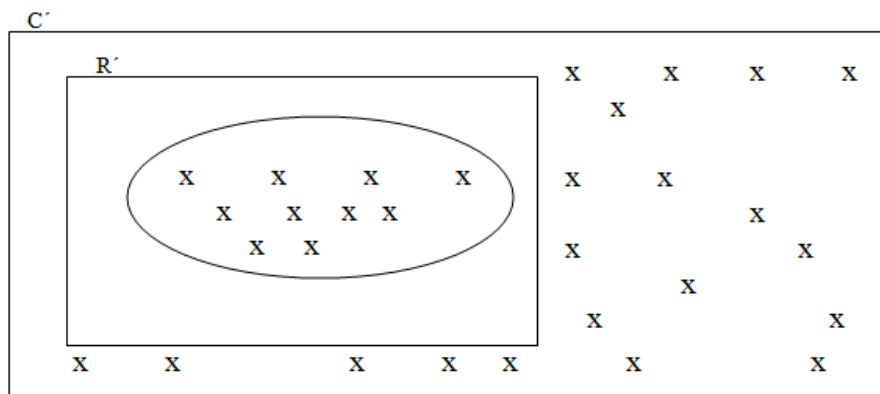


Figure 75 Rule R' with low recall, which classifies 50% of the members of class C' correctly. Outer box indicates the class C' , inner box indicates coverage of rule R' , and circle indicates correct classification.

In order to evaluate the rules, for each of the 295 rules, we count the numbers of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) produced by each rule, according to the classification matrix in Table 25. A perfect rule has positive values along the TP/TN diagonal, null-values along the FP/FN diagonal and maximizes TP.

	Classified as class C	Not classified as class C
Belongs to class C	TP	FN
Does not belong to class C	FP	TN

Table 25 Classification matrix

6.8.1 The 10 most Precise Rules

Precision, or the ratio between the number of true positives and number of matches for a given rule, is computed as:

$$\text{prec}(R) = \frac{|TP|}{|M|}$$

Where:

$|M|$ is the total number of instances matched by rule R .

$|TP|$ is the number of instances covered by rule R and in class C .

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Table 26 shows the 10 rules that have the largest precision score, ranked by largest |TP|.

Rank	Rule#	M	TP	FP	TN	FN	prec(R)
1	#127	11	11	0	2718	196	1,00
2	#388	8	8	0	2182	735	1,00
3	#6	7	7	0	2592	326	1,00
4	#55	7	7	0	2182	736	1,00
5	#339	7	7	0	2182	736	1,00
6	#7	6	6	0	2592	327	1,00
7	#128	6	6	0	2436	483	1,00
8	#533	5	5	0	2182	738	1,00
9	#149	4	4	0	2851	70	1,00
10	#625	4	4	0	2839	82	1,00

Table

26 Scores for the 10 best rules by precision, most covering first

Rank	Rule #	Rule
1	#127	if (prep = 'med' && firsttype = 'Experience') → rel eq WRT
2	#388	if (secondtype = 'Furniture') → rel eq LOC
3	#6	if (prep = 'mellem' && firsttype = 'UnboundedEvent') → rel eq INH
4	#55	if (prep = 'gennem' && secondtype = 'Place') → rel eq LOC
5	#339	if (prep = 'over' && firsttype = 'Natural') → rel eq LOC
6	#7	if (prep = 'mellem' && firsttype = 'Dynamic') → rel eq INH
7	#128	if (prep = 'med' && firsttype = 'Communication') → rel eq PNT
8	#533	if (prep = 'over' && firsttype = 'Human') → rel eq LOC
9	#149	if (prep = 'med' && secondtype = 'Property') → rel eq CHR
10	#625	if (secondtype = 'Physical' && firsttype = 'Physical') → rel eq POF

Table

27 The 10 most precise rules

All of the 10 most precise rules have a precision score of 1, meaning that they classify all matched instances correctly. In all, 81 rules have a precision score of 1, and in combination these rules cover 187 instances or 6.4% of

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

the data set. 8 of the 10 rules have restrictions on the form of the preposition (3 require *med*, 2 *mellem*, 2 *over*, 1 *gennem*). 7 rules have restrictions on the ontological type of the head of the first NP (1 requires COMMUNICATION, 1 DYNAMIC, 1 EXPERIENCE, 1 HUMAN, 1 NATURAL, 1 PHYSICAL, 1 UNBOUNDEDEVENT) and 4 rules have restrictions on the ontological type of the head of the second NP (1 requires FURNITURE, 1 PHYSICAL, 1 PLACE, 1 PROPERTY).

Rule no. 127 says that if the form of the preposition is *med* and the ontological type of the first NP is EXPERIENCE, then the relation denoted by the preposition is classified as WRT. The rule matches text chunks such as (74) with the corresponding instance in the data set in (75):

(74)
ingen problemer med den kulturelle arv
no problems with the cultural inheritance

(75)
med,WRT,problem,thirdOrderEntity;Mental;Experience,arbejdsmiljø,Property

Rule no. 388 says that, regardless of the form of the preposition, if the ontological type of the second NP is FURNITURE, then the relation denoted by the preposition is classified as LOC. The rule matches text chunks such as (76) with the corresponding instance in the data set in (77):

(76)
halvtaget over gyngestolen
the porch roof above the rocking chair

(77)
over,LOC,halvtag,Building;Object;Part,gyngestol,Furniture;Artifact;Object

Rule no. 6 says that if the form of the preposition is *mellem* and the ontological type of the first NP is UNBOUNDEDEVENT, then the relation denoted by the preposition is classified as INH. The rule matches text chunks such as (78) with the corresponding instance in the data set in (79):

Uncovering of the Semantic Relations Denoted by a Selection of Danish
Prepositions

(78)

Konflikten mellem de to lande

The conflict between the two countries

(79)

mellem,INH,konflikt,UnboundedEvent;Agentive,land,Human;Object;Group

Rule no. 55 says that if the form of the preposition is *gennem* and the ontological type of the second NP is PLACE, then the relation denoted by the preposition is classified as LOC. The rule matches text chunks such as (80) with the corresponding instance in the data set in (81):

(80)

lysens lange rejse gennem universet

light's long journey through the universe

(81)

gennem,LOC,rejse,UnboundedEvent;Cause;Physical;Location,univers,Place
;Object

Rule no. 339 says that if the form of the preposition is *over* and the ontological type of the first NP is NATURAL, then the relation denoted by the preposition is classified as LOC. The rule matches text chunks such as (82) with the corresponding instance in the data set in (83):

(82)

armene over hovedet

arms above the head

(83)

over,LOC,arm,Natural;Object;BodyPart,hoved,Part

Rule no. 7 says that if the form of the preposition is *mellem* and the ontological type of the first NP is DYNAMIC, then the relation denoted by the preposition is classified as INH. The rule matches text chunks such as (84) with the corresponding instance in the data set in (85):

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

(84)

transmissionen mellem centralerne
the transmission between the centrals

(85)

mellem,INH,transmission,Dynamic;Agentive;Communication,central,Building;Object

Rule no. 128 says that if the form of the preposition is *med* and the ontological type of the first NP is COMMUNICATION, then the relation denoted by the preposition is classified as PNT. The rule matches text chunks such as (86) with the corresponding instance in the data set in (87):

(86)

en snak med børnene
a talk with the children

(87)

med,PNT,snak,Dynamic;Agentive;Communication,barn,Human;Object

Rule no. 533 says that if the form of the preposition is *over* and the ontological type of the first NP is HUMAN, then the relation denoted by the preposition is classified as LOC. The rule matches text chunks such as (88) with the corresponding instance in the data set in (89):

(88)

teenagere over hele kloden
teenagers throughout the planet

(89)

rule533:1759,_over_,LOC,teenager,Human;Object,klode,Object;Natural

Rule no. 149 says that if the form of the preposition is *med* and the ontological type of the second NP is PROPERTY, then the relation denoted by the preposition is classified as CHR. The rule matches text chunks such as (90) with the corresponding instance in the data set in (91):

(90)

børn med autisme
children with autism

(91)

med,CHR,barn,Human;Object,autisme,Property;Condition;Physical

Rule no. 625 says that if the ontological type of both the first and the second NP is PHYSICAL, then the relation denoted by the preposition is classified as POF. The rule matches text chunks such as (92) with the corresponding instance in the data set in (93):

(92)

Finalerne ved de sjællandske mesterskaber
The finals at the Zealandic championships

(93)

ved,POF,finale,UnboundedEvent;Agentive;Physical;Social,mesterskab,
Dynamic;Agentive;Physical;Purpose;Social

6.8.2 The 10 most Covering Rules

We here define coverage as correct coverage, because we are interested in finding the rules that cover the most instances correctly. We thus compute a correct coverage score, $C_{cov}(R)$, as the ratio between the number of true positives and number of instances in the data set:

$$C_{cov}(R) = \frac{|TP|}{|D|} \cdot 10$$

Where:

$|D|$ is the total number of instances in a data set D.

$|TP|$ is the number of instances covered by rule R and in class C.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Rank	Rule#	D	TP	FP	TN	FN	Ccov(R)
1	#27	2925	101	13	2169	642	0.35
2	#20	2925	89	18	2139	679	0.30
3	#9	2925	72	22	2160	671	0.25
4	#69	2925	67	66	2370	422	0.23
5	#37	2925	50	19	2417	439	0.17
6	#39	2925	44	26	2131	724	0.15
7	#134	2925	42	11	2581	291	0.14
8	#402	2925	41	18	2164	702	0.14
9	#70	2925	36	3	2154	732	0.12
10	#4	2925	32	2	2590	301	0.11

Table 28 Scores for the 10 best rules by correctly covered instances

Rank	Rule #	Rule
1	#27	if (prep = 'i' && secondtype = 'Human') → rel eq LOC
2	#20	if (prep = 'gennem' && secondtype = '3rdOrderEntity') → rel eq TMP
3	#9	if (prep = 'fra' && firsttype = 'Human') → rel eq LOC
4	#69	if (prep = 'til' → rel eq PNT
5	#37	if (prep = 'til' && firsttype = '3rdOrderEntity') → rel eq PNT
6	#39	if (prep = 'efter' && secondtype = '3rdOrderEntity') → rel eq TMP
7	#134	if (prep = 'for' && firsttype = 'Human') → rel eq INH
8	#402	if (prep = 'ved' && secondtype = 'Human') → rel eq LOC
9	#70	if (prep = 'efter' && secondtype = 'BoundedEvent') → rel eq TMP
10	#4	if (prep = 'mellem' && firsttype = 'Purpose') → rel eq INH

Table 29 10 best rules by correctly covered instances

All of the 10 most precise rules have a precision score of 1, meaning that they classify all matched instances correctly. In all, 81 rules have a precision score of 1, and in combination these rules cover 187 instances or 6.4% of the data set. 8 of the 10 rules have restrictions on the form of the preposition

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

(3 require *med*, 2 *mellem*, 2 *over*, 1 *gennem*). 7 rules have restrictions on the ontological type of the head of the first NP and 4 rules have restrictions on the ontological type of the head of the second NP.

The rule that covers most instances in the data set, rule no. 27, says that if the form of the preposition is *i* and the ontological type of the second NP is HUMAN, then the relation is classified as LOC.

DanNet contains metonymic senses of place denoting words; for example, the word *land* (country) is both in a synset with the ontological type PLACE+OBJECT, as well as in a synset with the ontological type HUMAN+OBJECT+GROUP. As it turns out, systematically, all countries and cities have been annotated with HUMAN+OBJECT+GROUP in the data set. In some cases, of course, this is an adequate annotation, but for other instances it is not the preferred reading. Thus, in this case, the rule has been inferred on the grounds that place names have been uniquely annotated with the ontological type HUMAN+OBJECT+GROUP. If we assume that in the ontological type annotation, we assign all possible types to a given lemma, then this does not present a problem since any given country would be assigned both types and the rule would match no matter what. However, going through the matched instances for this rule, for all instances, the most appropriate ontological type for the second NPs would be PLACE+OBJECT, and thus the rule would be intuitively more correct if it had the form presented in (97).

The rule matches text chunks such as (94) with the corresponding instance in the data set in (95), which should have been as in (96).

(94)

FNs særlige udsending i Burma
the UN special emissary in Burma

(95)

i,LOC,udsending,Human;Object,Burma,Human;Object;Group

(96)

i,LOC,udsending,Human;Object,Burma,Place;Object

(97)

if (prep = 'i' && secondtype = 'Place') → rel eq LOC

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Rule no. 20 says that if the form of the preposition is *gennem* and the ontological type of the second NP is 3RDOORDERENTITY, then the relation is classified as TMP. The rule matches text chunks such as (98) with the corresponding instance in the data set in (99):

(98)

hans elskerinde gennem fire år
his mistress of four years

(99)

gennem,TMP,elskerinde,Human;Object,år,thirdOrderEntity;Time

Rule no. 9 says that if the form of the preposition is *fra* and the ontological type of the first NP is HUMAN, then the relation is classified as LOC. The rule matches text chunks such as (100) with the corresponding instance in the data set in (101). Note the inappropriate annotation of the place name *Sønderborg*, as described above regarding rule no. 27. In this case, however, the error is not significant as the rule does not put restrictions on the ontological type of the second NP in which the error occurs.

(100)

Blank Jørgensen fra Sønderborg
Blank Jørgensen from Sønderborg

(101)

fra,LOC,jørgensen,Human;Object,Sønderborg,Human;Object;Group

Rule no. 69 says that if the form of the preposition is *til*, then the relation is classified as PNT. The rule matches text chunks such as (102) with the corresponding instance in the data set in (103):

(102)

hjælpen til akut syge
the assistance for the acutely ill

(103)

til,PNT,hjælp,Dynamic;Agentive;Purpose;Social,syg,Human;Object

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Rule no. 37 says that if the form of the preposition is *til* and the ontological type of the first NP is 3RDOORDERENTITY, then the relation is classified as PNT. The rule matches text chunks such as (104) with the corresponding instance in the data set in (105):

(104)
anledning til bekymring
cause for worry

(105)
til,PNT,anledning,3rdOrderEntity;Mental;Purpose;Social,bekymring,Dynamic;Experience;Mental

Rule no. 39 says that if the form of the preposition is *efter* and the ontological type of the second NP is 3RDOORDERENTITY, then the relation is classified as TMP. The rule matches text chunks such as (106) with the corresponding instance in the data set in (107):

(106)
Otte minutter efter pausen
Eight minutes after the break

(107)
efter,TMP,minut,3rdOrderEntity;Time,pause,3rdOrderEntity;Time

Rule no. 134 says that if the form of the preposition is *for* and the ontological type of the first NP is HUMAN, then the relation is classified as INH. The rule matches text chunks such as (108) with the corresponding instance in the data set in (109):

(108)
En talsmand for græsrodsbevægelserne
A spokesperson for the NGOs

(109)
for,INH,talsmand,Human;Object,græsrodsbevægelse,Human;Object;Group

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Rule no. 402 says that if the form of the preposition is *ved* and the ontological type of the second NP is HUMAN, then the relation is classified as LOC. This rule suffers from the same problem as rule no. 27 above: The instances have been inappropriately annotated with the ontological type for the metonymic sense, HUMAN+OBJECT+GROUP, instead of the locative sense, PLACE+OBJECT, as would be preferred in this case. The rule matches text chunks such as (110) with the corresponding instance in the data set in (111), which should have been as in (112). For rule no. 402, the form in the final rule set will be: if (prep = 'ved' && secondtype = 'Place') → rel eq LOC

(110)

Ejby Skov ved Køge

Ejby Forest near Køge

(111)

ved_,LOC,skov, Place;Object;Group,Køge,Human;Object;Group

(112)

ved_,LOC,skov, Place;Object;Group,Køge,Place;Object;Natural

Rule no. 70 says that if the form of the preposition is *efter* and the ontological type of the second NP is BOUNDEDEVENT, then the relation is classified as TMP. The rule matches text chunks such as (113) with the corresponding instance in the data set in (114):

(113)

situationen efter næste valg

the situation after the next election

(114)

efter,TMP,situation,2ndOrderEntity,valg,BoundedEvent;Agentive;Purpose

Rule no. 4 says that if the form of the preposition is *mellem* and the ontological type of the first NP is PURPOSE, then the relation is classified as INH. The rule matches text chunks such as (115) with the corresponding instance in the data set in (116):

(115)

en aftale mellem et flertal af medlemmerne

an agreement between a majority of the members

(116)

mellem,INH,aftale,BoundedEvent;Agentive;Purpose;Communication,flertal
,thirdOrderEntity;Quantity

6.8.3 The 10 'best' Rules

If we wish to rank the rules by a quality score other than precision or coverage as above, we need to evaluate the quality of individual rules by some measure. Many measures of rule quality have been proposed in the literature, cf. e.g. (An & Cercone, 2000; Dean & Famili, 1997; Freitas, 1999) We here choose to apply a measure that combines rule coverage and rule accuracy, here named $Q(R)$ (cf. (Dean & Famili, 1997)).

For each of the 295 rules, we thus compute rule accuracy, rule coverage and the rule quality-score $Q(R)$, by which we can rank the rules. This rule quality measure favors rules that classify a large percentage of the instances belonging to a class correctly. The choice of this particular quality measure is to some extent arbitrary. For the present purpose of selecting a subset of the rules for individual descriptions in this dissertation, it is adequate. However, if we wish to rank the outcome of the rules according to the rule quality in an application, we would have to give the choice of a rule quality measure thorough consideration.

First, we compute a rule accuracy score, $acc(R)$, which takes into account the number of true positives (TP) as well as the number of true negatives (TN) and the total number of instances in the data set. This score is computed as:

$$acc(R) = \frac{|TP| + |TN|}{|D|}$$

Where:

$|D|$ is the total number of instances in a data set D .

$|TP|$ is the number of instances covered by rule R and in class C .

$|TN|$ is the number of instances not covered by rule R and not in class C .

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Then, we compute an estimated rule coverage score, $EC(R)$, which takes into account the number of true positives as well as the number of instances belonging to the class that the rule classifies instances as. This score is computed as:

$$EC(R) = \exp\left(\frac{|TP|}{|C|} - 1\right)$$

Where:

$|C|$ is the total number of instances belonging to class C.

$|TP|$ is the number of instances covered by rule R and in class C.

We now combine these two scores and compute the rule quality score, $Q(R)$, which is a value between 0 and 10, where 0 is a poor quality and 10 is a perfect quality. The score is computed as:

$$Q(R) = (acc(R) \cdot EC(R)) \cdot 10$$

Table 30 shows the scores for the top 10 best rules ranked by $Q(R)$, and Table 31 shows the top 10 best rules.

Rank	Rule#	Matches	TP	TN	C	D	acc(R)	EC(R)	Q(R)
1	#564	15	5	2901	14	2925	0.99	0.53	5.22
2	#246	48	29	2820	86	2925	0.97	0.52	5.02
3	#347	30	15	2836	74	2925	0.97	0.45	4.39
4	#438	2	2	2911	14	2925	1.00	0.42	4.23
5	#14	18	15	2831	91	2925	0.97	0.43	4.22
6	#405	8	4	2891	30	2925	0.99	0.42	4.16
7	#203	1	1	2917	8	2925	1.00	0.42	4.16
8	#201	2	1	2916	8	2925	1.00	0.42	4.16
9	#132	12	10	2849	74	2925	0.98	0.42	4.12
10	#265	17	13	2824	97	2925	0.97	0.42	4.08

Table 30 Scores for 10 best rules ranked by $Q(R)$

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

Rank	Rule #	Rule
1	#564	if (prep = 'under and secondtype = 'Human') → rel eq SUP
2	#246	if (prep = 'af and firsttype = 'Part') → rel eq POF
3	#347	if (prep = 'med') → rel eq CHR
4	#438	if (prep = 'under and firsttype = 'UnboundedEvent') → rel eq SUP
5	#14	if (prep = 'fra and firsttype = 'Mental') → rel eq SRC
6	#405	if (prep = 'for and secondtype = 'Building') → rel eq QUA
7	#203	if (prep = 'med and firsttype = 'Underspecified') → rel eq RLO
8	#201	if (prep = 'med and firsttype = 'Furniture') → rel eq RLO
9	#132	if (prep = 'med and secondtype = 'Dynamic') → rel eq CHR
10	#265	if (prep = 'på and firsttype = '3rdOrderEntity') → rel eq MEA

Table 31 The 10 ‘best’ rules.

None of the ten highest ranked rules in Table 31 have restrictions on the ontological types of both NP heads. Six rules have restrictions on the ontological type of the first NP head, three have restrictions on the ontological type of the second NP head, and one rule only has a restriction on the form of the preposition.

The rule that has the highest ranking, rule no. 564, says that if the preposition has the form *under* and the ontological type of the second NP head is HUMAN, then the relation denoted by the preposition is SUP. This rule matches 15 instances of which 5 are correctly classified. The data set contains just 14 instances of the relation SUP, and the rule thus correctly classifies 35.7% of the instances of this relation. The rule incorrectly classifies 3 instances of COM, 3 of LOC, 1 of MNR, 1 of POF and 2 of TMP as SUP.

The rule matches text chunks such as (117) with the corresponding instance in the data set in (118):

(117)

Danmark under Fogs ledelse

Denmark under the governance of Fog

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

(118)

under,SUP,Danmark,Human;Object;Group,ledelse,Human;Object;Group

The next rule, rule no. 246, says that if the preposition is *af* and the ontological type of the first NP head is PART, then the relation denoted by the preposition is POF. This rule matches 48 instances, of which 29 are correctly classified. The data set contains 86 instances of the relation POF, and the rule thus correctly classifies 33.7% of the instances of this relation. The rule incorrectly classifies 18 instances of INH and 1 of TMP as POF. The rule matches text chunks such as (119) with the corresponding instance in the data set in (120):

(119)

en del af det gamle, statslige postvæsen
a part of the old, national postal service

(120)

af,POF,del,Artifact;Object;Part,postvæsen,thirdOrderEntity;Mental;Purpose
;Social;Institution

The next rule, rule no. 347, says that if the preposition is *med*, then the relation denoted by the preposition is CHR. This rule matches 30 instances, of which 15 are correctly classified. The data set contains 74 instances of the relation CHR, and the rule thus correctly classifies 20.3% of the instances of this relation. The rule incorrectly classifies 1 instance of BMO, 5 of CMP, 1 of CUM, 1 of INH, 1 of MEA, 2 of MNR, 2 of PNT, 1 of RLO and 1 of WRT as CHR.

The rule matches text chunks such as (121) with the corresponding instance in the data set in (122):

(121)

patienter med depression
patients with depression

(122)

med,CHR,patient,Human;Object,depression,Property;Condition;Physical

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

The next rule, rule no. 438, says that if the preposition is *under* and the ontological type of the first NP head is UNBOUNDEDEVENT, then the relation denoted by the preposition is SUP. This rule matches 2 instances both of which are correctly classified. The data set contains 14 instances of the relation SUP, and the rule thus correctly classifies 14.3% of the instances of this relation.

The rule matches text chunks such as (123) with the corresponding instance in the data set in (124):

(123)
ulidelige forhold under en ledelse
unbearable conditions under a management

(124)
under,SUP,forhold,UnboundedEvent;Agentive;Physical;Social,ledelse,Human;Object;Group

The next rule, rule no. 14, says that if the preposition is *fra* and the ontological type of the first NP head is MENTAL, then the relation denoted by the preposition is SRC. This rule matches 18 instances, of which 15 are correctly classified. The data set contains 91 instances of the relation SRC, and the rule thus correctly classifies 16.5% of the instances of this relation.

The rule incorrectly classifies 2 instances of LOC and 1 of AGT as SRC.

The rule matches text chunks such as (125) with the corresponding instance in the data set in (126):

(125)
klar besked fra den australske regering
clear message from the Australian government

(126)
fra,SRC,besked,BoundedEvent;Agentive;Mental;Communication,regering,Human;Object;Group

The next rule, rule no. 405, says that if the preposition is *for* and the ontological type of the second NP head is BUILDING then the relation denoted by the preposition is QUA. This rule matches 8 instances of which 4 are correctly classified. The data set contains 30 instances of the relation

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

QUA, and the rule thus correctly classifies 13.3% of the instances of this relation. The rule incorrectly classifies 1 instance of CMP, 1 of PNT and 2 of WRT as QUA. The ontological type annotation is again inappropriate: the word form *arbejde* belongs to more than one synset, one of which has the ontological type BUILDING+OBJECT (glossed: *sted hvor man udøver denne virksomhed* or ‘place where this activity is exercised’), and another DYNAMIC+AGENTIVE+PURPOSE+SOCIAL. For the example given in (128) as well as the other instances of the relation QUA that give rise to this rule, the appropriate ontological type annotation would have been DYNAMIC+AGENTIVE+PURPOSE+SOCIAL.

The rule matches text chunks such as (127) with the corresponding instance in the data set in (128), which ideally should have been as in (129). For rule no. 402, the form in the final rule set will be: if (prep = 'for' && secondtype = 'Agentive') → rel eq QUA

(127)

Belønning for hårdt arbejde

A reward for hard work

(128)

for,QUA,belønning,thirdOrderEntity;Quantity,arbejde,Building;Object

(129)

for,QUA,belønning,thirdOrderEntity;Quantity,arbejde,Dynamic;Agentive;Purpose;Social

The next rule, rule no. 203, says that if the preposition is *med* and the ontological type of the first NP head is UNDERSPECIFIED, then the relation denoted by the preposition is RLO. This rule matches 1 instance which is correctly classified. The data set contains 8 instances of the relation RLO, and the rule thus correctly classifies 12.5% of the instances of this relation.

The rule matches text chunks such as (130) with the corresponding instance in the data set in (131): The ontological type UNDERSPECIFIED means that any given synset that has this ontological type is not yet fully analyzed (B. S. Pedersen, PC, March, 2010), and thus, this rule is uninformative. However, we can hypothesize that the inverse locative relation RLO will restrict the same ontological type for its first NP as its inverse relation LOC

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

does for its second NP, e.g. PLACE as in rule no. 147: if (prep = 'på' && secondtype = 'Place') → rel eq LOC.

(130)

en sættevogn med en container
a semi-trailer with a container

(131)

med,RLO,sættevogn,Underspecified,container,Container;Artifact;Object

The next rule, rule no. 201, says that if the preposition is *med* and the ontological type of the first NP head is FURNITURE, then the relation denoted by the preposition is RLO. This rule matches 2 instances of which 1 is correctly classified. The data set contains 8 instances of the relation RLO, and the rule thus correctly classifies 12.5 % of the instances of this relation. The rule incorrectly classifies 1 instance of CHR as RLO.

The rule matches text chunks such as (132) with the corresponding instance in the data set in (133). It is our belief, based on rules no. 203 and 201, that a more general rule for the RLO relation would restrict the ontological type of the first NP to OBJECT. Thus, in the final rule set, these two rules will be replaced by a new rule given in (134).

(132)

keramiske borde med et kunstnerisk tilsnit
ceramic tables with an artistic appearance

(133)

med,CHR,bord,Furniture;Artifact;Object,tilsnit,Property

(134)

if (prep = 'med' and firsttype = 'Object') → rel eq RLO

The next rule, rule no. 132, says that if the preposition is *med* and the ontological type of the second NP head is DYNAMIC, then the relation denoted by the preposition is CHR. This rule matches 12 instances of which 10 are correctly classified. The data set contains 74 instances of the relation CHR, and the rule thus correctly classifies 13.5% of the instances of this

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

relation. The rule incorrectly classifies 1 instance of CBY and 1 of PNT as CHR.

The rule matches text chunks such as (135) with the corresponding instance in the data set in (136):

(135)

patienter med et mere akut behov
patients with more acute needs

(136)

med,CHR,patient,Human;Object,behov,Dynamic;Experience;Mental

The next rule, rule no. 265, says that if the preposition is *på* and the ontological type of the first NP head is 3RORDERENTITY, then the relation denoted by the preposition is MEA. This rule matches 17 instances, of which 13 are correctly classified. The data set contains 97 instances of the relation MEA, and the rule thus correctly classifies 13.4% of the instances of this relation. The rule incorrectly classifies 1 instance of LOC, 1 of PNT, 1 of TMP and 1 of WRT as MEA. The rule matches text chunks such as (137) with the corresponding instance in the data set in (138):

(137)

en årlig indtægt på 198.400 kroner
an annual income of 198,400 kroner

(138)

på,MEA,indtægt,3rdOrderEntity;Quantity,krone,Static;Location

There is no overlap between the 10 most precise, 10 most covering and 10 best rules described above.

Not all the relations that were the result of the analysis of dictionary entries, as described in section 6.4, were used in the annotation, as described in section 6.5.2, and not all relations that were used in the annotation resulted in rules. The matrix in Figure 75 shows:

- For (x,y): The preposition *x* denoting the relation *y* is identified in corpus, and at least one rule is inferred

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

- For (x,y): The preposition x denoting the relation y is identified in corpus, but no rules are inferred
- For (y): The relation y is identified in corpus, but no rules are inferred
- For (y): The relation y is not identified in corpus

	af	efter	for	fra	gennem	hos	i	med	mellem	over	på	til	under	ved
ADD		●	○											
AGT	●			○		○	○	○						○
BMO					●		○	●		○	○			○
CAU														
CBY	○	●		○	○			○		●				○
CHR	○						●	●						
CMP	○		●					●			●			
COM	○	○								●			○	○
CUM								○						
INH	●	○	●	○		●	●	○	●	●	○	○		●
LOC	○		●	●	●	●	●		●	●	●	○	●	●
MEA	○		●	○				○	○	●	●	○	●	○
MNR								○			○		○	●
MTH		○												
PNT	●	○	●		●	●	●	●	○	●	●	●	○	●
POF	●			○		○	○				●		●	○
PRP												○		
QUA		●	●										○	
RCH	○					○				○				○
RLO	○							●						
RST												○		
SBT														
SRC	○	○		●	○	○								
HPR														
HPO														
SUP													●	
TAR		●										○		
TMP	○	●	○	●	●		●		○	●	●	○	●	○
WRT	●		●	○		○	●	●	○	●	●	○		●

Figure 76 Matrix of relations and prepositions, showing which combinations resulted in rules.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

6.9 Dictionary of Prepositions

As a result, using the rules produced by the machine learning algorithm, and the modifications described above, we can now produce a dictionary of prepositions. The dictionary consists of entries for the 14 Danish prepositions with a specification of the relations they can denote when they occur in the specific syntactic construction NP-PREP-NP. The dictionary contains ontological restrictions on the arguments of the preposition where such have been inferred, and for each entry marks a default relation that associates the given preposition with the most frequent relation for that given preposition. For each entry, if we apply the restrictions to our data set, we get the precision scores as shown in Figure 76.

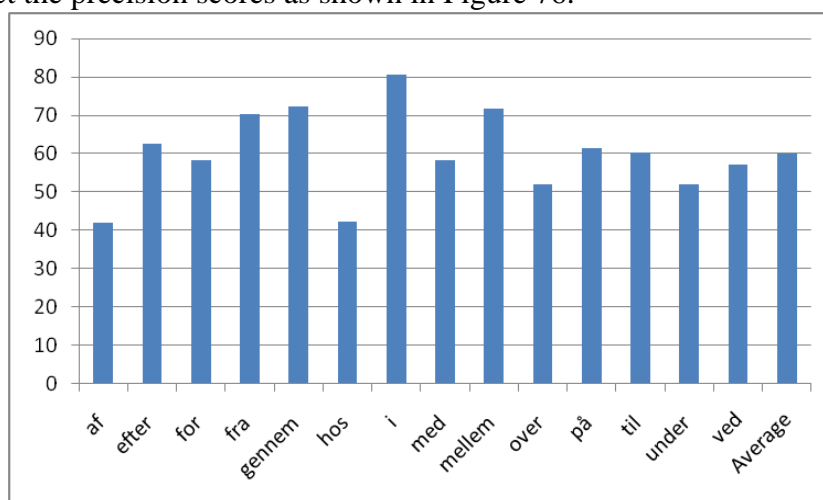


Figure 77 Precision scores for the individual entries in the dictionary of prepositions

The relations listed in an entry occur in order of frequency of the relation for the given preposition in the corpus.

Next to the relation, a corpus-evidence is given. The corpus-evidence merely exemplifies the relation, but is not necessarily an exemplar of the ontological types mentioned.

Below the relation and the corpus-evidence, ontological restrictions are given for the arguments, 1^{st} *onto* (the ontological type of the head of the first NP) and 2^{nd} *onto* (the ontological type of the head of the second NP). These

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

restrictions are the result of machine learning on an annotated corpus consisting of the aforementioned subset of Korpus 2000, followed by a human analysis.

Some relations have restrictions for one argument, some for both, and some do not have any restrictions.

A (u) in front of a relation means that no rules have been inferred for this relation, but it is part of the relation inventory for the given preposition.

A (d) in front of a relation means that this can be viewed as a default relation, because it is the most frequent relation for a given preposition.

6.9.1 Example Entries

AF 'æresmedlem af Dansk Brygmester Forening'

INH

1st onto: 'Human'
'Occupation'

This means that for a given instance of the preposition 'af' in a NP-PREP-NP construction, if the ontological type of the first NP is 'Human' or 'Occupation', then the relation is INH. No restrictions are put on the ontological type of the second NP. The corpus-evidence 'æresmedlem af Dansk Brygmester Forening' exemplifies the relation INH.

GENNEM/IGENNEM

(d) **TMP** 'dansk kunst fra det 18. århundrede'
2nd onto: '3rdOrderEntity'
'Purpose'

This means that for a given instance of the preposition 'gennem' or 'igennem' in a NP-PREP-NP construction, if the ontological type of the second NP is '3rdOrderEntity' or 'Purpose', then the relation is TMP. The flag (d) means that TMP is the most frequent relation for the preposition 'gennem/igennem'. No restrictions are put on the ontological type of the first NP.

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

FOR

LOC		'nord for Viborg'			
	<i>1st onto:</i>	'3rdOrderEntity'	<i>2nd onto:</i>	'Place'	

This means that for a given instance of the preposition ‘for’ in a NP-PREP-NP construction, if the ontological type of the first NP is '3rdOrderEntity' and the ontological type of the second NP is 'Place', then the relation is LOC.

VED

(u)	BMO	'formering ved tilfældig knopskydning'
-----	-----	--

This means that the preposition ‘ved’ can denote the relation BMO, but no rules have been inferred for this.

The full dictionary of prepositions is given in Appendix A.

6.10 Summary

In this chapter, we have described an experiment that aims at uncovering the semantic relations denoted by a selection of Danish prepositions. We have described the selection process for the prepositions to receive further treatment: this process includes an analysis of the extensional descriptions in a selection of dictionaries and reference works. Further, we have described the analysis of the selected prepositions through dictionary definitions. This resulted in a preliminary relation set that was used in the annotation of corpus evidences. As a result, we have a final set of relations that can be denoted by the selection of 14 Danish prepositions occurring in the specific syntactic pattern NP-PREPOSITION-NP. We have given a detailed description of each of these relations, as well as of the prepositions. Further, we have annotated the corpus with various features, including ontological types. The ontological types stem from the Danish wordnet DanNet. Subsequently, we performed machine-learning on the resulting dataset, and as a result, we have set of rules that may predict the semantic relation denoted by a preposition given the ontological types of one or both arguments. This knowledge is presented in a dictionary of prepositions rendered in Appendix A. Provided that a given preposition heads a noun modifying PP, and that any of the NP heads for which the ontological type

Uncovering of the Semantic Relations Denoted by a Selection of Danish Prepositions

that is restricted by the rule can be mapped to a synset in DanNet, the final rule set will correctly assign the correct semantic relation to 61.6% of the instances of these prepositions in running text. The difference in the figures from 59.5% as the precision for the rules prior to reduction to 61.5% as the precision for the final reduced rule set is significant.

Uncovering of the Semantic Relations Denoted by a Selection of Danish
Prepositions

Chapter 7

Conclusion and Future Work

The general goal for our work is to provide better search possibilities in large text collections. Part of this goal can be achieved by improving or enabling semantic analysis of documents, and semantic analysis of prepositions is a small part of this. The experiments presented in this dissertation concern analysis of Danish language texts, and this aspect is important. Small languages such as Danish receive less attention than larger languages in general research, notably less than English. However, if we wish to be able to use our language in the age of the internet, we need to conduct research on it.

The specific goal for the work presented in this thesis was to uncover the senses of Danish prepositions. The senses, in this context, are semantic relations denoted by prepositions. In order to give an account of this topic, we first needed to define the essence of the class of prepositions. Next, we defined a set of possible relations that prepositions can denote, and finally, we discovered the senses that prepositions in Danish texts in fact do express, and inferred ontological affinity rules for these.

Thus, the questions that this dissertation has sought to answer are the following:

1. What is an adequate definition of the class of prepositions?
2. Which semantic relations can a subset of Danish prepositions denote?

Conclusion and Future Work

3. Can we infer ontological affinity rules for the relations denoted by a subset of Danish prepositions from an annotated corpus?

For the first part of the research question, we reviewed definitions of the class of prepositions from a variety of reference works on the Danish language; general as well as specific. A specific reference work is that of Brøndal which provides a thorough treatment of a theory of prepositions. However, we found Brøndal's theory opaque, and thus, did not find it applicable to our treatment of prepositions. Other definitions, however, provided valuable insight into the essence of the class. We concluded by phrasing a definition that adequately defines the class of prepositions for our further work:

- a) *The class consists of uninflectable words which may be of simple, compound or complex form.*
- b) *Prepositions are transitive. Their complement may be of various forms but is typically a noun, a pronoun or a clause (including infinitives).*
- c) *Prepositions are pure relators that denote binary relations.*

For the second part of the research question concerning an analysis of the semantic relations that a subset of Danish prepositions can denote, we first selected a subset of 14 Danish prepositions based on their common inclusion in a number of Danish dictionaries. These prepositions were then analysed, and a preliminary list of possible senses was produced based on their definitions in the selection of dictionaries. Subsequently, ~3500 corpus items were analyzed, and as a result, a final list of 29 possible senses for the subset of prepositions was produced.

For the third part of the research question concerning whether it is possible to infer ontological affinity rules for the relations denoted by a subset of Danish prepositions from an annotated corpus, we performed two experiments.

In chapter 5, we described our introductory experiments with a machine learning approach to disambiguation of semantic relations denoted by prepositions. We asserted that the task is similar to, but not identical to, word sense disambiguation. The difference lies primarily in the purpose, which was to produce conceptual feature structures representing the conceptual content of textual expressions of the syntactic form NP-PREP-NP. The corpus was annotated with various features including ontological type and semantic relation. The ontological types in this experiment were top-ontology concepts from the SIMPLE-ontology, and the semantic relations from a set of 12 relations previously used in the OntoQuery project. We then performed machine-learning experiments with the inclusion of a variety of feature combinations, and even though these experiments were performed on a limited test corpus of ~900 corpus items, our results showed that it is indeed possible to infer rules that predict the relation denoted by a preposition – at least within the domain of nutrition. Despite the limited size of our dataset, we achieved an encouraging precision of 86.5% for the feature space consisting of ontological types and prepositions.

In chapter 6, we described a larger experiment, for which we marked up a subpart of the Danish general language corpus Korpus 2000 with various features, including ontological type information and semantic relations based on our analysis of prepositional senses. The corpus consisted of ~3000 corpus items. We fed this dataset to a machine-learning algorithm, and the resulting rule set consisted of 687 rules. We performed various evaluations on different subsets of this rule set, and performed evaluations based on different quality measures. In the course of these evaluations, as many as 392 of the rules were discarded because they did not produce any true positives. This left us with 295 rules which yielded a precision of 61.6% and a recall of 100%. These rules were transformed into a dictionary of prepositional senses in which, given a preposition and a sense, ontological affinities are expressed as restrictions on the ontological types of the arguments.

Thus, we can conclude that it is possible to infer ontological affinity rules for relations denoted by a subset of Danish prepositions by means of an annotated corpus. The fact that the results of our first experiments yielded a

Conclusion and Future Work

better precision than the results of the second is neither to be taken as proof that one relation set is better than the other, nor that affinities are more pronounced in a nutrition domain than in general language texts. We believe that the better results stem from two main causes:

- 1) The relation set in the first experiments was smaller, which alone means that a relation assignment based on pure chance would yield a better precision.
- 2) The annotation process in the second experiment was performed with greater insight following the detailed analysis of prepositional senses. In the first experiment, a large number of prepositions were annotated with the underspecified aboutness relation WRT, which made this relation a trivial rejector with a high precision. In the second experiment, there was a balanced distribution of the 14 included prepositions, which was not the case in the first experiment.

In addition to being useful in an information search system, the results of this research have provided new knowledge about the relations that the subset of Danish prepositions can denote as well as of the ontological affinities for these relations.

We would like to pursue this approach, and analyze other syntactic forms than NP-PREPOSITION-NP, in particular V-PREPOSITION-NP constructions. We would also like to investigate other relation denoting word classes.

Conclusions and Future Work

Appendix A

A Rule-based Dictionary of Danish Prepositions

Introduction

This is a dictionary of 14 Danish prepositions and the relations they can denote when they occur in the specific syntactic construction NP-PREP-NP.

For each preposition, a number of possible relations are given. The relations are identified through 1) an analysis of a range of monolingual Danish and bilingual Danish-English dictionaries and 2) an analysis of a corpus comprising 2925 text excerpts of the form NP-PREP-NP from the citation version of Korpus 2000.

The relations listed in an entry occur in order of frequency of the relation for the given preposition in the corpus.

Next to the relation, a corpus evidence is given. The corpus evidence merely exemplifies the relation, but is not necessarily an exemplar of the ontological types mentioned.

Below the relation and the corpus evidence, ontological restrictions are given for the arguments, *1st onto* (the ontological type of the head of the first NP) and *2nd onto* (the ontological type of the head of the second NP). These restrictions are the result of machine learning on an annotated corpus consisting of the aforementioned subset of Korpus 2000, followed by a human analysis.

Some relations have restrictions for one argument, some for both, and some do not have any restrictions.

A (u) in front of a relation means that no rules have been inferred for this relation, but it is part of the relation inventory for the given preposition.

A (d) in front of a relation means that this can be viewed as a default relation, because it is the most frequent relation for a given preposition.

Example entries

AF 'æresmedlem af Dansk Brygmester Forening'

INH
1st onto: 'Human'
 'Occupation'

This means that for a given instance of the preposition 'af' in a NP-PREP-NP construction, if the ontological type of the first NP is 'Human' or 'Occupation', then the relation is INH. No restrictions are put on the ontological type of the second NP. The corpus evidence 'æresmedlem af Dansk Brygmester Forening' exemplifies the relation INH.

GENNEM/IGENNEM

(d) **TMP** 'dansk kunst fra det 18. århundrede'
2nd onto: '3rdOrderEntity'
 'Purpose'

This means that for a given instance of the preposition 'gennem' or 'igennem' in a NP-PREP-NP construction, if the ontological type of the second NP is '3rdOrderEntity' or 'Purpose', then the relation is TMP. The flag (d) means that TMP is the most frequent relation for the preposition 'gennem/igennem'. No restrictions are put on the ontological type of the first NP.

FOR

LOC
1st onto: 'nord for Viborg'
 '3rdOrderEntity' *2nd onto:* 'Place'

This means that for a given instance of the preposition 'for' in a NP-PREP-NP construction, if the ontological type of the first NP is '3rdOrderEntity' and the ontological type of the second NP is 'Place', then the relation is LOC.

VED

(u) **BMO** 'formering ved tilfældig knopskydning'

Appendix A

This means that the preposition 'ved' can denote the relation BMO, but no rules have been inferred for this.

A Dictionary of Prepositions

AF

(d)	PNT	‘bearbejdning af nye, ukendte input’
	<i>1st onto:</i>	'1stOrderEntity' 'Artwork' 'Existence' 'LanguageRepresentation' 'UnboundedEvent'
	INH	‘æresmedlem af Dansk Brygmester Forening’
	<i>1st onto:</i>	'Human' 'Occupation'
	POF	‘en filial af Danske Bank’
	<i>1st onto:</i>	'Part'
	AGT	‘De tidlige sange af Alban Berg’
	<i>1st onto:</i>	'Property'
	WRT	‘lørdagens udgave af JyllandsPosten’
	<i>1st onto:</i>	'Communication'
(u)	TMP	‘udgangen af det 20. århundrede’
(u)	MEA	‘den beskedne sum af 30.000 kroner’
(u)	LOC	‘udkanten af Paris’
(u)	CMP	‘flokke af svaner’
(u)	COM	‘i skikkelse af høje, unge kvinder’
(u)	CBY	‘syge af Salmonella’
(u)	SRC	‘resultatet af forhandlingerne’
(u)	RCH	‘en djævelinde af en kone’
(u)	CHR	‘modeller af typen A’

Appendix A

(u) **RLO** 'et bæger af syre'

EFTER

(d) **TMP** 'et stykke tid efter VM'
1st onto: '3rdOrderEntity'
2nd onto: '3rdOrderEntity'
'BoundedEvent'
'Dynamic'
'Physical'
'Purpose'

QUA 'erstatning efter de gældende regler'
2nd onto: 'Property'

ADD 'den ene celle efter den anden'
1st onto: 'Artwork'

CBY 'graviditet efter et ubeskyttet samleje'
1st onto: 'BoundedEvent'

TAR 'det uendelige begær efter penge'
1st onto: 'Mental'

(u) **PNT** 'Den øgede interesse efter den ægte vare'
(u) **INH** 'enke efter arkitekt Preben Hansen'
(u) **COM** 'en kamp efter vestligt forbillede'
(u) **MTH** 'boligudgiften efter fradrag'
(u) **SRC** 'arven efter Nielsen'

FOR

(d) **INH** 'formand for Aarhus sejlklub'
1st onto: 'Human'
'Property'

	<i>2nd onto:</i>	'Dynamic'
WRT	<i>1st onto:</i>	'Landsforeningen for Autisme' '2ndOrderEntity' 'Animal' 'Building' 'Container' 'Dynamic' 'Part'
	<i>2nd onto:</i>	'UnboundedEvent' '3rdOrderEntity' 'Mental' 'Purpose' 'UnboundedEvent'
PNT	<i>1st onto:</i>	'dansk lobbyisme for Baltikum' 'LanguageRepresentation' 'BoundedEvent' 'Comestible' 'Purpose'
	<i>2nd onto:</i>	'Human'
MEA	<i>2nd onto:</i>	'en 28 dages kur for 200 kr' 'Location'
LOC	<i>1st onto:</i>	'nord for Viborg' '3rdOrderEntity'
	<i>2nd onto:</i>	'Place'
CMP	<i>1st onto:</i>	'en nystartet bogklub for sygeplejersker' 'Institution'
(u) TMP		'dronning for en dag'
QUA	<i>2nd onto:</i>	'Belønning for hårdt arbejde' 'Agentive'
(u) ADD		'dag for dag'

Appendix A

FRA

(d)	LOC		‘Michael Nielsen fra Bjerringbro’
		<i>1st onto:</i>	'Building' 'Human' 'Object'
		<i>2nd onto:</i>	'Artifact'
	SRC		‘opbakningen fra den borgerlige gruppe’
		<i>1st onto:</i>	'3rdOrderEntity' 'BoundedEvent' 'Dynamic' 'Mental' 'Purpose'
		<i>2nd onto:</i>	'Human'
	TMP		‘dansk kunst fra det 18. århundrede’
		<i>2nd onto:</i>	'3rdOrderEntity'
(u)	POF		‘blade fra en busk’
(u)	INH		‘en repræsentant fra Novo’
(u)	MEA		‘enhver distance fra 1500 m’
(u)	AGT		‘et flot sololøb fra Mads’
(u)	WRT		‘forskellene fra Bush’
(u)	CBY		‘det dunkle skær fra bålet’

GENNEM/IGENNEM

(d)	TMP		‘adskillige forhøjelser gennem året’
		<i>2nd onto:</i>	'3rdOrderEntity' 'Purpose'
	LOC		‘Boringer gennem Indlandsisen’
		<i>2nd onto:</i>	'Animal' 'Artifact' 'Building'

		'Dynamic' 'Human' 'Place'
BMO	<i>2nd onto:</i>	'kvælstof gennem kunstgødning' 'BoundedEvent' 'Comestible' 'Instrument' 'Physical' 'UnboundedEvent'
PNT	<i>2nd onto:</i>	'en vej gennem systemet' 'Container'
(u) SRC		'oplysninger gennem Interpol'
(u) CBY		'leukæmi gennem stråling'

HOS

(d) PNT	<i>1st onto:</i>	'Forsinket sårheling hos de 3 førstnævnte patientgrupper' 'BoundedEvent' 'Communication' 'Mental' 'Physical' 'Property' 'Purpose'
LOC	<i>1st onto:</i>	'En læreplads hos den internationalt berømte Georg Jensen' '3rdOrderEntity' 'Artifact' 'Comestible' 'Part' 'Social' 'UnboundedEvent'

Appendix A

	INH		‘behandlinger hos en fysioterapeut’
		<i>1st onto:</i>	‘Human’
(u)	AGT		‘en negativ reaktion hos tilhørerne’
(u)	WRT		‘den ansattes fremtidige stilling hos arbejdsgiveren’
(u)	SRC		‘trøst hos andre mænd’
(u)	RCH		‘den øgede økologiske bevidsthed hos forbrugerne’
(u)	POF		‘en afdeling hos Warner’

I

(d)	LOC		‘afdelingskontorerne i Skanderborg’
		<i>1st onto:</i>	‘1stOrderEntity’ ‘Artifact’ ‘Building’ ‘Group’ ‘Purpose’
		<i>2nd onto:</i>	‘Place’
		<i>1st onto:</i>	‘Object’
		<i>1st onto:</i>	‘Human’
		<i>2nd onto:</i>	‘Artifact’
		<i>2nd onto:</i>	‘Place’
	INH		‘afdelingschef i juridisk afdeling’
		<i>1st onto:</i>	‘Human’ ‘Occupation’
	TMP		‘affæren i efteråret’
		<i>2nd onto:</i>	‘3rdOrderEntity’
	PNT		‘en afdæmpet vækst i den amerikanske økonomi’
		<i>1st onto:</i>	‘3rdOrderEntity’ ‘Physical’

(u)	POF		‘medlemsstater i Den Europæiske Union’
	WRT	<i>1st onto:</i>	‘En nagende mistanke i det nye forhold’ 'Dynamic'
(u)	AGT		‘,’
	CHR	<i>1st onto:</i>	‘,’ 'BoundedEvent'
(u)	BMO		‘,’

MED

(d)	PNT	<i>1st onto:</i>	‘telefonsamtaler med Miguel’ '2ndOrderEntity' 'BoundedEvent' 'Communication' 'Dynamic' 'Location' 'Part' 'Property' 'Purpose' 'UnboundedEvent'
		<i>2nd onto:</i>	'Animal'
		<i>1st onto:</i>	'3rdOrderEntity' <i>2nd onto:</i> 'Human'
		<i>1st onto:</i>	'Social' <i>2nd onto:</i> 'Human'
	CHR	<i>1st onto:</i>	‘personer med nedsat arbejdsevne’ 'Artwork' 'Building' 'Container' 'Group' 'Institution' 'Object' 'Occupation'
		<i>2nd onto:</i>	'Building' 'Comestible'

Appendix A

		'Communication'
		'Dynamic'
		'Plant'
		'Property'
		'Purpose'
		'UnboundedEvent'
	<i>1st onto:</i>	'Human'
		<i>2nd onto:</i> '3rdOrderEntity'
CMP		'De fleste biler med 15 tommers fælde'
	<i>1st onto:</i>	'Comestible'
		'Instrument'
		'Vehicle'
	<i>2nd onto:</i>	'LanguageRepresentation'
WRT		'sager med somaliske ansøgere'
	<i>1st onto:</i>	'Experience'
RLO		'biler med tunesiske fans'
	<i>1st onto:</i>	'Furniture'
		'Underspecified'
BMO		'en flot fejende bevægelse med hånden'
	<i>1st onto:</i>	'Animal'
	<i>2nd onto:</i>	'Instrument'
(u) CUM		'En mor med en lille dreng'
(u) INH		'kontrol med levende dyr'
(u) CBY		'i sengen med feber'
(u) MNR		'deres liv med andre børn'
(u) MEA		'største procentvise fald med knap 20%'
(u) AGT		'pigtrådmusik med The Donkeys'

MELLEM/IMELLEM

(d) INH		'relationerne mellem de øvrige EU-lande'
	<i>1st onto:</i>	'3rdOrderEntity'
		'Artifact'
		'BoundedEvent'
		'Dynamic'

		'Physical'	
		'Purpose'	
		'Social'	
		'UnboundedEvent'	
	<i>2nd onto:</i>	'3rdOrderEntity'	
LOC		'vejen mellem Holbæk og Sjællands Odde'	
	<i>1st onto:</i>	'Building'	
		'Group'	
		'Object'	
	<i>2nd onto:</i>	'Artifact'	
	<i>1st onto:</i>	'Human'	<i>2nd onto:</i> 'Human'
(u) PNT		'samværet mellem børn og forældre'	
(u) WRT		'de store forskelle mellem de forskellige parceller'	
(u) MEA		'alle børn mellem to uger og et år'	
(u) TMP		'natten mellem lørdag og søndag'	

OVER

(d) LOC		'udsigt over hele Kattegat'	
	<i>1st onto:</i>	'2ndOrderEntity'	
		'Animal'	
		'Comestible'	
		'Container'	
		'Garment'	
		'Human'	
		'Instrument'	
		'Natural'	
	<i>2nd onto:</i>	'Artifact'	
		'Building'	
		'Natural'	
		'Place'	
		'Property'	
PNT		'en tilintetgørende dom over Schröder'	
	<i>1st onto:</i>	'BoundedEvent'	

Appendix A

		'Purpose'	
	<i>2nd onto:</i>	'Dynamic'	
		'Group'	
		'Plant'	
	<i>1st onto:</i>	'Place'	<i>2nd onto:</i> '3rdOrderEntity'
MEA		'danskere over 80 år'	
	<i>2nd onto:</i>	'3rdOrderEntity'	
		'Human'	
	<i>1st onto:</i>	'Human'	<i>2nd onto:</i> '3rdOrderEntity'
	<i>1st onto:</i>	'Occupation'	<i>2nd onto:</i> '3rdOrderEntity'
WRT		'et kort over London'	
	<i>1st onto:</i>	'Location'	
		'Physical'	
	<i>2nd onto:</i>	'LanguageRepresentation'	
TMP		'24 millioner kroner over en fireårig periode'	
	<i>1st onto:</i>	'Institution'	
		'Plant'	
		'Static'	
CBY		'sorg over det store antal dræbte'	
	<i>1st onto:</i>	'Place'	
	<i>1st onto:</i>	'Property'	<i>2nd onto:</i> 'Physical'
COM		'et pænt stykke over den bogførte værdi'	
	<i>1st onto:</i>	'Artwork'	
		'Colour'	
		'Object'	
	<i>2nd onto:</i>	'UnboundedEvent'	
INH		'et monument over to brødres skæbnesvangre samlermani'	
	<i>1st onto:</i>	'Building'	
(u) BMO		'bedre kommunikation over Internettet'	
(u) RCH		'stil over moden'	

PÅ

(d)	LOC		‘fem medarbejdere på 13. sal’
		<i>1st onto:</i>	'Building' 'Occupation'
		<i>2nd onto:</i>	'Artifact' 'Building' 'Furniture' 'Natural' 'Place' 'Purpose'
		<i>1st onto:</i>	'Artifact'
		<i>2nd onto:</i>	'Human'
		<i>1st onto:</i>	'Object'
		<i>2nd onto:</i>	'Human'
	MEA		‘en lønforhøjelse på 100 kr’
		<i>1st onto:</i>	'3rdOrderEntity' 'Artifact' 'Dynamic' 'Part' 'Property'
		<i>1st onto:</i>	'Human'
		<i>2nd onto:</i>	'3rdOrderEntity'
	PNT		‘stor indflydelse på blodets indhold af stoffet homocystein’
		<i>1st onto:</i>	'Physical' 'Purpose'
		<i>2nd onto:</i>	'Plant'
		<i>1st onto:</i>	'UnboundedEvent'
		<i>2nd onto:</i>	'Human'
	WRT		‘valgmulighederne på andre områder’
		<i>2nd onto:</i>	'Dynamic' 'Physical'
(u)	INH		‘chefredaktør på Der Spiegel’
	POF		‘stroppe på badedragten’
		<i>2nd onto:</i>	'Garment' 'LanguageRepresentation'

Appendix A

	TMP		'den mest romantiske dag på hele året'
		<i>1st onto:</i>	'Imagerepresentation'
(u)	MNR		'blikkenslagere på akkord'
	CMP		'Front Nationals gruppe på 11 medlemmer'
		<i>1st onto:</i>	'Group'
(u)	BMO		'handel med varer på computeren'

TIL

(d)	PNT		'Forfatteren til bogen'
		<i>1st onto:</i>	'3rdOrderEntity' 'Mental'
(u)	LOC		'dyre rejser til New York'
(u)	WRT		'tid til en gåtur'
(u)	PRP		'midler til både olie og lønninger'
(u)	INH		'nedtællingen til et historisk vendepunkt'
(u)	TMP		'11. marts til 10. April'
(u)	TAR		'uddannelsen til speciallæge'
(u)	RST		'Isminurs forvandling til Lone'
(u)	MEA		'verdens dyreste frimærke til 13 millioner kroner'

UNDER

(d)	TMP		'løn under barsel'
		<i>2nd onto:</i>	'BoundedEvent' 'Communication' 'Dynamic' 'Purpose' 'UnboundedEvent'
		<i>1st onto:</i>	'3rdOrderEntity'
		<i>2nd onto:</i>	'3rdOrderEntity'

	LOC		‘det lille bord under vinduet’
		<i>1st onto:</i>	'Animal' 'Artifact' 'Building' 'Comestible' 'Container' 'Part' 'Substance'
		<i>2nd onto:</i>	'Artifact' 'Container' 'Furniture'
		<i>1st onto:</i>	'Object'
		<i>2nd onto:</i>	'Building'
(u)	MNR		‘vin under kontrollerede former’
	SUP		‘et Europa under tysk herredømme’
		<i>1st onto:</i>	'UnboundedEvent'
		<i>2nd onto:</i>	'Human'
	MEA		‘fonde med formuer under fem millioner’
		<i>1st onto:</i>	'Human'
		<i>2nd onto:</i>	'3rdOrderEntity'
	POF		‘Et udvalg under justitsministeriet’
		<i>1st onto:</i>	'3rdOrderEntity' 'Institution'
(u)	COM		‘et enkelt slag under par’
(u)	QUA		‘sin pligt under sædvaneretten’
(u)	PNT		‘vel meget fut under økonomien’

VED

(d)	LOC		‘en dame ved skranken’
		<i>1st onto:</i>	'Building' 'Container' 'Group' 'Natural'

Appendix A

		'Place'
	<i>2nd onto:</i>	'Artifact'
		'Container'
		'Place'
	<i>1st onto:</i>	'Natural'
		'3rdOrderEntity' <i>2nd onto:</i> 'Building'
MNR		'pæne ord ved festlige lejligheder'
	<i>1st onto:</i>	'3rdOrderEntity'
		'Communication'
WRT		'flere ulemper ved en model'
	<i>1st onto:</i>	'Mental'
		'Property'
INH		'deltagelsen ved OL'
	<i>1st onto:</i>	'Occupation'
PNT		'start ved et stævne'
	<i>1st onto:</i>	'Artifact'
		'Part'
(u) BMO		'formering ved tilfældig knopskydning'
(u) POF		'kvartfinalerne ved VM'
(u) CBY		'skade ved branden'
(u) TMP		'ni procent ved overenskomstperiodens udløb'
(u) AGT		'en helt ny oversættelse ved mag. art. Ole Vesterholt'
(u) COM		'en gammel græker ved navn Iannis'
(u) MEA		'en forvarmet ovn ved 225 grader'
(u) RCH		'den gode stemning ved Gaimars hof'

Bibliography

- Agirre, E., & Martinez, D. (2001). Learning class-to-class selectional preferences. 2005
- Allan, R., Holmes, P., & Lundskaer-Nielsen, T. (1995). *Danish - A Comprehensive Grammar*. London: Routledge.
- An, A., & Cercone, N. (2000). *Rule Quality Measures Improve the Accuracy of Rule Induction: An Experimental Approach*. Paper presented at the Proceedings of the 12th International Symposium on Foundations of Intelligent Systems.
- Andreasen, T., Bulskov, H., Jensen, P., & Lassen, T. (2009). Conceptual Indexing of Text Using Ontologies and Lexical Resources. In *Flexible Query Answering Systems* (Vol. Volume 5822/2009, pp. 323-332): Springer Berlin / Heidelberg.
- Andreasen, T., Bulskov, H., Lassen, T., Zambach, S., Jensen, P. A., Madsen, B. N., et al. (2009). *SIABO: Semantic Information Access through Biomedical Ontologies*. Paper presented at the KEOD 2009 – First International Conference on Knowledge Engineering and Ontology Development, Madeira, Portugal.
- Andreasen, T., Jensen, P. A., Fischer Nilsson, J., Paggio, P., Sandford Pedersen, B., & Erdman Thomsen, H. (2002). Ontological Extraction of Content for Text Querying. In *Lecture Notes in Computer Science* (Vol. 2553, pp. 123 -136): Springer-Verlag.
- Andreasen, T., Jensen, P. A., Fischer Nilsson, J., Paggio, P., Sandford Pedersen, B., & Erdman Thomsen, H. (2004). Content-based text

Bibliography

- querying with ontological descriptors. *Data & Knowledge Engineering*, 48(2), 199-219.
- Becker-Christensen, C. (Ed.) (2005) Politikens Nudansk ordbog (19. udgave ed.). København: Politiken.
- Becker-Christensen, C., & Widell, P. (2003). *Nudansk Grammatik*. København: Politikens Forlag A/S.
- Brøndal, V. (1928). *Ordklasserne*. København: Gad.
- Brøndal, V. (1940). *Præpositionernes Theori - Indledning til en rationel Betydningslære*. København: Københavns Universitet.
- Chen, J. Y., & Lonardi, S. (Eds.). (2009). *Biological Data Mining* Chapman & Hall/CRC
- Chomsky, N. (1957). *Syntactic structures*. The Hague :: Mouton.
- Christensen, I. (2000). Silken, rummet, sproget, hjertet. In *Hemmelighedstilstanden*. København: Gyldendal.
- Clark, P., Fellbaum, C., Hobbs, J., , , & (2008). *Using and Extending WordNet to Support Question-Answering*. Paper presented at the Fourth Global WordNet Conference (GWC'08) from <http://www.cs.utexas.edu/users/pclark/papers/gwa08-extending-wordnet.pdf>
- Clark, P., Fellbaum, C., Hobbs, J. R., Harrison, P., Murray, W. R., & Thompson, J. (2008). *Augmenting WordNet for deep understanding of text*. Paper presented at the Proceedings of the 2008 Conference on Semantics in Text Processing.
- Cohen, W. (1995). *Fast Effective Rule Induction*. Paper presented at the In Proceedings of the Twelfth International Conference on Machine Learning.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.

- Cruse, D. A. (2002). Hyponymy and its Varieties. In R. Green, C. A. Bean & S. H. Myaeng (Eds.), *The semantics of relationships : an interdisciplinary perspective*. Dordrecht: Kluwer Academic Publishers.
- Dahlerup, V. (1918-56). *Ordbog over det Danske Sprog* København: Det Danske Sprog- og Litteraturselskab.
- DanNet. (2010). DanNet - det danske wordnet. Retrieved February, 2010, from <http://wordnet.dk>
- Dean, P., & Famili, A. (1997). Comparative Performance of Rule Quality Measures in an InductionSystem. *Applied Intelligence*, 7(2), 113-124.
- Descartes, R. (1903). *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences*
Arc Manor.
- Di Eugenio, B., & Glass, M. (2004). The kappa statistic: a second look. *Comput. Linguist.*, 30(1), 95-101.
- Diderichsen, P. (1946). *Elementær Dansk Grammatik* (3rd ed.). København: Gyldendal.
- Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3), 547-615.
- DSL. (2010). Den Danske Ordbog - ordnet.dk (Publication. Retrieved February, 2010, from Det Danske Sprog- og Litteraturselskab: <http://ordnet.dk/ddo>
- Dyvik, H. (1998). A translational basis for semantics. In S. Johansson & S. Oksefjell (Eds.), *Corpora and Crosslinguistic Research: Theory, Method and Case Studies* (pp. pp. 51-86). Amsterdam: Rodopi B. V.
- Eibe, F., & Witten, I. H. (1998). *Generating Accurate Rule Sets Without Global Optimization*. Paper presented at the Proceedings of the Fifteenth International Conference on Machine Learning.

Bibliography

- Fausto, G., Biswanath, D., & Vincenzo, M. (2009). Faceted Lightweight Ontologies. In *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos* (pp. 36-51): Springer-Verlag.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*: The MIT Press.
- Fellbaum, C., & Miller, G. A. (2003). Morphosemantic Links in WordNet. *Traitement automatique des langues*, vol. 44(2), p. 69–80.
- Fillmore, C. J. (1968). The Case for Case. *Universals in Linguistic Theory*, pp. 1-88.
- Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280, 20-32.
- Freitas, A. A. (1999). On rule interestingness measures. *Knowledge-Based Systems*, 12(5/6), 309-315.
- Genesereth, M. R., & Nilsson, N. J. (1987). *Logical foundations of artificial intelligence*: Morgan Kaufmann Publishers Inc.
- Gildea, D., & Jurafsky, D. (2000). *Automatic labeling of semantic roles*. Paper presented at the Proceedings of the 38th Annual Meeting on Association for Computational Linguistics.
- Graff, C. (Ed.) (2009) *Den Satiriske Encyklopædi*. Lyngby: StemningsHotellet.dk.
- Greenbaum, S., & Quirk, R. (1990). *A student's grammar of the English language*. London: Longman.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2), 199-220.
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(4-5), 907-928.

- Guarino, N. (1997). *Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration*. Paper presented at the International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology.
- Guarino, N. (1998). *Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*: IOS Press.
- Guarino, N., & Giaretta, P. (Eds.). (1995). *Ontologies and Knowledge Bases: Towards a Terminological Clarification*. Amsterdam: IOS Press.
- Guarino, N., & Welty, C. A. (2000). *A Formal Ontology of Properties*. Paper presented at the Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management.
- Hansen, E., & Heltoft, L. (2003). Grammatik over det Danske Sprog, Kap 2 Ordklasserne (pp. 97 pages). Roskilde: Kompendium.
- Hjort, E., & Kristensen, K. (Eds.). (2003-5). København: Det Danske Sprog- og Litteraturselskab/Gyldendal.
- Hjorth, E., & Kristensen, K. (Eds.). (2003-2005) Den Danske Ordbog. København: Det Danske Sprog- og Litteraturselskab/Gyldendal
- Ide, N., & Véronis, J. (1998). Special issue on word sense disambiguation: Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24.
- Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, Mass: MIT Press.
- Jackendoff, R. (1990). *Semantic structures*. Cambridge, Mass: MIT Press.
- Jensen, P. A. (1985). *Principper for grammatisk analyse*. København: Nyt Nordisk Forlag.

Bibliography

- Jensen, P. A., & Fischer Nilsson, J. (2006). Ontology-Based Semantics for Prepositions. In *Syntax and Semantics of Prepositions* (Vol. 29): Springer.
- Jensen, P. A., & Vikner, C. (2006). Leksikalsk semantik og omverdensviden. In A. Braasch (Ed.), *Sprogteknologi i dansk perspektiv : En samling artikler om sprogforskning og automatisk sprogbehandling* (pp. 229-248). København: C.A.Reitzel.
- Jespersen, O. (1924). *The Philosophy of Grammar*. London: George Allen & Unwin Ltd.
- Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3), 637-649.
- Kilgarriff, A. (2000). Review of wordnet : An electronic lexical database. *Language Resources and Evaluation*, 76, 3.
- Kipper-Schuler, K. (2006). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Unpublished PhD thesis, University of Pennsylvania, Philadelphia.
- Lassen, T. (2006). *An Ontology Based View on Prepositional Senses*. Paper presented at the Third ACL-SIGSEM Workshop on Prepositions, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento Italy.
- Lassen, T., & Terney, T. V. (2006a). *An Ontology-Based Approach to Disambiguation of Semantic Relations*. Paper presented at the Learning Structured Information in Natural Language Applications, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy.
- Lassen, T., & Terney, T. V. (2006b). *Ontology-based Disambiguation of the Semantic Relation Between the Heads of Two Noun Phrases*. Paper presented at the The 19th International FLAIRS Conference, Melbourne, Florida.

- Lassila, O., & McGuinness, D. (2001). The role of frame-based representation on the semantic web.
- Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., et al. (2000). SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13(4), 249-263.
- Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., et al. (1999). Linguistic Specifications, *SIMPLE Deliverable D2.1: ILC* and University of Pisa.
- Levin, B. (1993). *English Verb Classes And Alternations: A Preliminary Investigation*: The University of Chicago Press.
- Litkowski, K. C., & Hargraves, O. (2005). *The Preposition Project*. Paper presented at the ACL-SIGSEM Workshop on “The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications”, University of Essex - Colchester, United Kingdom.
- Litkowski, K. C., & Hargraves, O. (2007). *Word-Sense Disambiguation of Prepositions*. Paper presented at the The Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic.
- Locke, J. (1690). *An essay concerning human understanding*.
- Lund, J. (Ed.). (1994-2002). *Den store Danske Encyklopædi*. København: Danmarks Nationalleksikon A/S.
- Lyons, J. (1977). *Semantics* (Vol. 2). Cambridge: Cambridge University Press.
- Madsen, B., & Thomsen, H. (2008). *Terminological Principles used for Ontologies*. Paper presented at the Terminology and Knowledge Engineering 2008 (TKE'08).

Bibliography

- Madsen, B. N., Pedersen, B. S., & Thomsen, H. E. (2000). Semantic Relations in Content-based Querying Systems: a Research Presentation from the OntoQuery Project. *Ontologies and Lexical Knowledge Bases. Proceedings of the 1st International Workshop, OntoLex 2000*.
- Madsen, B. N., Pedersen, B. S., & Thomsen, H. E. (2001). Defining Semantic Relations for OntoQuery. *Proceedings of the First International OntoQuery Workshop Ontology-based interpretation of NP's*.
- Madsen, B. N., & Thomsen, H. E. (2006). *Terminological ontologies in normative terminology work*. Paper presented at the International Conference on Terminology, Standardization and Technology Transfer, TSTT'06.
- Madsen, B. N., & Thomsen, H. E. (2009). Ontologies vs. classification systems.
- Madsen, B. N., Thomsen, H. E., & Vikner, C. (2004). *Comparison of Principles Applying to Domain Specific versus General Ontologies*. Paper presented at the OntoLex 2004: Ontologies and Lexical Ressources in Distributed Environments.
- Madsen, B. N., Thomsen, H. E., & Vikner, C. (2004). *Principles of a system for terminological concept modelling*. Paper presented at the The 4th International Conference on Language Resources and Evaluation.
- Madsen, B. N., Thomsen, H. E., & Vikner, C. (2005). *Multidimensionality in terminological concept modelling*. Paper presented at the Terminology and Content Development, TKE 2005, 7th International Conference on Terminology and Knowledge Engineering
- Mikkelsen, K. (1911). *Dansk Ordføjningslære*. København: Hans Reitzels Forlag.

- Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database*. *Int J Lexicography*, 3(4), 235-244.
- Miller, G., & Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation*, 41(2), 209-214.
- Miller, G. A., & Hristea, F. (2006). WordNet Nouns: Classes and Instances. *Computational Linguistics*, 32(1), 1-3.
- Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., et al. (1991). Enabling technology for knowledge sharing. *AI Mag.*, 12(3), 36-56.
- Nielsen, J. L. (1995). En syntaktisk og semantisk undersøgelse af præpositionen. *Odense Working Papers in Language and Communication*, no. 9.
- Nilsson, J. F. (1999). Ontological Typing of Natural Language Phrases. Unpublished Incomplete working draft. DTU.
- Nilsson, J. F. (2001). *A Logico-Algebraic Framework for Ontologies, ONTOLOG*. Paper presented at the The First International OntoQuery Workshop.
- Nilsson, J. F., & Jensen, P. A. (2003). *Ontology-based Semantics for Prepositions*. Paper presented at the {ACM-SIGSEM} Workshop on the Linguistic Dimension of Prepositions and their use in Computational Linguistics, Toulouse, France.
- O'Hara, T., & Wiebe, J. (2009). Exploiting semantic role resources for preposition disambiguation. *Comput. Linguist.*, 35(2), 151-184.
- Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics*, 31(1).
- Palmer, M., Gildea, D., & Xue, N. (2010). *Semantic Role Labeling* (Vol. 6): Morgan & Claypool.

Bibliography

- Pedersen, B., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3), 269-299.
- Pedersen, B., & Sørensen, N. H. (2006). *Towards Sounder Taxonomies in Wordnets*. Paper presented at the OntoLex 2006 Workshop, Genoa, Italy.
- Pedersen, B. S. (1999). *Den Danske SIMPLE-ordbog - En semantisk, ontologibaseret ordbog*. København: Center for sprogteknologi.
- Pedersen, B. S. (2009). *DanNet - A Network of Words*. Paper presented at the DanNet Symposium. from http://wordnet.dk/dannet/dannet/DanNet_symposium2009_slides_pedersen.pdf.
- Pedersen, B. S., Braasch, A., Nimb, S., Asmussen, J., Sørensen, N., Lorentzen, H., et al. (2009). *Lingvistiske specifikationer for DanNet Version 1.0*. København: DanNet.
- Pedersen, B. S., & Paggio, P. (2004). The Danish SIMPLE Lexicon and its Application in Content-based Querying. *Nordic Journal of Linguistics*, 27(1), 97-127.
- Pedersen, V. H. (Ed.) (1999) *Dansk-Engelsk Ordbog - Vinterberg & Bodelsen* (4. udgave ed.). København: Gyldendal.
- Princeton_University. (2010). WordNet. Retrieved February 20 2010, from <http://wordnet.princeton.edu>
- Pustejovsky, J. (1991a). The generative lexicon. *Computational Linguistics*, 17(4), 409-441.
- Pustejovsky, J. (1991b). The syntax of event structure. *Cognition*, 41(1-3), 47-81.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382-439.
- Saussure, F. D. (1983). *Course in General Linguistics* (R. Harris, Trans.): Duckworth.
- Schwarz, H. (Ed.) (2007) Dansk-Engelsk ordbog over præpositioner (1. udgave ed.). København: Handelshøjskolens Forlag.
- Spang-Hanssen, E. (1996). Sprog og betydningsindlæring, Informations- og Dokumentationscenteret for Fremmedsprogspædagogik ved Danmarks Pædagogiske Bibliotek og Foreningen for Anvendt Sprogvidenskab i Danmark (ADLA). *Sprogforum, Vol. 2*(ekstra), pp 27-31.
- Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge Engineering: Principles and methods. *Data and Knowledge Engineering*, 25, 161-197.
- Swartout, B., Ramesh, P., Knight, K., & Russ, T. (1997). *Toward Distributed Use of Large-Scale Ontologies*. Paper presented at the AAAI Symposium on Ontological Engineering.
- Swift, J. (1726). *Gulliver's Travels*.
- Terney, T. V. (2009). *The Combined Usage of Ontologies and Corpus Statistics in Information Retrieval*. Roskilde University, Roskilde.
- Togebj, O. Stiltræk. Retrieved September, 2005, from <http://www.hum.au.dk/dk/nordisk/norot/stiltrak.htm>
- Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca, N.Y.: Cornell University Press.

Bibliography

- Verkuyl, H. J. (1972). *On the Compositional Nature of the Aspects*: Reidel.
- Verkuyl, H. J. (1989). Aspectual classes and aspectual composition. *Linguistics and Philosophy*, 12(1).
- Vikner, C., & Jensen, P. A. (2002). A Semantic Analysis of the English Genitive Interaction of Lexical and Formal Semantics. *Studia Linguistica*, 56(2).
- Voorhees, E. M. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. Pittsburgh, PA, USA, June 27 - July 1, 1993, 171-180.
- Vossen, P. (2001). EuroWordNet Website. Retrieved October, 2009, from <http://www.illc.uva.nl/EuroWordNet>
- Vossen, P., Bloksma, L., Rodriguez, H., Climent, S., Calzolari, N., Roventini, A., et al. (1998). *The EuroWordNet Base Concepts and Top Ontology*: Centre National de la Recherche Scientifique.
- Vossen, P., Díez-Orzas, P., & Peters, W. (1997). *Multilingual design of EuroWordNet*. Paper presented at the Automatic Information Extraction and Building of Lexical Semantic Resources, Workshop at ACL/EACL-97.
- W3C. (2004). OWL Web Ontology Language Overview. Retrieved February, 2010, 2010, from <http://www.w3.org/TR/owl-features/>
- Winston, M. E., Chaffin, R., & Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11(4), 417-444.
- Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd Edition ed.). San Francisco: Morgan Kaufmann.
- Yarowsky, D. (1992). Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora. *Proceedings of COLING-92*, 454-460.

Bibliography

Zwarts, J. (1997). Vectors as relative positions: A compositional semantics of modified PPs. *Journal of Semantics*, 14, 57-86.

RECENT RESEARCH REPORTS

- #130 Gourinath Banda. *Modelling and Analysis of Real Time Systems with Logic Programming and Constraints*. PhD thesis, Roskilde, Denmark, August 2010.
- #129 Maren Sander Granlien. *Participation and Evaluation in the Design of Healthcare Work Systems — A participatory design approach to organisational implementation*. PhD thesis, Roskilde, Denmark, April 2010.
- #128 Thomas Bolander and Torben Braüner, editors. *Preliminary proceedings of the 6th Workshop on Methods for Modalities (M4M-6)*, Roskilde, Denmark, October 2009.
- #127 Leopoldo Bertossi and Henning Christiansen, editors. *Proceedings of the International Workshop on Logic in Databases (LID 2009)*, Roskilde, Denmark, October 2009.
- #126 Thomas Vestskov Terney. *The Combined Usage of Ontologies and Corpus Statistics in Information Retrieval*. PhD thesis, Roskilde, Denmark, August 2009.
- #125 Jan Midtgaard and David Van Horn. Subcubic control flow analysis algorithms. 32 pp. May 2009, Roskilde University, Roskilde, Denmark.
- #124 Torben Braüner. Hybrid logic and its proof-theory. 318 pp. March 2009, Roskilde University, Roskilde, Denmark.
- #123 Magnus Nilsson. *Arbejdet i hjemmeplejen: Et etnometodologisk studie af IT-støttet samarbejde i den københavnske hjemmepleje*. PhD thesis, Roskilde, Denmark, August 2008.
- #122 Jørgen Villadsen and Henning Christiansen, editors. *Proceedings of the 5th International Workshop on Constraints and Language Processing (CSLP 2008)*, Roskilde, Denmark, May 2008.
- #121 Ben Schouten and Niels Christian Juul, editors. *Proceedings of the First European Workshop on Biometrics and Identity Management (BIOID 2008)*, Roskilde, Denmark, April 2008.
- #120 Peter Danholt. *Interacting Bodies: Posthuman Enactments of the Problem of Diabetes Relating Science, Technology and Society-studies, User-Centered Design and Diabetes Practices*. PhD thesis, Roskilde, Denmark, February 2008.
- #119 Alexandre Alapetite. *On speech recognition during anaesthesia*. PhD thesis, Roskilde, Denmark, November 2007.
- #118 Paolo Bouquet, editor. *CONTEXT'07 Doctoral Consortium Proceedings*, Roskilde, Denmark, October 2007.