Lilja Øvrelid

Argument Differentiation

Data linguistica

<http://hum.gu.se/institutioner/svenska-spraket/publ/datal/>

Editor: Lars Borin

Språkbanken • Språkdata
Department of Swedish Language
University of Gothenburg

20 • 2008

Lilja Øvrelid

Argument Differentiation

Soft constraints and data-driven models

Gothenburg 2008

ISBN 978-91-87850-35-6
ISSN 0347-948X

Printed in Sweden by
Intellecta Docusys Västra Frölunda 2008

Typeset in LaTeX 2ε by the author

Cover design by Kjell Edgren, Informat.se

Front cover illustration:
How to describe the world is still an open question
by Randi Nygård ©

Author photo on back cover by Rudolf Rydstedt

ABSTRACT

The ability to distinguish between different types of arguments is central to syntactic analysis, whether studied from a theoretical or computational point of view. This thesis investigates the influence and interaction of linguistic properties of syntactic arguments in *argument differentiation*. Cross-linguistic generalizations regarding these properties often express probabilistic, or soft, constraints, rather than absolute requirements on syntactic structure. In language data, we observe frequency effects in the realization of syntactic arguments.

We propose that argument differentiation can be studied using data-driven methods which directly express the relationship between frequency distributions in language data and linguistic categories. The main focus in this thesis is on the formulation and empirical evaluation of linguistically motivated features for data-driven modeling. Based on differential properties of syntactic arguments in Scandinavian language data, we investigate the linguistic factors involved in argument differentiation from two different perspectives.

We study automatic acquisition of the lexical semantic category of animacy and show that statistical tendencies in argument differentiation supports automatic classification of unseen nouns. The classification is furthermore robust, generalizable across machine learning algorithms, as well as scalable to larger data sets.

We go on to perform a detailed study of the influence of a range of different linguistic properties, such as animacy, definiteness and finiteness, on argument disambiguation in data-driven dependency parsing of Swedish. By including features capturing these properties in the representations used by the parser, we are able to improve accuracy significantly, and in particular for the analysis of syntactic arguments.

The thesis shows how the study of soft constraints and gradience in language can be carried out using data-driven models and argues that these provide a controlled setting where different factors may be evaluated and their influence quantified. By focusing on empirical evaluation, we come to a better understanding of the results and implications of the data-driven models and furthermore show how linguistic motivation in turn can lead to improved computational models.

ACKNOWLEDGEMENTS

This thesis has been a big part of my life for several years and to think that it is actually finished now is truly beyond my grasp. I do know, however, that there are numerous people who have helped and supported me and whom it is my undivided pleasure to thank.

I want to express my gratitude to my two supervisors, Elisabet Engdahl and Joakim Nivre. Elisabet welcomed me to Gothenburg over four years ago and has since then been a person to be counted with in my life. She has provided advice and pointed criticism on all aspects of my work, made me think and rethink linguistic issues small and large and pushed me to move on when I was frozen. Thank you for your enthusiasm and interest, for truly caring, for always making time, for reading into the last hours, and for being such an open-minded, outstanding linguist. Joakim has been involved almost from the very beginning and has provided invaluable insight and inspiration in the writing of this thesis. Thank you so much for taking time out of your busy schedule, for always showing a genuine interest in my work, for your clarity of thought, formal expertise and for new ideas. Thank you both for believing in me when I did not!

There are several other people who have read and commented on parts of this thesis along the way and whom I would like to give my warmest thanks to: Maia Andréasson, Harald Hammarström, Fredrik Heinat, Helen de Hoop, Jerker Järborg, Ida Larsson, Benjamin Lyngfelt, Malin Petzell and Annie Zaenen. A special thanks to Beáta Megyesi for scrutinizing a first draft of this thesis for my final seminar and providing very useful comments.

I want to thank Helen de Hoop, Monique Lamers, Peter de Swart, Sander Lestrade and everyone in the PIONIER project at Radboud University, Nijmegen for welcoming me as a guest researcher and for sharing thoughts on the ever-fascinating topic of animacy. I would also like to thank Gemma Boleda for discussions about classification, Ryan McDonald for advice on the MST-Parser experiments and Johan Hall for help with MaltTagger.

No woman is an island and I have been fortunate to be part of several stimulating research environments. I would like to thank the Graduate School of Language Technology (GSLT) for providing top-quality courses and an inspiring setting in which to meet fellow PhD-students and senior researchers and

discuss and get feedback. I have benefited immensely from being a part of GSLT. A special thanks to Atelach Alemu and Karin Cavallin for a most memorable trip to Tuscany, Eva Forsbom for discussions on annotation, to Ebba Gustavii, with whom I started exploring dependency parsing, and to Harald Hammarström, Hans Hjelm, Maria Holmqvist, Svetoslav Marinov and all the other PhD-students for all the good times. In Gothenburg I have had the pleasure of being part of the NLP-unit at the Dept. of Swedish as well as the newly started Center for Language Technology (CLT). I want to express a big thanks to Lars Borin for the work he has spent editing my thesis and for being such a friendly boss, Dimitrios Kokkinakis for being so helpful and letting me use his eminent suite of Swedish NLP-tools, to Rudolf Rydstedt for letting me take up a lot of disk space and for help with photography, to Robert Andersson for all technical assistance and to Dana Dannells, Karin Friberg, Jerker Järborg, Sofie Johansson-Kokkinakis, Leif-Jöran Olsson, Torgny Rasmark, Maria Toporowska-Gronostaj, Karin Warmenius and everyone else at Språkdata for being such a great group of colleagues. At the Dept. of Swedish, I also want to give a special thanks to the members of the OT reading group for inspiring discussions about linguistics.

Moving to Gothenburg from Oslo, I could never have asked for better colleagues, who soon became close friends. Annika Bergström, Ida Larsson and Karin Cavallin, thank you for your endless support and friendship. I want to give a very special thanks to Ida for giving me daily doses of porridge, perfect matters and perspective on thesis-writing, linguistics and life in general. I want to thank my fabulous friends in Oslo, Madrid and New York for keeping me grounded. Thanks to Randi Nygård for letting me use her lovely drawing on the cover of this book. I want to extend the warmest thanks possible to my dear family for all the love and support through what has been a life-altering time. And finally, Fredrik, I could have written this thesis without you, but I certainly would not have wanted to.

Thank you all!

Lilja Øvrelid
Gothenburg, April 20th, 2008

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

The main goal of syntactic analysis is often bluntly summarized as figuring out "who does what to whom?" in natural language. At the core of this simplification, however, is the idea that central to the understanding of a natural language sentence is the understanding of the predicate-argument structure which it expresses, and, in particular, the syntactic relationship which holds between the predicate and its individual arguments. The study of the relationship between meaning and form, how the syntactic expression of a certain semantic proposition precisely reflects the meaning which we wish to convey, can be seen to unite current syntactic theories. In the field of computational linguistics, syntactic parsing constitutes a central topic, where the main focus is on the automatic assignment of syntactic structure to natural language. The relation between syntax and semantics is furthermore exploited in work on automatic acquisition of lexical semantics, where the syntactic distribution of an element is seen as indicative of certain semantic properties. In psycholinguistics, the understanding of how we as language users perform this mapping in real-time comprehension has been widely studied. The study of *argument differentiation* focuses on the distinguishing properties of syntactic arguments which are central to syntactic analysis, whether studied from a theoretical, experimental or computational point of view. This is the central topic of this thesis.

1.1 Argument differentiation

Syntactic arguments express the main participants in an event, hence are intimately linked to the semantics of a sentence. Syntactic arguments also occur in a specific discourse context where they convey linguistic information. For instance, the subject argument often expresses the agent of an action, hence will tend to refer to a human being. Moreover, subjects typically express the topic of the sentence and will tend to be realized by a definite nominal. These types of generalizations regarding the linguistic properties of syntactic arguments express probabilistic, or 'soft', constraints, rather than absolute requirements

on syntactic structure. In language data, we observe frequency effects in the realization of syntactic arguments and a range of linguistic studies emphasize the correlation between syntactic function and various linguistic properties, such as animacy and definiteness. These properties are recurring also in cross-linguistic studies where they determine argument differentiation to varying degrees in different languages.

The realization of a predicate-argument structure is furthermore subject to surface-oriented and often language-specific restrictions relating to word order and morphology. In many languages, the structural expression of syntactic arguments exhibits variation. The Scandinavian languages, for instance, are characterized by a rigid verb placement and a certain degree of variation in the positioning of syntactic arguments. Work in syntactic theory which separates the function-argument structure from its structural realization highlights exactly the mediating role of arguments between semantics and morphosyntax.

An understanding of the influence of different linguistic factors and their interaction in argument differentiation clearly calls for a principled modeling of soft constraints and the frequency effects which these incur in language data. Semantic properties of verbs and their relation to syntactic realization have been given much attention both in theoretical and computational linguistic studies. The central status of the predicate as syntactic head, selecting and governing its arguments, is hardly under dispute. However, a focus on linguistic properties of syntactic arguments is important, both from a theoretical and a more practical or applied point of view. The study of properties of arguments and their influence in argument differentiation highlights cross-linguistic tendencies in the relation between syntax and semantics. It furthermore raises theoretically relevant questions regarding the modeling of these insights, the interaction between levels of linguistic analysis and the relation between theoretical results and practical applications.

1.2 Data-driven models

Recent decades have witnessed an empirical shift in the field of computational linguistics. New types and quantities of data have enabled new types of generalizations, and empirical, data-driven models are by now widely used. A defining property of these models is found in the systematic combination and weighting of different sources of evidence. In the processing of natural language, the ability to generalize over complex interrelationships has provided impressive results for a range of different NLP tasks.

A central theorem in machine learning theory emphasizes the fact that all learning requires a bias, that is, the learning problem must be defined in such a

way as to make generalization possible. Different machine learning algorithms come with different biases and an understanding of the way in which the search for the most likely hypothesis is performed is important in order to understand the results. Moreover, in order for learning to take place, the input data must be represented in such a way as to capture useful distinctions. The selection of features employed in the representation of the training data can have dramatic effects on results.

There exists a pronounced interest in a deeper understanding of the results obtained using data-driven methods and how these relate to generalizations from more theoretically oriented work. Empirical methods have gained momentum also in theoretical linguistics in recent years, where important insights revolve around the role and theoretical interpretation of language data and the modeling thereof. The exchange of insights and results constitutes an important step for further advancement of the study of natural language processing and linguistics in general. It is clear, however, that such an understanding requires an understanding of the data-driven models themselves as well as the implications of various representational choices. In the modeling of natural language, it is certainly not always the case that the most linguistically informed system is also the best performing system. Data-driven models, largely being probabilistic, furthermore have a reputation for being chaotic and difficult to interpret. In this respect, theoretically motivated hypotheses regarding linguistic analysis may provide a clarifying perspective.

1.3 Modeling argument differentiation

In this thesis, we propose that argument differentiation should be studied using data-driven methods which highlight the direct relationship between frequency distributions in language data and linguistic categories. The commitment is strictly empirical in that we will not explicitly formulate a set of constraints or a grammar for the interpretation of syntactic arguments. Rather, the focus will be on an explicit formulation and evaluation of a learning bias in terms of linguistically motivated features and evaluation of these. We will investigate the linguistic factors involved in argument differentiation, from two different perspectives, both highlighting different aspects of syntactic argumenthood and the relation between linguistic theory and model.

Animacy is a linguistic property which has been claimed to be an important factor in argument differentiation both in cross-linguistic studies and in psycholinguistic work. If this assumption is correct, we may hypothesize that differentiated arguments should provide important clues with respect to the property of animacy. In this thesis, we will investigate lexical acquisition of

animacy information based on syntactic, distributional features. By generalizing over the syntactic distribution of individual noun tokens, we may study linguistic properties of syntactic arguments irrespective of their specific realization in a particular sentence. In this way we may capture empirical frequency effects in the mapping between syntax and semantics. Through the application and evaluation of data-driven machine learning methods, we will investigate theoretical claims regarding the relationship between syntactic arguments and the property of animacy, as well as the robustness and reliability of such correlations. The focus is thus on the relation of syntactic arguments to lexical semantics, and the types of generalizations which can be obtained under current distributional approaches to computational semantics.

The more abstract task of argument differentiation can be directly linked to the practical task of automatic syntactic parsing. We propose that the task of argument disambiguation in a data-driven system provides us with a setting where the effect of various linguistic properties may be tested, and their interaction studied experimentally. In this respect, the property of being data-driven, as opposed to grammar-driven, allows for argument differentiation to be directly acquired through frequency of language use and with minimal theoretical assumptions. It enables an investigation of the relation of syntactic arguments to semantic interpretation, as well as to explicit, formal marking such as case and word order. Moreover, we may investigate whether the task of argument disambiguation can be improved by theoretically informed features and error analysis.

The overall research questions addresses in this thesis may be formulated as follows:

1. How are syntactic arguments differentiated?
    - Which linguistic properties differentiate arguments?
    - How do linguistic properties interact to differentiate an argument?
2. How may we capture argument differentiation in data-driven models of language? What are the effects?

The two main questions posed above are addressed throughout this thesis and can be viewed as constituting the central motivation behind the work presented here. Following from these, several more specific research questions will be posed during the course of the thesis which serve to further elucidate the topic of argument differentiation and its data-driven modeling.

1.4 Assumptions and scope of the thesis

The main languages in focus in this thesis are Scandinavian *type* languages, exemplified primarily by Swedish and Norwegian. The phenomena studied are not, however, limited to Swedish or Norwegian and we provide examples from a range of languages. The Scandinavian type languages exhibit some properties which make them interestingly different from English, while still being similar enough to warrant comparison. The case of argument differentiation touches upon issues that are relevant for several other languages and on methodological and theoretical issues which are of interest to linguists and computational linguists alike.

We aim throughout the thesis at a fairly theory-neutral investigation of arguments and argument differentiation. However, due to the nature of the problems which the thesis addresses, a certain bias will be present in the theories which are most readily used for exemplification and comparison. These will include *lexicalist* theories, due to the link to lexical semantics and *non-modular* theories, due to the mixed nature of the constraints taken from the syntax-semantics interface.

1.5 Outline of the thesis

The thesis is organized into three parts, where the two central parts, Part II and III, are largely independent and may be read separately.

Part I: Background

provides the relevant background by introducing the theoretical terminology, as well as models and resources employed in the ensuing parts of the thesis.
Chapter 2: Soft constraints addresses notions of soft, probabilistic constraints in linguistic theory. We discuss the role of frequency in the study of language and introduce the notion of soft, probabilistic constraints on language. The effect of incrementality on linguistic generalizations further leads us to the notion of linguistic ambiguity which is central to computational language processing, and syntactic parsing in particular. Finally, we discuss the notion of gradience and, more specifically, gradience in linguistic categories.
Chapter 3: Linguistic dimensions of argument differentiation starts out by introducing the notion of argumenthood in linguistics, as well as establishing a set of central distinctions within the group of arguments. We further introduce linguistic properties which have been proposed to differentiate syntactic arguments, in particular the property of animacy, as well as definiteness and

referentiality. We present evidence from linguistic studies providing cross-linguistic, as well as psycholinguistic and empirical support for the role of these properties in argument differentiation.
Chapter 4: Properties of Scandinavian morphosyntax describes some relevant properties of the Scandinavian languages, with a particular focus on the morphological and structural expression of syntactic arguments.
Chapter 5: Resources describes the corpora and resources employed for machine learning and parsing in the following two parts of the thesis. We provide a brief introduction to dependency representations, which will be central in Part III of the thesis. We also discuss some important distinctions in machine learning of linguistic data and present decision tree learning, memory-based learning and clustering.

Part II: Lexical Acquisition

concerns lexical acquisition of animacy information, with focus on the task of animacy classification. We briefly introduce the area of lexical acquisition and previous work which has focused on the relation between syntax and semantics.
Chapter 6: Acquiring animacy – experimental exploration presents a detailed study of animacy classification which investigates theoretical and practical issues including a definition of the learning task, feature selection and extraction, results, robustness to data sparseness and implications for the choice of machine learning algorithm.
Chapter 7: Acquiring animacy – scaling up deals with the scaling up of lexical acquisition of animacy information. We discuss schemes for animacy annotation and our requirements on such annotation. We experiment with a generalization of the results from chapter 6 in the application of animacy classification to a new data set in a different, although closely related, language. We discuss issues of data representation, data sparsity, class distribution and machine learning algorithm further and provide a quantitative evaluation of the method, as well as in-depth feature and error analysis.

Part III: Parsing

presents experiments in argument disambiguation, with a focus on linguistic features relating to argument differentiation. We introduce data-driven dependency parsing and motivate its use in the study of argument differentiation.
Chapter 8: Argument disambiguation in data-driven dependency parsing starts

out by defining a methodology for error analysis of parse results. We proceed to apply the methodology to a baseline parser for Swedish. We discuss the types of generalizations which are acquired regarding syntactic arguments and furthermore relate the errors to properties of argument expression in Scandinavian type languages.
Chapter 9: Parsing with linguistic features investigates the effect of theoretically motivated linguistic features on the analysis of syntactic arguments. We present a range of experiments evaluating the effect of different linguistic dimensions in terms of overall parse results, as well as on argument disambiguation in particular. We furthermore evaluate the effect of different parser properties on the results and discuss scalability in terms of parsing with automatically acquired features.

Chapter 10: Concluding remarks concludes the thesis by outlining its main contributions and directions for future work.

Part I

Background

2 SOFT CONSTRAINTS

The surge of empiricism characterising the last decades in the field of computational linguistics has also influenced the field of theoretical linguistics. The availability of large corpora and fairly good automatic annotation thereof provides the possibility to make new types of generalizations about language and language use. Dealing with real language with all its imperfections and massive variation has sparked an interest in more empirically motivated methods and models also within theoretical linguistics. In particular, the strict competence-performance dichotomy has been called into question. The main concern is that the traditional categorical distinctions are unsatisfactory in their coverage: "there is a growing interest in the relatively unexplored gradient middle ground, and a growing realization that concentrating on the extremes of continua leaves half the phenomena unexplored and unexplained" (Bod, Hay and Jannedy 2003: 1).

Based on work in both computational, theoretical and experimental linguistics, this chapter discusses a discernable shift in the view of human language and the modeling thereof. In particular, this shift is characterized by an acknowledgement that bridging the divide between studies of competence and studies of performance can be fruitful in unifying insights obtained in the various subfields of linguistics. Empirical investigations of language rely on the use of new types of data, in particular *frequency* of language use. The modeling of these results express probabilistic grammars of *soft constraints* on linguistic structure. The role of constraints in language processing and, in particular, the notion of *incrementality* raise further questions about the nature of constraints and their interaction. A probabilistic view of language furthermore entails *gradience* of grammaticality, as well as linguistic categories in general.

2.1 Frequency

The data-driven methods prevalent in current computational linguistics rely to a large extent on statistical modeling where frequency of usage is employed

to approximate probabilities. An interesting question is whether frequency in language and modeling thereof expresses generalizations of interest to more theoretically oriented linguists as well. Frequency has first and foremost been viewed as a property of performance or language *use* and frequency effects are found within all areas of linguistic realization. In the following we examine the role of frequency in linguistic theory, with particular focus on frequency as theoretical data, its role in language processing and in modeling of both practical, theoretical and experimental results.

2.1.1 Frequency as linguistic evidence

The view of what constitutes linguistic evidence is one distinguishing factor between largely rationalist and empiricist approaches to the study of human language. The rationalist view of linguistic theory, with inspiration taken from the natural sciences, sees the main task as the modeling of our internal linguistic knowledge, or competence, and introspection is considered sufficient evidence to this end. Strictly empiricist approaches, on the other hand, consider real language data to be paramount and the primary object of study, not necessarily attempting generalization across data sets. Within the area of corpus linguistics, the study of linguistic phenomena is synonymous with the study of frequency distributions in language use and corpus data is widely employed within a range of sub-disciplines of linguistics, e.g. lexicography, sociolinguistics, spoken language etc. (McEnery and Wilson 1996). This empiricist focus on properties of naturally occurring data has been viewed as irreconcilable with the rationalist goals. The strict division between rationalism and empiricism is admittedly an oversimplification. Most current day linguists employ both kinds of data in their theoretical and/or descriptive work. However, the extent to which properties observed in the data form part of a comprehensive model with testable consequences is not always explicitly clear.

Recent syntactic work within Optimality Theory (OT)[1] has exploited a gradient notion of *markedness* expressed through a set of ranked, universal constraints and has promoted the idea that "soft constraints mirror hard constraints" (Bresnan, Dingare and Manning 2001: 1); linguistic generalizations which incur categorical effects in some languages show up as strong statistical tendencies in other languages. This certainly calls the competence-performance dichotomy into question and in particular, the effect that the very same generalizations should form part of linguistic competence for the speakers of one language but be considered mere performance effects in another. The proposal

[1] See the introductory sections in Kager 1999 for an introduction to the main tenets of OT.

that a probabilistic grammar might be an alternative which provides a comprehensive model of these facts and thus cuts across the traditional competence-performance divide has emerged.

The idea that some linguistic generalizations are reducible to frequency of use is not new. The work within OT mentioned above, has adopted from functional and typological work the notion of markedness, which is based on "asymmetrical or unequal grammatical properties of otherwise equal linguistic elements" (Croft 2003: 87), where the more unmarked an element is, the more natural and typical it is. Frequency is clearly related to the notion of markedness and often figures as a criterion for this distinction (Croft 1990). It has been argued, however, that this notion of markedness may simply be reduced to differential frequency of language use (Haspelmath 2006). Rather than introducing the additional notion of markedness to account for these frequency effects, we should refer directly to frequency as the determining factor.[2]

Frequency as the central explaining factor is found in largely non-generative, usage-based accounts (Barlow and Kemmer 2000; Bybee and Hopper 2001), where the key role of frequency is linked to linguistic induction or learning. Starting from the same generalization that phenomena are frequent to varying degrees in different languages and calling the competence-performance distinction into question, we see that it is possible to arrive at an alternative conclusion, namely that it is all performance.

In general we can see that the role of frequency effects in language raises the issue of the balance between learning and innateness, i.e. how much of our linguistic knowledge is acquired and how much is innate? In this respect we may view the mainstream generative paradigm and the usage-based approaches mentioned above as representing extreme oppositions. Recent work discussing the theoretical implications of data-driven models, highlights the use of machine learning to assess hypotheses regarding language acquisition and the so-called 'poverty of the stimulus' argument for innateness (Lappin and Shieber 2007). Investigations into the relationship between syntactic structure and lexical semantics, and, in particular verbal semantic classes, have furthermore highlighted the use of machine learning methods over frequency distributions in language to test linguistic hypotheses (Merlo and Stevenson 2004).

[2]The type of markedness argument certainly has a flair of circularity: an element is unmarked because it is frequent and frequent because it is unmarked.

2.1.2 The mental status of frequency

Within psycholinguistics it has long been recognized that frequency plays a key role in human language processing and, furthermore, it is largely believed that language processing is probabilistic (Jurafsky 2003). Frequency has been shown to be an important factor in several areas of language comprehension (Jurafsky 2003):[3]

Access Frequent lexical items are accessed, hence processed, faster.

Disambiguation The frequency of various interpretations influences processing of ambiguity.

Processing difficulty Low-frequent interpretations cause processing difficulties.

These frequency effects are mostly connected to lexical form, i.e., word form or category, or lexical semantics. For instance, it has been shown that frequent words are processed faster. With respect to lexical ambiguities, studies indicate that use of the most frequent morphological category or most frequent sense of a lexeme stands in a direct relation to processing time. With respect to structural ambiguities in language comprehension, subcategorization frame probabilities have been related to parsing difficulties in notorious garden-path sentences, such as, e.g., *The horse raced past the barn fell*, see section 2.3.1.

Efforts to link results from empirically oriented, theoretical work with psycholinguistic evidence have highlighted the role of frequency also in production, in particular with respect to variation or syntactic *choice*. Bresnan (2006) presents results from forced continuation experiments on the dative alternation and argues that the same set of soft, probabilistic, constraints which were shown to correlate with the choice of dative construction in corpus studies (Bresnan and Nikitina 2007; Bresnan et al. 2005) are also active in the judgements of language users. This indicates that language users have detailed knowledge on the interaction of constraints and Bresnan (2006) concludes, somewhat controversially, that syntactic knowledge is in fact probabilistic in nature.

2.1.3 Frequency and modeling

Frequency effects in language lend themselves readily to probabilistic modeling and provide empirical estimates for probabilistic model parameters. In

[3] Jurafsky (2003) reasons that these phenomena are influenced by *probability* and goes on to present evidence from experiments showing the effect of raw frequencies or conditional probabilities estimated by frequencies.

computational linguistics, probabilistic modeling based on language frequencies has permeated practically all areas of analysis.[4] Stochastic models, such as Hidden Markov models (HMMs) and Bayesian classifiers have been widely employed in word-based tasks such as part-of-speech tagging and word sense disambiguation. In parsing, probabilistic extensions of classical grammar formalisms, such as probabilistic context-free grammars (PCFGs) (Charniak 1996) and the lexicalized successors in various incarnations (Collins 1996; Charniak 1997; Bikel 2004), have dominated the constituent-based approaches to parsing. Central to this development has been the use of syntactically annotated corpora, or *treebanks* (Abeillé 2003) and parameter estimation from treebanks.[5] The use of statistical inference in induction of information from corpus data constitutes an integral part of most NLP systems, recasting a range of complex problems, such as named-entity tagging (Tjong Kim Sang 2002b), phrase detection/chunking (Tjong Kim Sang and Buchholz 2000), parsing (Buchholz and Marsi 2006; Nivre et al. 2007) and semantic role labeling (Carreras and Màrquez 2005) as classification problems.

Probabilistic models have also been widely employed to model human language processing. The primary concern is that these models should provide realistic approximations of the language processing task and, in particular, be predictive of the types of processing effects indicated by experimental results. For the processing of lexical ambiguities, HMMs have been employed and syntactic ambiguities have been modeled employing probabilistic extensions of grammars, such as probabilistic context-free grammars (PCFGs). The processing difficulties observed in conjunction with the garden-path sentences mentioned above, so-called 'reanalysis', can then be directly related to the presence of an additional rule with a small probability in the reanalysis. Furthermore, within the area of language acquisition, probabilistic modeling is common and the learning problem can be formulated as acquisition of a set of weighted constraints through exposure to linguistic data, expressing a connectionist, functionalist view of language, (see, e.g., Seidenberg and MacDonald 1989).

Within theoretical linguistics, the probabilistic modeling of frequencies has mostly been descriptive, for instance in testing statistical significance of distributional differences. To a certain extent, probabilistic models have also been employed to test the strength of various correlations by means of logistic regression models in particular, (see, e.g., Bresnan et al. 2005; Rahkonen 2006;

[4] See Manning and Schütze 1999 for an overview.

[5] Note however that lexicalized parsers necessarily rely on advanced techniques for smoothing of sparse data, hence maximum likelihood estimation is not sufficient for parameter estimation. One common technique is to markovize the rules (Collins 1999; Klein and Manning 2003).

Bouma 2008). Probabilistic models also provide a method for modeling the interaction of probabilities over syntactic structure without necessarily demanding a rebuttal of the tools of formal syntactic models and frameworks developed over a long period of time. A simple example is a probabilistic context-free grammar which conditions the probability of a sentence on the probabilities of its subtrees. However, more sophisticated theories of syntax based on a notion of probability have also been proposed (Bod 1998). In theories where grammatical generalizations are expressed as constraints on structure, these constraints may themselves be associated with probabilities (or 'weights') and their interaction modeled using probabilistic models. Within the framework of Optimality Theory there has been a substantial amount of work in recent years on probabilistic formulations of constraint interaction.

2.2 Constraints

Generally speaking, a constraint restricts a solution, usually by providing a condition which must be fulfilled. Constraint-based theories are central in the theoretical and psycholinguistic modeling of syntactic structure. However, properties of the constraints employed differ in a way that corresponds with the object of study and the data employed to do so. In theoretical linguistics, the constraints are generally assumed to be absolute and based on strict grammaticality judgements, whereas experimental results indicate the use of probabilistic constraints in human language processing. Recent work in theoretical linguistics, however, opens up for a reconsideration of properties of constraints as a reflection of linguistic knowledge.

2.2.1 The status of constraints

Within the discourse of syntactic formalisms, the term 'constraint' has been widely used. Constraint-based theories such as Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag 1994; Sag, Wasow and Bender 2003) and Lexical Functional Grammar (LFG) (Kaplan and Bresnan 1982; Bresnan 2001), are often contrasted with derivational theories, such as Government and Binding (Chomsky 1981) and Minimalism (Chomsky 1995). One of the main differences between the two is situated in the view of syntactic structure as constructed or ultimately constrained. Central to a notion of constraint-based syntax is the idea that constraints limit the number of possible grammatical structures in a way that corresponds to the system modeled, namely our linguistic competence. The constraint-based theories place much of the con-

straining power in the lexicon, where constraints in lexical entries restrict the possible combinatory space in syntactic structure. In much the same way that derivational theories associate restrictions in terms of structural positions along with movement, constraint satisfaction in constraint-based theories is assured by means of unification. The constraints are *absolute* in the sense that they impose requirements on structure which must be fulfilled.

Optimality Theory (OT) operates with a somewhat different view of constraints. Here the constraints are *violable*, or 'soft', but strictly ranked with respect to each other and a violation of a constraint is possible only to fulfil a constraint that is higher in rank. The interaction of constraints in a ranking is therefore the key to understanding the difference between the two notions of constraints. The principal notion of a constraint as a "structural requirement that may be either satisfied or violated by an output form" (Kager 1999: 9) is thus not shared by the two directions outlined above, since constraint violation excludes any output form in the constraint-based theories.

The effect of the constraints on linguistic structure, whether absolute or ranked and violable, however, is common to both of the types of constraint-based theories outlined above. In OT-terms, there is only one output for any given input – both the constraint-based theories and OT operate with a categorical notion of grammaticality. It does not make sense within these theories to speak about varying acceptability of different constructions or outputs.

2.2.2 Soft constraints

In contrast to the view of constraints presented above, recent work within Optimality Theory has focused on the use of soft, in the sense 'probabilistic' or 'weighted', constraints. In line with the shift towards empirical methods in computational linguistics, focus on the relationship between language data and (OT) grammars has resulted in work on acquisition of constraint rankings from corpus data.

Constraints in an OT grammar are ranked in a hierarchy of dominance, related through *strict domination* (Kager 1999: 22):

Strict domination: Violation of higher ranked constraints cannot be compensated for by satisfaction of lower-ranked constraints.

It follows from the above definition that i) constraint ranking is strict, not variable and ii) constraint violations are non-cumulative. The work on soft, weighted constraints in OT challenges both of these entailments.

Soft constraints were initially introduced in OT to model linguistic variation (Boersma and Hayes 2001; Goldwater and Johnson 2003), but has also been

applied to syntactic variation (Bresnan and Nikitina 2007; Bresnan, Dingare and Manning 2001; Øvrelid 2004). In order to account for more than one possible output for a given input, i.e., linguistic variation, constraints may be defined over a continuous scale, where the distance between the constraints is proportional to their fixedness in rank. The ranks or weights of constraints are acquired from language data and thus reflect the frequency distributions found in the data.[6] Goldwater and Johnson (2003) make use of a Maximum Entropy model to learn constraint weights and model constraint interaction.

The use of a Maximum Entropy model for modeling constraint interaction brings us to the second entailment above, namely the issue of *cumulativity*. It is one of the main tenets of OT that no amount of violations of a lower ranked constraint can cancel out a violation of a higher ranked constraint. This is not, however, a property of most probabilistic models where cost computations often are additive. Jäger and Rosenbach (2006) discuss models for variation in OT and put forward empirical evidence for cumulativity in the syntactic variation of the English genitive alternation. The view is of the alternation as probabilistic variation and statistical tendencies in language data are employed as evidence.A distinction between soft and hard constraints has furthermore been introduced in modeling of experimental judgement data, where these are proposed to differ in the observable effect that their violations incur on the relative acceptability of a sentence (Keller 2000).[7]

We thus observe two notions of 'soft constraint' emerging in recent discourse, where the main difference between the two is found in constraint interaction:

Standard OT Constraints are soft in the sense that they may be violated and are strictly ranked. This is the standard sense of a soft constraint which distinguishes between the view of constraints within OT and other constraint-based theories.[8]

[6]We may note, however, that the modeling of linguistic variation does not necessarily demand the introduction of probabilistic constraints, although, within an OT setting, it does entail relaxation of the demand for strict ranking. Proposals have been made that employ unranked constraints, however, still ordinal as in standard OT (Anttila 1997). Furthermore, the introduction of probabilistic constraints does not necessitate variable ranking. A categorical OT system with a strict ranking of constraints within a probabilistic setting simply constitutes an extreme where all constraints are ranked so far apart as to be non-interacting.

[7]Keller (2000) proposes a version of Optimality Theory, Linear Optimality Theory (LOT), where constraints come in two flavours – soft and hard. The weighting of constraints in LOT models numerical acceptability data from Magnitude Estimation experiments (Bard, Robertson and Sorace 1996). Unlike the work discussed above, however, Keller argues that the status of a constraint as soft/hard is not susceptible to cross-linguistic variation; if a constraint is soft in one language, it is soft in another too. So rather than allowing for the soft/hard distinction to follow directly from the weighting of constraints, it is stipulated independently as a universal property of the constraints.

[8]We may note, however, that OT and constraint-based theories like HPSG and LFG should not be viewed as competitors due to the fact that they operate on different levels. OT is a theory of constraint interaction and not representation and is fully compatible with other representational theories, see for instance work on OT-LFG (Choi 2001; Kuhn 2001).

Probabilistic OT Constraint interaction is furthermore probabilistic, in the sense that

- constraints are weighted,
- constraint interaction is stochastic (not strictly ranked),
- constraint interaction is (possibly) cumulative.

Probabilistic OT is thus an extension of Standard OT. We may note that a very similar development can be found in work on automatic, syntactic parsing. As an equivalent to the hard notion of constraints discussed above, a line of work in dependency parsing proposes disambiguation by boolean constraints taken from various linguistic levels of analysis through constraint propagation in a constraint network (Maruyama 1990). Extensions of Maruyama's approach has included a notion of soft, weighted constraints (Schröder 2002) and some work has also been done on machine learning of grammar weights for these hand-crafted constraints (Schröder et al. 2001). Parsing with a set of weighted constraints, where hard constraints are simply constraints located at the extreme end of the scale, recasts the parsing problem as an *optimization problem*, i.e. locating the best of all possible solutions which maximizes/minimizes a certain scoring function. The parallel to the constraint interaction proposed in OT is obvious when parsing is modeled as an optimization problem where the search space consists of all possible linguistic analyses (Buch-Kromann 2006).

2.3 Incrementality

Human language processing and modeling thereof is characterized by *incrementality*; data is presented bit by bit, hence analyses are necessarily based on incomplete evidence. Probabilistic models are typically employed in modeling, providing a model of decision making under uncertainty and based on incomplete evidence. Effects of incremental language processing have typically been attributed to performance, along with extra-linguistic factors such as memory load. However, the interest in probabilistic grammars as discussed above, opens for a reevaluation of the competence-performance distinction and its bearing on linguistic theory building:

> We believe not only that grammatical theorists should be interested in performance modeling, but also that empirical facts about various aspects of performance can and should inform the development of the theory of linguistic competence. That is, compatibility with performance

> models should bear on the design of competence grammars. (Sag and Wasow 2008: 2)

In the following we discuss processing of ambiguity, a problem which has been widely studied in both theoretical, computational and experimental linguistics, hence may be employed to illustrate the demands of incrementality on the nature of constraints and constraint interaction.

2.3.1 Ambiguity processing

Ambiguity is a property which is characteristic of natural language, distinguishing it from formal languages. It consists of a mismatch in the mapping between form and meaning, where one form corresponds to several meanings (Beaver and Lee 2004). Ambiguities in natural language have been widely studied within theoretical linguistics, psycholinguistics and computational linguistics. It is a notorious problem within NLP, in particular within the areas of part-of-speech tagging, syntactic parsing and word sense disambiguation. Ambiguity has is seen as one of the main reasons "why NLP is difficult" (Manning and Schütze 1999: 17) and is prevalent at all levels of linguistic analysis. In psycholinguistics, ambiguities have been claimed to increase processing difficulty (Frazier 1985) and the study of ambiguity processing has been performed under the assumption that it can be indicative of the underlying architecture and mechanisms of the human language faculty.

2.3.1.1 Types of ambiguity

As mentioned, ambiguity is found at all levels of linguistic analysis, ranging from the level of morphemes, so-called *syncretism*, to semantic and pragmatic ambiguities. Ambiguity with respect to syntactic arguments is, however, in a majority of cases caused by ambiguity in lexical form or in the syntactic environment.[9]

Lexical ambiguities are ambiguities associated with lexical units which have more than one interpretation or meaning. These types of ambiguities are extremely common, and especially frequent words tend to be polysemous. *Categorial* ambiguity is found where a word has several meanings, each associated with a distinct category or word class. For instance, *time* is both a noun and a

[9] In section 4.1 we examine examples of syncretism in morphological case marking, which directly contribute to functional ambiguity.

verb. Function words are notoriously ambiguous, e.g. *to* may be both an infinitival marker and a preposition and *that* may be a determiner, a demonstrative pronoun and a complementizer (Wasow, Perfors and Beaver 2005). Categorial ambiguity has syntactic consequences since the category of a lexical item clearly influences its syntactic behaviour. The example in (1) illustrates the polysemy of the English noun *case*, and (2) the categorial ambiguity of *strikes* and *idle*, which both can be used as a as verb, as well as noun or adjective (Mihalcea 2006):

(1) Drunk gets nine years in violin case

(2) Teacher strikes idle kids

Structural ambiguities are found when a sentence may be assigned more than one structure. These include PP-attachment ambiguities, as in (3), coordination ambiguities, as in (4) and noun phrase bracketing ambiguities, as in (5):

(3) The prime minister hit the journalist with a pen

(4) Choose between peas and onions or carrots with the steak

(5) He is a Danish linguistics teacher

2.3.1.2 Global and local ambiguity

Orthogonal to the types of ambiguity discussed above, and hence regardless of the source of ambiguity, we may distinguish between global and local ambiguity. In the processing of ambiguity in language, and with reference to a sentence, local ambiguity obtains when parts of a sentence is ambiguous, whereas global ambiguity is found when the whole sentence is ambiguous, cf. (3)-(5) above. Since human language processing is incremental in nature, local ambiguities can cause processing difficulties, for instance in so-called garden path sentences:

(6) I knew the solution to the problem was correct

A garden-path effect is observed when interpretation changes during the incremental exposure to a sentence. In (6), the postverbal argument is initially interpreted as an object, but must be reanalyzed as subject of a complement clause when the second verb is encountered.

2.3.1.3 Ambiguity resolution

Disambiguation is the process of resolving ambiguities and within NLP many tasks involve disambiguation in some form. Word sense disambiguation, for instance, is solely devoted to the resolution of lexical ambiguities, whereas part-of-speech tagging deals with the subclass of categorial ambiguities. In syntactic parsing, disambiguation is a crucial task which is dealt with in a variety of ways. Irrespective of the particular approach to parsing, disambiguation can be defined as a "process of reducing the number of analyses assigned to a string" (Nivre 2006: 23). In most current approaches to parsing this is achieved by assigning probabilities to the syntactic structure(s), approximated by frequency data from language use. Disambiguation is then performed either as a post-processing step over the total of analyses, or as an integral part of the parsing process itself, often in combination with deterministic processing.

The processing of ambiguity has been studied extensively in psycholinguistic experiments and has been argued to provide evidence for the mechanisms of the human language processor. Important topics in this respect have been the role of frequency in lexical ambiguity resolution and the role of various types of linguistic information in the processing of structural ambiguities. In a seminal article, MacDonald, Pearlmutter and Seidenberg (1994) propose that resolution of lexical and structural ambiguities, contrary to earlier assumptions, follows the same types of strategies. In particular, language processing can be viewed as a constraint satisfaction problem, where interpretation is constrained by a set of largely lexical, probabilistic constraints. Needless to say, frequency plays an important role in ambiguity resolution in such a model.

2.3.2 Constraining interpretation

We have earlier discussed how frequency effects can affect sentence comprehension, as well as how language-specific frequency effects, typically assigned to the realm of performance, have been claimed to provide evidence for probabilistic grammars of universal competence-oriented constraints. The study of language comprehension raises further questions regarding properties of a comprehensive model of grammar, unifying insights from the study of competence and performance alike.

Results from psycholinguistics suggest several properties that are relevant for grammatical constraints to be "performance-compatible" (Sag and Wasow 2008):

Surface oriented Processing deals with "what is actually there".

Non-modular Information from all linguistic levels should interact.

Lexical Individual words should carry information on their combinatory potential, as well as their semantic interpretation.

With respect to the theories of constraints discussed in section 2.2 above, we find that both the constraint-based theories employing absolute constraints, and OT, which uses violable ranked constraints, are compatible with these demands. LFG and HPSG, being theories of representation, are explicit lexicalist theories, whereas all three are non-modular in not placing any restrictions on the type of information which may interact in parallel.[10]

One might also take the integration of performance-compatible constraints further and suggest that not only should a grammatical model of competence be compatible with the processing of performance data, but it should in fact be one and the same model (Bod 1995, 1998). An important property is then found in the ability to provide an analysis for sentence fragments and a main concern is that incrementality is incompatible with a categorical notion of grammaticality, at least one that is defined by hard, global constraints over complete sentences. OT provides one possible approach for such a model (Stevenson and Smolensky 2005; de Hoop and Lamers 2006), due to the fact that constraints under this approach are violable and therefore provide an analysis for any input, including sentence-fragments.

2.4 Gradience

Gradience is employed to refer to a range of continuous phenomena in language, ranging from morphological and syntactic categories to phonetic sounds. The idea that the language system is non-categorical has been promoted within several subdisciplines of linguistics – phonology, sociolinguistics, typology and gradient categories have been examined at all levels of linguistic representation (Bod, Hay and Jannedy 2003).

2.4.1 Grammaticality

We have discussed the implications of a probabilistic grammar expressed in terms of constraints on linguistic structure. One implication of such a view is a gradient notion of grammaticality.

[10] These theories are 'lexicalist' in the sense that they place much of the explanatory burden in the lexicon, i.e. the lexical entries contain a majority of the information needed to interpret a sentence. They are also lexicalist in the sense that they adhere to the principle of Lexical Integrity (Bresnan 2001); words are the smallest units of syntactic analysis and the formation of words is subject to principles separate from those governing syntactic structures.

Whereas, 'degrees of grammaticalness' (Chomsky 1965, 1975), has played a certain role in generative theoretical work, there has been no systematic incorporation of such notions in the proposed grammatical models. Manning (2003) argues for the use of probabilistic models to explain language structure and motivates his claims by the following observation:

> Categorical linguistic theories claim too much. They place a hard categorical boundary of grammaticality where really there is a fuzzy edge, determined by many conflicting constraints. (Manning 2003: 297)

The concern that introduction of probabilities into linguistic theory will introduce chaos is unfounded, according to Manning (2003). Rather, a probabilistic grammar can be seen to broaden the scope of linguistic inquiry, and doing so in a principled manner. A probabilistic view of grammaticality can thus provide more fine-grained knowledge about language and the different factors which interact.

2.4.2 Categories

Linguistic category membership can also be gradient in the sense that elements are members of a category to various degrees. In general, we find gradience between two categories α and β when their boundaries are blurred. By this we mean that some elements clearly belong to α and some to β, whereas a third group of elements occupy a middle ground between the two. The intermediate category possesses both α-like and β-like properties (Aarts 2004).

In work on descriptive grammar it is often recognized that taxonomic requirements of linguistic categories are problematic; elements do not all neatly fall into a category and some elements have properties of several categories. For instance, it is well known that providing necessary and sufficient criteria for membership in part-of-speech classes is difficult and a view of these criteria as graded, or weighted, was proposed as early as in Crystal 1967. *Prototype theory*, following influence from psychology, has been influential in cognitive linguistics (Lakoff and Johnson 1980; Lakoff 1987) and promotes precisely the idea that membership in a category is not absolute, but rather a matter of gradience. Moreover, gradience is defined with reference to a prototypical member of a category.

One response to graded phenomena which maintains a sense of categoricity is the introduction of split categories. For instance, in LFG, phrasal categories may be both functional and lexical in terms of the notion of 'co-heads', and HPSG allows for multiple inheritance in type hierarchies.

2.5 Conclusion

The empirical shift mentioned initially is evident in work ranging from theoretical and experimental approaches to computational modeling of natural language. The work described in this thesis adheres to an empiricist methodology, focusing on the essential role of language data in linguistic investigations. Furthermore, we ascribe to a view of language where linguistic structure is determined by a set of, possibly conflicting, constraints. In chapter 3 we examine the linguistic dimensions of argument differentiation, an area which has been proposed to be influenced by constraints on linguistic structure which show up as frequency effects in a range of different languages. The main parts of this thesis will be devoted to the investigation and computational modeling of argument differentiation. In particular, we employ data-driven models taken from computational linguistics, which support a direct relation between frequency of language use and linguistic categories.

Data-driven models rely on statistical inference over language data, combining different sources of information and can in this respect be seen to express soft, probabilistic constraints. Within the area of syntactic parsing, computational models of incremental parsing may be studied to elucidate properties of constraints further. In chapter 8, we introduce data-driven dependency parsing (Nivre 2006) as an instantiation of such a model. We will study argument disambiguation and investigate the effect of various types of linguistic information. The linguistic features employed in the study of argument disambiguation in chapter 9 are theoretically motivated and furthermore surface-oriented, lexical and non-modular.

The direct relationship in data-driven models between frequency of language use and categories furthermore enables a study of gradience. We will in the following chapters discuss categorial gradience in several places, and in particular with respect to semantic properties, such as animacy and selectional restrictions.

3 LINGUISTIC DIMENSIONS OF ARGUMENT DIFFERENTIATION

This chapter presents argument differentiation and its linguistic dimensions. We start out by briefly introducing the notion of argumenthood and discuss some further distinctions within the category of arguments. The introduction of the term 'argument differentiation' is motivated and we go on to discuss several linguistic factors which have been proposed to differentiate between the arguments of a sentence. We discuss the factors independently, as well as their interaction in the context of argument differentiation. This chapter thus introduces terminology which will be employed in the following and provides theoretical motivation for the linguistic properties which will be investigated in Part II and Part III of the thesis.

3.1 Arguments

A distinction between *arguments* and *non-arguments* is made in some form or other in all syntactic theories.[11] The distinction can be expressed through structural asymmetry or stipulated for theories where grammatical functions are primitives in representation. For instance, in LFG (Kaplan and Bresnan 1982; Bresnan 2001), grammatical functions are primitive concepts and arguments or governable functions (SUBJ, OBJ, OBJ_θ, OBL_θ, COMP, XCOMP) are distinguished from non-arguments or modifiers (ADJ, XADJ). HPSG (Pollard and Sag 1994; Sag, Wasow and Bender 2003) similarly distinguishes the valency features (SPR, COMPS) from modifiers (MOD). In most versions of dependency grammar, (see, e.g, Mel'čuk 1988; Hudson 1990), grammatical functions are also primitive notions and not derived through structural position.[12]

Regardless of notation, the notion of argumenthood is important in syntactic theory and is closely related to the semantic interpretation of a sentence.

[11] We adopt the more theory-neutral term of 'non-argument', rather than 'adjunct', which is closely connected to the structural operation of adjunction.

[12] For a brief introduction to dependency grammar, see section 5.1.1.

Dalrymple (2001) cites Dowty (1982) in proposing two tests for argumenthood:

(7) Tests for argumenthood (Dowty 1982):

(i) Entailment - the existence of an argument is entailed by the predicate

(ii) Subcategorization - arguments are obligatory, non-arguments are optional

These two tests decompose the notion of an argument, positioning it in the syntax-semantics interface. The entailments of a predicate are closely related to the argument structure of a predicate, which characterize the core participants, or thematic roles, involved in an event. The subcategorization of a verb relates to the obligatoriness of an argument, hence constrains the syntactic realization of the event. Neither test, however, provides a sufficient criterion by which to distinguish arguments from non-arguments. The entailment test is not strict enough, for instance allowing for time adverbials to be arguments since all events entail a location in time and space. The subcategorization test, on the other hand, is too strict in excluding, for instance, arguments of verbs like *eat*, which may function intransitively. As Dalrymple (2001) notes, both of these tests still make some valid predictions: "if a phrase is an argument, it is either obligatorily present or it is entailed by the predicate. If a phrase is a modifier, it can be omitted" (Dalrymple 2001: 12).[13] Other tests for the argument/non-argument distinction include iteration and reordering of non-arguments (Sag, Wasow and Bender 2003).

Cross-linguistic generalizations relating to grammatical functions often make reference to a hierarchy, such as the one in (8) below (Keenan and Comrie 1977; Bresnan 2001):[14]

[13]Due to the amount of variation exhibited by different verbs in their subcategorization frames, Manning (2003) proposes a probabilistic view of argumenthood, according to which the exceptions from tests like the ones in 7 simply represent less prototypical, or likely, arguments.

[14]The hierarchy in (8) is taken from Bresnan 2001 and differs from the original hierarchy (Keenan and Comrie 1977) in the inventory of object functions. Keenan and Comrie (1977) employ the distinction between direct and indirect objects, and impose the ordering OBJ $>$ IOBJ on these. This distinction takes semantic role to be indicative of function, and groups together the theme object of a ditransitive verb with the object of a monotransitive verb. An alternative distinction is made between *primary* and *secondary* objects in typological work on grammatical functions (Dryer 1986) and is also the main distinction made in LFG, where OBJθ is the secondary object. It is argued that in many languages indexation and case-marking group the indirect object with the monotransitive object (primary objects) and treats these as distinct from the secondary (direct) object in ditransitive constructions. English has been argued to follow both of these, with evidence in the dative alternation illustrated in (10)–(11).

(8) SUBJ > OBJ > OBJ_θ > OBL > COMP > ADJ

The main idea behind such a hierarchy is that a generalization which applies to an element on the scale will also apply to the elements to the left of it.[15] Grammatical hierarchies may also be interpreted as expressing the relative *prominence* of the ranked elements. In this case, prominence can be defined structurally, but recent proposals have highlighted highly ranked elements as being cognitively accessible, see section 3.6 below.

We distinguish between *core* and *non-core* argument functions (Bresnan 2001). Subjects and objects (direct and indirect) are the core functions, whereas various oblique functions, as well as clausal complements, are non-core. Phenomena which differentiate core from non-core arguments are possibilities for verb agreement, anaphoric binding patterns and control (Dalrymple 2001).

There are also reasons to distinguish the subject relation from those of the other argument relations, as the *external* argument. The external argument is in theories such as HPSG assigned a relation (SPR), which groups it with determiners. This is a clear feature structure translation of the structural asymmetry expressed in $\bar{X}$ theory as holding between specifiers and complements. The differentiation of the subject from the other argument functions, however, is not only based on structural assumptions. Subjects exhibit linguistic properties which differentiate them from the other argument functions, such as direct objects. Phenomena which only the subject participates in thus have a "cut-off point" after the first element in the hierarchy. These phenomena include verb agreement in a range of languages (including English), honorification in Japanese, as well as raising and control phenomena.[16]

We introduce the term *argument differentiation* and will in the following employ it as a neutral cover term to denote the process by which arguments are distinguished along one or more linguistic dimensions. The rationale behind the introduction of this term reflects the mediating status of arguments between syntax and semantics. First of all, argument differentiation will be employed as neutral with respect to theory or application, as opposed to terms like 'interpretation' or 'disambiguation', which are more or less theoretical and applied terms, respectively. We also, as mentioned initially, wish to maintain a non-modular orientation in the following, and argument differentiation reflects this orientation in not taking syntactic or semantic evidence to be primary. This allows us to generalize over mapping from meaning to form, as expressed by 'realization', and from form to meaning, known as 'interpretation', as well as

[15] The particular hierarchy in 8 relates to accessibility of grammatical functions for relativization, e.g. if direct objects may be relativized in a language, then it will also be possible to relativize the grammatical subject etc. The original hierarchy also includes genitive modification.

[16] It is only the subject of the subordinate clause which may be raised or controlled.

terminology employed in work on psycholinguistic processing, e.g., 'production' and 'comprehension'. We will employ the above terms when appropriate to make clear the exact application in the specific context.

3.2 Animacy

The dimension of animacy roughly distinguishes between entities which are alive and entities which are not, however, other distinctions are also relevant and the animacy dimension is often viewed as a continuum. Animacy is a grammatical factor in a range of languages and is closely related to argument realization and differentiation. In this section we examine how animacy influences language, with focus on argument differentiation, and we furthermore examine some properties of the category of animacy itself.

The effect of animacy in linguistic phenomena has been noted several places in the literature and we provide a few examples of this below. For a more detailed overview, see Yamamoto 1999. Typological work on animacy often makes reference to an animacy hierarchy or scale, following Silverstein 1976. An example of an animacy hierarchy, taken from Aissen 2003, is provided in (9):[17]

(9) Human > Animate > Inanimate

Evidence for this hierarchy comes from cross-linguistic examination of the realization of animacy in different languages, and especially of how animacy motivates morphological and/or functional "splits" in various ways. The scale in (9) generalizes over phenomena which are influenced by the animacy of the referents involved by providing a set of implications following from different cut-off points on the hierarchy. For instance, number marking may be sensitive to animacy in the sense that elements with human or animate reference exhibit number distinctions not possible for elements with inanimate reference, as found in, e.g., Tiwi and Kharia (Yamamoto 1999). The phenomenon known as Differential Object Marking (Aissen 2003; Comrie 1989) provides another example, where the morphological case marking of direct objects may be determined by animacy and where different languages exhibit different cut-off points on the above hierarchy. For instance, in the Dravidian language Malayalam, we find case marking of objects which are human and animate referring and objects with inanimate reference are unmarked for case.

[17] Comrie (1989) calls the middle category in the hierarchy *Animal*, whereas Aissen uses the term *Animate*. As we shall see, the intermediate category need not be limited to animals, but rather highlights the gradient nature of the animacy dimension, see section 3.2.4.

3.2.1 Animacy of arguments

A recent special issue of the linguistic journal *Lingua* was dedicated to the topic of animacy and discusses the role of animacy in natural language from rather different perspectives, ranging from theoretical and typological to experimental studies (de Swart, Lamers and Lestrade 2008). These various perspectives all highlight animacy as an influencing factor in argument differentiation.

There is a cross-linguistic tendency for external arguments or subjects to be human or animate and for objects to be inanimate (Comrie 1989). de Swart, Lamers and Lestrade (2008) cite examples from languages like Jakaltek, where inanimate subjects are simply ungrammatical, but where human/animate subjects are perfectly grammatical.

We may distinguish between the effect of *isolated* versus *relative* animacy, that is, whether the animacy of an isolated element determines an effect or whether it is the animacy of one argument relative to another which creates an effect in a language. For instance, in the Mayan language MamMaya, a transitive sentence is ungrammatical if the object is higher in animacy than the subject, as in *The dog sees the woman* (de Swart, Lamers and Lestrade 2008). In Navajo, such a construction is clearly avoided and an alternative construction (*The woman is seen by the dog*) is chosen instead.[18] In many languages this tendency is reflected in language data as a frequency effect, even though these types of transitive constructions are perfectly grammatical (Dahl and Fraurud 1996; Øvrelid 2004). Following a corpus study of animacy in Swedish, Dahl and Fraurud (1996) conclude that:

> [M]ore than 97% of all transitive sentences obey the constraint that the subject should not be lower than the object in animacy. Thus, this constraint, which is grammaticalized in a language such as Navajo, could be said to be approximated statistically in Swedish texts. (Dahl and Fraurud 1996: 53)

Animacy furthermore has an effect on the differentiation of core and non-core arguments. Bresnan et al. (2005), for instance, argue that animacy is an important factor in the so-called dative alternation in English, clearly influencing the choice between expression the double object construction in (10) and the prepositional dative structure in (11):

(10) . . . gave the prime minister a pen

[18]The *inverse* construction in Navajo can be paraphrased by our passive construction and is expressed by the verbal affix *bi* and employed when the subject is lower in animacy than the object (Dahl and Fraurud 1996).

(11) . . . gave a pen to the prime minister

3.2.2 Ambiguity resolution

The influence of animacy in both language production and comprehension has been widely investigated in psycholinguistic studies. By manipulating the animacy of elements in otherwise controlled environments, the effect of animacy on syntactic structure may be studied experimentally.

Animacy effects have played an important role in the debate in psycholinguistic theory between two different views of language processing and the modeling thereof – a serial, modular or "syntax-first" model versus a single-stage model, where different kinds of information from different linguistic levels, such as syntax, semantics and pragmatics, interact. In particular, the fact that animacy, being a lexical and semantic property, influences syntactic interpretation has been taken as evidence for the latter type of model. Abandonment of modular processing models has characterized psycholinguistic work in the last decade, and the use of new types of processing evidence[19] has enabled even more detailed results on the use of various information sources during language processing (Sag and Wasow 2008).

In comprehension studies, animacy has been shown to have a clear effect on the resolution of grammatical function ambiguities. The tendency for animate elements to be syntactically prominent, as discussed in the preceding section, is shown to provide an important information source in disambiguation. Weckerly and Kutas (1999) report results from ERP experiments on the comprehension of English object relatives[20] and argue that a probabilistic constraint-based, interactional model is most appropriate for modeling the influence of animacy on choice of syntactic structure. They find an early effect of animacy, independently of the verb, which expresses the clear correlation between animacy and syntactic function assignment. Inconsistencies between syntactic and semantic information, following the cross-linguistic tendencies outlined in the previous section, result in clear experimental effects. Mak, Vonk and Schriefers (2006) come to similar conclusions after studying the effect of animacy on the processing of Dutch relative clauses using a self-paced reading task. They find that the relative animacy of the entities in question is most

[19] Event-Related brain Potentials (ERP) and eye-movement tracking are examples of experimental methods which allow for on-line and more precise measures of processing activity, through electrical and muscular activity, respectively.

[20] Object relative constructions are tested with differing animacy of the extracted object, as well as the subject of the relative clause, as in *the poetry that the editor recognized* . . . vs. *the editor that the poetry baffled* . . . (Weckerly and Kutas 1999).

important and can even counteract the usual preference for subject-relatives which has a strong influence on processing.[21]

3.2.3 The nature of animacy effects

There are languages where animacy creates hard, categorical effects on the realization of arguments. These hard effects are found in particular in the encoding of core arguments through morphology, as in the phenomenon of Differential Object Marking mentioned above. In most theoretical and experimental work where animacy figures, however, it is claimed to be a soft, probabilistic constraint on structure, with evidence in frequency effects either in corpus data or in experimental results. Animacy is then argued to influence the choice of syntactic structure and the realization or interpretation of syntactic arguments.

Theoretical studies have examined the influence of animacy in a range of syntactic constructions in various languages. The focus has in particular been on various grammatical alternations, such as the active-passive alternation (Bresnan, Dingare and Manning 2001), the dative alternation (Bresnan and Nikitina 2007; Bresnan et al. 2005; Bresnan 2006) and the genitive alternation (Rosenbach 2003, 2005, 2008). Here, the choice of construction is seen as depending on several factors and central themes in this work is investigating the effect of these factors by assessing their predictive strength and teasing them apart with reference to the particular construction under scrutiny.

Experimental studies involving animacy share with the theoretical work discussed above the assumption that animacy influences the choice of syntactic structure, and, in particular, the differentiation of arguments. Whether from the perspective of language production, where the outcome is directly observable, or from that of comprehension where the chosen analysis shows up as a response to language data, the probabilistic effect of animacy has been reported in numerous studies (Branigan, Pickering and Tanaka 2008; Weckerly and Kutas 1999; Mak, Vonk and Schriefers 2006).

As we have seen, animacy has a hard effect on the realization of arguments in many languages and it is also evident in distributional tendencies in a range of languages. One possibility which has been explored in recent studies is that the hard and soft effects of animacy are instances of the same constraints on language and that "soft constraints mirror hard constraints" by simply hav-

[21] Dutch subject and object relative clauses do not differ in word order, e.g. *de wandelaars, die de rots beklommen hebben* 'the hikers, who climbed the rock' versus *de rots, die de wandelaars beklommen hebben* 'the rock, that the hikers climbed' (Mak, Vonk and Schriefers 2006). This means that in processing these, the choice of analysis as subject or object relative has to be made, unlike English where these are structurally unambiguous.

ing varying strength in different languages (Bresnan, Dingare and Manning 2001). For instance, a language like Lummi has grammaticalized the preference for local (1st/2nd person) subjects, where a passive construction with a demoted local subject is ungrammatical. In English, the avoidance of this type of construction, as in *The car was bought by me*, constitutes a strong statistical tendency (Bresnan, Dingare and Manning 2001). With respect to animacy, the constraint on relative animacy grammaticalized in for instance Jakaltek and MamMaya, is observed as a statistical tendency in corpus data (Dahl and Fraurud 1996; Øvrelid 2004).

Even if categorical animacy effects in language are rare, it is clear that probabilistic effects of animacy have been observed in numerous languages and that frequency data from language use can be employed as linguistic evidence for such a claim. A view of linguistic structure as determined by a set of interacting, soft constraints captures these observations in a comprehensive model.

3.2.4 Gradient animacy

The animacy hierarchy presented in (9) above consists of three categories – human, animate and inanimate. However, these are by no means static and given a priori. We are interested in animacy first and foremost as a linguistic category and how it is reflected in language. The hierarchy therefore reflects the categories which are deemed relevant in linguistic phenomena which are sensitive to the dimension of animacy. We shall see later on that animacy interacts with several other linguistic dimensions in argument differentiation. However, it can be useful to be able to separate out animacy as an independent property which is inherent of nouns. In this respect, we make a distinction between *denotational* as opposed to *referential* properties.

Denotational properties hold for lexemes and are context-independent, i.e., independent of the particular linguistic context, whereas referential properties are determined in context and hold for referring expressions, rather than lexemes (Lyons 1977). These terms are clearly related and are often used interchangeably. Denotational properties are important in reference, and what is referred to in a given context is always within the denotation of at least one lexeme (Lyons 1977). For instance, ‘doctor’ contributes a bulk of semantic information in a referring expression such as ‘her doctor’. Nevertheless, this distinction is useful in discussing gradience within the category of animacy. We thus distinguish between animacy as a denotational property of lexemes, e.g. it is in the denotation of ‘doctor’ that it refers to a human, and animacy as a referential property of referring expressions, e.g., that ‘her doctor’ refers to a

particular human being in a particular context. In particular, we shall see that these need not always coincide.

In section 2.4, we discussed gradience of linguistic categories. We find that the animacy dimension exhibits gradience cross-linguistically, as well as within languages. For linguistic phenomena which are sensitive to animacy, there is a certain degree of variation between languages that seems to be culturally determined. For instance, Persian has been cited to treat trees linguistically as animates (Rosenbach 2002) and number marking in the Papuan language Manam is sensitive to the categories of human and 'higher' animals (Croft 2003). It is suggested then that the animacy hierarchy is "not an ordering of discrete categories, but rather a more or less continuous category ranging from most animate to least animate" (Croft 2003: 130). Collective nouns, like *committee* and *family* pose some interesting problems for semantic theories, due to their dual nature in denoting both a group and a collection of individuals. This duality is reflected in the possibility for collective and distributive predication, respectively. With respect to animacy, collective nouns are candidates for an intermediate animacy status, something which is reflected in annotation schemes for animacy, see section 7.1.1. They also vary in their status cross-linguistically, and Yamamoto (1999) finds that Japanese tends to treat collectives more like inanimates than English.[22]

The fact that a referring expression may be employed to refer to entities of varying animacy, should be kept separate from the gradience of denotation discussed above. With respect to various processes of referential shifts, such as *metaphor* and *metonymy*, the context may override denotational properties in determining reference for an expression:

(12) The **ham sandwich** is sitting at table 20

The classic example in (12), taken from Nunberg 1979, employs metonymy in that it uses "one entity to refer to another that is related to it" (Lakoff and Johnson 1980: 35). It is clear, however, that the inherent, denotational animacy of the noun *ham sandwich* is not gradient, and it is exactly this property which makes non-literal language possible, since metonymy often involves a violation of the semantic selectional restrictions of the verb (Fass 1988).

The example in (12) is an instance of creative metonymy. However, metonymy is also a regular process with a set of conventionalized patterns which recur in language (Lakoff and Johnson 1980). In fact, a corpus study of metonymy in English found that 20% of country names and 30% of organization names were employed metonymically (Markert and Nissim 2006). Commonly occur-

[22] Yamamoto (1999) examines an English-Japanese parallel corpus and performs a contrastive corpus study of the use of referring expressions in the two languages.

ring patterns are 'place-for-people', as in (13), and 'organization-for-members', as in (14) below (Markert and Nissim 2006):

(13) **America** did once try to ban alcohol

(14) Last February **NASA** announced [. . .]

Note, however, that proper names are not traditionally assumed to have a denotation separate from their reference, neither are pronouns (Lyons 1977). In the parlance of truth-conditional semantics, referring expressions do not predicate of their referent. However, processes of metonymy, seem to presuppose a denotation of some kind. We furthermore find these types of regular metonymical extensions for common nouns:[23]

(15) **Kyrkan** *menar att båda dessa riktningar var positiva tillgångar*
church-DEF thinks that both these directions were positive assets
'The church feels that both of these directions were positive assets'

With respect to gradience of animacy, we may ask whether regular metonymic patterns of this type influence the semantics of nouns to such an extent that this is rather a matter of polysemy than referential shifting in a particular context. It is interesting to note that these nouns, as well as the proper nouns, have a collective meaning, a category with gradient animacy properties, as noted above.

In computational semantics, the distributional hypothesis represents a common assumption about meaning which proposes that words with a similar distribution also have similar meanings. On this hypothesis, the denotation of a linguistic element then is reduced to the set of possible contexts in which the element may occur. Chapters 6 and 7 deal with the acquisition of animacy information for nouns based on linguistic distribution. In particular, we will see that a denotational treatment of the category of animacy enables an investigation into the gradience of the animacy dimension which abstracts over individual linguistic contexts.

3.3 Definiteness

Central to a notion of definiteness is the property of *identifiability*, see, e.g., the discussion in Lyons 1999. A referent is identifiable if the hearer is familiar with

[23]The example in (15) is taken from the Swedish treebank Talbanken05. See section 5.1 for a general overview of corpora employed in this thesis and section 5.1.1 for more detail on the Talbanken05 corpus in particular.

identified by	definite specific	indefinite specific	indefinite non-specific
speaker	+	+	−
hearer	+	−	−

Figure 1: The 'identifiability' criterion for definiteness and specificity (von Heusinger 2002: 249)

the referent, or, based on the situation of the utterance, the previous discourse or general background knowledge, the hearer is able to work out the referent of the noun. Another often mentioned characteristic of definiteness, is that it involves an implication of *uniqueness*, i.e. that the referent is in some sense unique in a certain context (Lyons 1999):

(16) I've just been to a wedding. **The bride** wore blue.

Clearly, this is not a matter of the hearer identifying the referent of the definite noun phrase, but rather acknowledging that there is usually only one bride at a wedding.

The following grammatical hierarchy for definiteness is presented in Croft (2003: 132):

(17) Definite > Specific Indefinite > Non-specific Indefinite

Specificity has been a widely studied subject in formal semantics, see von Heusinger 2002 for an overview and references therein. With respect to the earlier mentioned criterion of identifiability, the difference between the three categories in the hierarchy can be schematized as in figure 3.3. Under this criterion, the specificity distinction for indefinites is linked to the speaker having a more precise conception of the referent.

The notion of identifiability has, however, been debated as a criterion for definiteness. Different approaches have provided a more discourse-oriented view of definiteness, highlighting properties such as familiarity and salience, which underline the role of the discourse context in definiteness and how the degree of definiteness may be equated with *cognitive status* (Gundel, Hedberg and Zacharski 1993). A hierarchy focusing on these discourse-oriented properties is represented by the 'givenness hierarchy' in (18) (Gundel, Hedberg and Zacharski 1993: 275):

(18) in focus > activated > familiar > uniquely identifiable > referential > type identifiable

As we shall see in section 3.4, languages exhibit highly conventionalized ways of referring, depending on definiteness, and cognitive status affects the formal realization of referring expressions in a systematic manner.

We distinguish between semantic and formal definiteness. Semantic definiteness can be marked formally in a language in different ways, for instance through morphological marking, but the two are not necessarily isomorphic. In the following we will discuss notions of semantic definiteness and the way definiteness interacts with argumenthood.[24]

3.3.1 Definite arguments

Definiteness is not as commonly recognized as a factor in argument differentiation as animacy. A tendency towards definite subjects has, however, been noted for several languages, both as a categorical constraint influencing morphological marking and as a statistical tendency. Common to these is the same generalization, namely a tendency for subjects to be definite or specific and for objects to be indefinite. In Turkish and Persian, we find Differential Object Marking which is sensitive to definiteness and where definite objects are marked with accusative case, but indefinite objects are not (Croft 2003). A range of languages have been noted to categorically exclude or strongly disprefer non-specific indefinite subjects (Aissen 2003).

There are clear correlations between information-flow in a sentence and argumenthood; subjects tend to represent old information and objects tend to introduce new information. Since subjects tend to precede objects as well, it can be difficult to establish this influence independent of ordering. Weber and Müller (2004) present a corpus study of word order variation in German main clauses, indicating that formal definiteness to a greater extent correlates with grammatical function than linear order, and furthermore that formal definiteness and givenness of information tend to coincide. A discourse-oriented definition of definiteness is also employed in the aforementioned study of the dative alternation in English, where givenness is shown to be a factor in the choice between core and non-core argument realization (Bresnan et al. 2005), cf. examples (10) and (11).

The avoidance of an indefinite subject has been argued to constitute one factor in the choice of existential or presentational constructions in the Scandinavian languages (Sveen 1996; Mikkelsen 2002), illustrated by a Swedish example in (19) and a Norwegian example in (20):[25]

[24] See chapter 4 for more on formal marking of definiteness in Scandinavian.

[25] The Swedish example in (19) is taken from the Talbanken05 treebank and the Norwegian example in (20) is taken from the Oslo Corpus. See chapter 5 for descriptions of these corpora.

(19) *Det finns olika slags* **barnhem**
it exists different sorts orphanages
'There are different kinds of orphanages'

(20) *Det oppsto* **brudd** *mellom stoffet og tankveggen*
it occurred break between substance-DEF and tank-wall-DEF
'A break occurred between the substance and the wall of the tank'

The presentational construction contains an expletive subject and a postverbal, logical subject occurring in object position. The object position in these constructions may only be occupied by an indefinite argument.[26]

3.4 Referentiality

The difference between the denotation and reference of an expression was discussed above and we may furthermore note that there are (at least) three related senses of 'referentiality' in the literature:

1. level of context-dependence
2. specificity
3. meaningfulness

First, referentiality may be employed to make the distinction between referring expressions and elements which are not referring in the sense that they rely only on denotational properties for semantic interpretation. A grammatical hierarchy of referentiality expressing its influence in various linguistic phenomena is presented in (21) (Croft 2003: 130):[27]

(21) pronoun > proper name > common noun

This sense of referentiality, then, relates to the extent to which semantic interpretation requires access to the context of the utterance. This is related to the expression of definiteness, or level of cognitive status, as discussed in section 3.3 above. Pronouns have to be resolved by the context, proper nouns rely on a conventional mapping to a referent, whereas the interpretation of common nouns relies the least on context and more on denotation. Sense 2 (Givón 1984) distinguishes between referential and non-referential indefinites, largely

[26] See section 8.3.1 and examples for more on the distribution of formal subjects in Swedish.

[27] Croft (2003) provides several examples of linguistic phenomena which support the hierarchy in (21). Number marking is often sensitive to this dimension; for instance, pronouns in Usan distinguishes number, whereas common nouns do not (Croft 2003: 128).

synonymous with the distinction of specificity discussed above. Sense 1 and 2 are thus related in that they focus on 'ways of referring' to an entity.

The term 'non-referential' is also employed in the sense 'semantically empty or null' (Sag, Wasow and Bender 2003) and with particular reference to the distinction between referential, as in (22) and non-referential, or expletive, pronouns, as in (23) below:

(22) **It** bothered us for days

(23) **It** is hard to sleep

In the following we will employ the term 'referentiality' in sense 1 above, expressing the degree of context-dependence of an expression. Whereas we recognize, in line with von Heusinger (2002), that specificity differs from definiteness, at least in a discourse-related sense, we will not delve further into a discussion of how to make this finer subdivision. We will furthermore make explicit when referentiality in sense 3 is employed by specifying the application of the term, e.g., 'non-referential *it*'.

3.4.1 Referentiality and arguments

Syntactic arguments differ with respect to their referentiality. As mentioned in section 3.3, the definiteness or cognitive status of an element influences its referentiality. In particular, subjects are likely to be pronominal and objects are more likely to express a lower referentiality (Keenan 1976). The category of pronouns may be further subdivided along the dimension of *person* which distinguishes reference to the speaker and hearer (i.e. discourse participants) from others (Croft 2003: 130):

(24) 1st/2nd (local) person $>$ 3rd person

We find that subjects cross-linguistically tend to be expressed by a local person, and more so than objects. This tendency has been attributed to the 'egocentricity' of human discourse (Dahl 2000) or, less cynically perhaps, to notions of empathy (Kuno and Kaburaki 1977); we tend to speak about ourselves, a fact which is reflected in our choice of referring expressions. The active/passive alternation has been shown to be influenced by person, where passive voice is strongly dispreferred, or even ungrammatical, when the subject is local (Bresnan, Dingare and Manning 2001). Furthermore, the dative alternation is influenced by referentiality and in particular on the distinction between pronominal and non-pronominal expression (Bresnan et al. 2005).

3.5 Relational properties

The above sections have focused on properties of arguments of verbs which are inherent to the arguments. However, an argument is defined as such because it stands in a syntactic relation to a particular predicate. *Relational* properties are properties which are not inherent to the argument itself, but rather describe facets of the relation of an argument to its predicate.

Semantic roles express the semantic relation between an argument and a particular predicate, often expressed as a lexical property of the predicate. There is a consistent relationship between thematic roles and syntactic functions, expressed in syntactic theory as a theory of mapping, e.g., Linking Theory (Baker 1997) and Lexical Mapping Theory in LFG (Bresnan and Kanerva 1989). These mapping theories express a direct relationship between prominence of thematic role and syntactic function.[28] In particular, we find that the most prominent thematic role is mapped to the subject function, the external argument. Central to a mapping of thematic roles to syntactic functions is therefore often a hierarchy of thematic roles, like the one in (25) below (Bresnan 2001: 307):[29]

(25) Agent > Benefactive > Experiencer > Instrument > Theme > Location

The semantic restrictions posed by the verb on its arguments are called *selectional restrictions*.[30] The idea that verbs select for arguments of a specific semantic category has been explored in semantic theories (Katz and Fodor 1963). In lexicalist theories, this notion has been somewhat more developed and rests on the idea that the verb is the head of the clause and is specified lexically for certain selectional restrictions. Animacy has figured in the expression of selectional restrictions from the very beginning. Chomsky (1965) operates with the categories of [±Animate] and [±Abstract] in selectional restrictions for verbs and Katz and Fodor (1963) distinguish the categories of 'Human' and 'Higher Animal' explicitly, as well as their hypernym 'Physical Object'.

[28] Syntactic prominence in theories of semantic role mapping may be expressed as phrase-structural prominence, i.e. c-command or with reference to a hierarchy of syntactic functions, such as the one in (8) above.

[29] Aissen (1999) also makes use of a hierarchy of thematic roles, albeit a simple one, inspired by Dowty (1991): Agent > Patient.

[30] There is variation in the terminology employed in the literature to refer to selectional restrictions. Other terms include 'selectional constraints' (Resnik 1996) and 'selectional preferences' (Erk 2007). We employ the term 'selectional restrictions' in the following as we we take it to be more neutral with respect to formulation (absolute vs. gradient) and application (theoretical vs. computational/applied).

In computational work, automatic acquisition of selectional restrictions, mainly from corpus data, has been further investigated. Resnik (1996) first proposed an approach to acquisition of selectional restrictions with class information taken from the English WordNet (Fellbaum 1998). It is based on an information-theoretic approach to verbal argument selection and quantifies the extent to which a predicate constrains the semantic class of its arguments as its *selectional preference strength*.[31] The approach highlights the gradient nature of the selectional restrictions posed by a verb on its arguments, rather than providing absolute constraints.

3.6 Interaction and generalization

The above sections have presented the various linguistic properties independently. However, it is clear that they interact, in particular with respect to argument differentiation. Various proposals in the literature have also attempted to reduce these factors to one general, more or less explanatory principle. We examine both the interaction and generalization over these linguistic properties below.

3.6.1 Interaction

In the above sections, we have observed very similar patterns for all of the three factors of animacy, definiteness and referentiality, showing differential tendencies with respect to several distinctions in argumenthood. The original animacy hierarchy proposed in Silverstein 1976 included not only information on animacy, but also on referentiality:

(26) 1/2 person pronoun > 3rd person pronoun > proper names > human common noun > animate common noun > inanimate common noun

The above hierarchy combines the factors of animacy and referentiality and provides generalizations over a range of linguistic phenomena, such as number marking (Croft 1990). The fact that this hierarchy has been called an animacy hierarchy is somewhat misleading however; the referents of pronouns are not more animate than the referents of human nouns. However, it is clear that a high level of referentiality often co-occurs with human reference and that this tendency is clearly observable in language use.

Aissen (2003) makes use of a more fine-grained scale for definiteness, incorporating information regarding definiteness and referentiality, as well as a

[31] More details on acquisition of selectional restrictions in section 9.2.8.

notion of specificity to account for Differential Object Marking in languages such as Turkish and Hebrew:

(27) Personal Pronoun > Proper Noun > Definite NP > Indefinite Specific NP > Indefinite Non-Specific NP

Object marking in these two languages exhibit differing cut-off points on the hierarchy; in Turkish object marking distinguishes definites from indefinites, whereas in Hebrew the same distinction is made with reference to specificity.

It is clear that these linguistic factors interact and are interdependent in a way that makes their effect difficult to reduce to a single, well-behaved hierarchy. Rather, these properties yield clusters of properties which tend to go together, exemplified by (28):

(28) Linguistic properties that tend to go together (Dahl 2008: 142)

Animate	Inanimate
Definite	Indefinite
Pronominal	Lexical
Subject	Non-subject
Count	Mass
Proper	Common
Rigid designation	Non-rigid designation
Independent reference	Dependent reference
Proximate	Obviative
Agent	Non-agent

The generalizations embodied in the different grammatical hierarchies presented above, as well as in the list of properties in (28) have been modeled as constraints in recent work in OT, which reinterprets the grammatical hierarchies as expressing the relative *prominence* of an element. Prominent elements on one hierarchy tend to attract prominent elements of another, hence subjects will tend to be animate, definite, agentive etc. This tendency has been modeled formally employing the technique of *harmonic alignment* of constraint hierarchies (Aissen 1999, 2003).[32] The generalizations mentioned in the preceding

[32] Harmonic alignment aligns the dominant elements of a scale with the dominant elements of another and the lower ranked elements of one scale with the lower ranked of another, expressing the idea that prominence on one scale will attract prominence on another. Markedness constraints are derived by reversing the output scales from the alignment and adding the 'avoid'-marker, '*', to them:

> Suppose given a binary dimension D1 with a scale X > Y on its elements {X, Y}, and another dimension D2 with a scale a > b ... > z on its elements. The harmonic alignment of D1 and D2 is the pair of harmony scales:

sections regarding properties of the subject, for instance, are thought to be a result of constraints which express the relative markedness of combinations of properties, such as the constraints on the animacy of subjects in (29):[33]

(29) SUBJ/HUM > SUBJ/ANIM > SUBJ/INAN

The main idea is that these constraints interact with other constraints on argument expression in various languages, for instance constraints on morphological marking or word order, but that their internal ranking is fixed. This predicts that inanimate subjects will cross-linguistically be more marked than animate subjects.

As mentioned in section 2.1.1, the relation between typological markedness and frequency in language is obvious and has even lead to proposals reducing this notion of markedness to frequency effects (Haspelmath 2006). As a consequence, the modeling of various factors in argument differentiation may employ constraints which are grounded directly in acquired frequency effects, such as the BIAS-constraint, first suggested by Zeevat and Jäger (2002) and employed by Jäger (2004); de Swart (2007), among others:

BIAS**:** prefer the normal reading, the reading that is available in most cases

The normal reading is thus the reading that is most likely according to the statistical tendencies described in (28).

3.6.2 A more general property

As mentioned above, the interaction of the various linguistic properties in argument differentiation has been attributed to their prominence in individual grammatical hierarchies.[34] It is not immediately clear, however, what this notion of prominence actually entails. Without a clear definition of prominence, generalizations which make reference to this are at risk of simply restating the

---

H_x: X/a > X/b > … > X/z
H_y: Y/z > … > Y/b > Y/a
The constraint alignment is the pair of constraint hierarchies:
C_x: *X/z ≫ … ≫ *X/b ≫ *X/a
C_y: *Y/a ≫ *Y/b ≫ … ≫ *Y/z
(Prince and Smolensky 1993, as quoted in Aissen (2003: 441))

[33] For illustrative purposes, we show only the harmonically aligned scales in (29), rather than the markedness constraints resulting from the reversal and negation of these scales. See Aissen 2003 for details.

[34] Formally, the interaction is modeled as the cross product of the constraint subhierarchies (Aissen 2003).

question. There have, however, been several proposals attempting to reduce the effects observed in conjunction with argument differentiation to a more general linguistic property or even properties of our cognitive abilities or language processing in general.

3.6.2.1 Individuation

The hierarchy in 26 above combines the categories of definiteness and referentiality, highlighting different aspects of linguistic reference and may be explained by appealing to a notion of *individuation*: "the degree to which the interpretation of an NP involves a conception of an individuated entity" (Fraurud 1996). Fraurud suggests that our cognitive ontology distinguishes between:

- Individuals, e.g., *Gabriel*
- Functionals, e.g., *the postman*
- Instances, e.g., *a glass of wine*

This ontology is at a more general level than the factors reviewed above but influences the choice of NP form. In particular, animate entities tend to be perceived as Individuals and inanimate as Instances. Individuals are the most individuated and are typically named, whereas Functionals on the other hand are conceived of in relation to some other entity in a part-whole relation and are typically definite. Instances, finally, are instantiations of general types and are thus the least individuated and they are typically indefinite 'type descriptions'.

Fraurud distinguishes between two main types of knowledge which are necessary for the interpretation of NPs – type knowledge and token knowledge. Token knowledge is contextually determined and requires previous knowledge about the referent of an expression, whereas type knowledge relies on lexico-encyclopedic knowledge. Token knowledge is relevant only for the identification of Individuals, whereas for the other two types of ontological categories, Functionals and Instances, type knowledge is sufficient. This relates to our earlier discussion on animacy and nouns in section 3.2.4. Nouns typically express Functionals or Instances which rely on type knowledge for interpretation, clearly related to the notion of being a denotational property.

3.6.2.2 Accessibility

In experimental work on language production, cognitive status has been proposed as an explanation for the tendencies observed above. Branigan, Pickering and Tanaka (2008) appeal to a notion of *conceptual accessibility*: "[T]he

ease with which the mental representation of some potential referent can be activated in or retrieved from memory" (Bock and Warren 1985: as cited in Branigan, Pickering and Tanaka 2008), and distinguish between *inherent* and *derived* accessibility. Inherent accessibility is invariant across contexts and is a direct consequence of the number of conceptual relations an entity may partake in, also known as its *predictability*. Animate entities are assumed to be more predictable than inanimate ones, hence have a high inherent accessibility.[35] The derived accessibility of an entity is temporary and context-specific, influenced by factors such as semantic priming and discourse status (Prat-Sala and Branigan 2000). Effects of animacy, both on word order and argument differentiation, are then explained with reference to conceptual accessibility; an animate entity is inherently accessible, and often also derived accessible by being definite and referential.

Following from this generalization, the factors influencing argument differentiation can be seen to amount to conceptual accessibility. Conceptually accessible entities are thought to be retrieved first and assigned syntactic function first. It is assumed that syntactic functions are assigned incrementally following a hierarchy of grammatical functions, like the one in (8) above, a fact which accounts for animacy effects found on word order, as well as syntactic function assignment.

3.6.2.3 Agentivity

In our earlier discussion on semantic roles in section 3.5, we noted that subjects often stand in an *agentive* relation to their predicate and animacy and agentivity are therefore strongly related. An important property of agents is their control over and sentience of an event (Dowty 1991). It has been argued that agentivity presupposes animacy (Hundt 2004). This depends somewhat on the notion of agentivity employed, however, and in particular on the treatment of causation. It is well known that many languages can have inanimate natural force subjects, e.g. *the storm broke the window*, and theories of thematic roles differ in the inclusion of these as agents proper.[36]

It is clear that there is no isomorphism between agentivity and animacy in general. As we have seen, animacy is an independent factor in a range of lin-

[35]The noted gradience of the animacy dimension is assumed to stem from gradience of predictability – some animate entities are more predictable, e.g. humans than others, e.g. jellyfish (Branigan, Pickering and Tanaka 2008).

[36]In Fillmore's case grammar (Fillmore 1968), for instance, inanimate causers are treated as agents. Yamamoto (1999), in contrast, operates with a notion of agenthood which does not include causers, following Dik (1989), hence is therefore more closely connected to animacy.

guistic phenomena, where it seems unlikely that agentivity is the determining factor, e.g. number and case marking of objects. However, with respect to the strong correlation between animacy and subjects, agentivity is clearly an important factor. Semantic roles are relational categories, just like syntactic functions, hence are closely related, but not overlapping. In this sense, explanation by means of agentivity adds a more semantic dimension to the generalization, but does not explain the correlations with definiteness and referentiality any further than the level of syntactic functions itself. Moreover, animacy provides a surface-oriented, lexical constraint on syntactic function assignment which embodies the semantic dimension of agentivity. This will prove to be important in the following.

4 PROPERTIES OF SCANDINAVIAN MORPHOSYNTAX

Languages differ in the way they encode grammatical functions. It has been noted that "morphology competes with syntax" (Bresnan 2001: 6) in that there is largely an inverse relationship between the extent of morphological marking and the degree of word order variation. Languages which encode arguments largely through morphological marking exhibit freer word orders, whereas languages which primarily employ structural positions to encode grammatical functions, so-called *configurational* languages, necessarily exhibit limited word order variation. Most languages, however, are somewhere in between these two extremes with respect to the balance between morphology and syntax in argument encoding. For instance, the Scandinavian languages have limited morphological marking of syntactic functions, but allow for variation in word order which makes for an interesting comparison with more configurational languages, like English.

In this chapter we examine some relevant characteristics of the Scandinavian languages. Particular focus will be on various properties of syntactic arguments and on variation in their categorial, morphological and structural encoding. For more detailed overviews of the Scandinavian languages see, e.g., the Norwegian reference grammar (Faarlund, Lie and Vannebo 1997), the Swedish reference grammar (Teleman, Hellberg and Andersson 1999) or an English overview in Holmes and Hinchliffe 2003.

4.1 Morphological marking

The distinction between various types of arguments is partially encoded through *case* marking in Scandinavian. Nominal arguments are furthermore inflected for other categories, such as *definiteness*.

4.1.1 Case

Morphological case explicitly encodes grammatical function. The Scandinavian languages make limited use of case marking, and, in this respect, resemble English. Pronouns are marked for case, but exhibit syncretism and syntactic variation, whereas nouns distinguish only genitive case and are otherwise invariant for case.

4.1.1.1 Core arguments

Personal pronouns distinguish nominative, accusative and genitive case.[37] In the Swedish examples in (30)–(32), we see that case unambiguously signals syntactic function for the first person pronoun *jag/mig* 'I/me'. It is a subject in (30) and a direct object in (32)). Assignment of syntactic function takes place irrespective of the position of this argument, which is preverbal in the case of (30) and postverbal in (31) and (32).

(30) *Jag* *såg* *den*
I-NOM saw it-Ø
SUBJ OBJ
'I saw it'

(31) *Den* *såg* *jag*
it-Ø saw I-NOM
OBJ SUBJ
'It, I saw'

(32) *Den* *såg* *mig*
it-Ø saw me-ACC
SUBJ OBJ
'It saw me'

[37] In the following we will adhere to a rather liberal definition of pronouns, following Teleman, Hellberg and Andersson 1999, which includes:

1. definite pronouns - personal, e.g. *han* 'he', *honom* 'him', demonstrative, e.g. *denna* 'this', reflexive e.g. *sig* 'him/her/itself', reciprocal pronouns e.g. *varandra* 'each other'
2. interrogative pronouns, e.g. *vem* 'who', *vilken* 'which'
3. quantifying pronouns, *alla* 'all', *någon* 'some'
4. relational pronouns - comparative, *samma* 'same', ordinal, *första* 'first'

Case marking is not, however, always unambiguously indicative of syntactic function. For instance, in both Swedish and Norwegian, the third person singular pronouns *det*, *den* 'it' have the same form for nominative and accusative case. Quantifying pronouns, like *alla* 'all', *många* 'many' are also invariant for case. In the examples in (33)–(34), in contrast to (30)–(32) above, case does not indicate syntactic functions for the pronominal argument *it* 'den' and the proper noun *Lisa*.

(33) *Lisa* *såg* *den*
Lisa-Ø saw it-Ø

(34) *Den* *såg* *Lisa*
it-Ø saw Lisa-Ø

In Norwegian, nominative form is preferred when a pronoun is stressed, regardless of syntactic function (Johannessen 1998). So, when followed by for instance a relative clause, the pronoun will be in its nominative form even when functioning as an object.[38] In example (35) we see a nominative pronoun functioning as object, whereas the same pronoun functions as subject in (36):[39]

(35) *Dette* *gjelder* *i* *tillegg* **de** *som* *håndterer* …
this concerns in addition they-NOM who handle …
'This also concerns those who handle …'

(36) **De** *som* *fortsatt* *tror* *at* *idyllen* *kan* *bevares*
they-NOM who think that the idyll can maintain-PASS
[…] *tar* *alvorlig* *feil*
[…] take seriously wrong
'Those who still believe that the idyll can be maintained […] are seriously mistaken'

The same tendency for the plural 3rd person with relative clause modification *de som* … 'they who' has been noted in written Swedish as well (Teleman, Hellberg and Andersson 1999: vol. 2, 299). In Swedish spoken language and casual writing, the 3rd person plural pronoun is realized as *dom* 'they/them' in both subject and object function.

[38] Note that the nominative form is employed also when the pronoun is modified by a prepositional postnominal modifier, so the preference for nominative case is not due to the argument's subject status in the relative clause.

[39] The examples in (35)–(36) are taken from the Norwegian Oslo Corpus, see section 5.1.3.

4.1.1.2 Arguments and determiners

Genitive case signals a nominal's status as determiner. Definite and personal pronouns distinguish genitive case formally, e.g. *min*, *hans*, *deras* 'my, his, their', whereas proper and common nouns do not distinguish nominative and accusative case, but may be marked for genitive case with the suffix *-s*, e.g., *Gabriels* 'Gabriel's', *doktorandens* 'the PhD-student's'.[40]

We may furthermore distinguish between nominal and attributive pronouns, where nominal pronouns are pronouns which function as independent arguments, and the attributive pronouns are determiners. However, this distinction is blurred by the fact that many pronouns may function as both, and are then formally identical. This is true for most of the pronouns which do not distinguish case, see section 4.1.1.1 above. For instance, the pronoun *den* 'it' may function as a subject, as in (37), and as a definite determiner, as in (38), where it modifies a common noun:[41]

(37) *Sedan somnar* **den**
later sleeps it
'Later, it falls asleep'

(38) **Den** *vetenskap som sysslar med dessa kallas psykiatri*
the science which deals with these call-PASS psychiatry
'The science which deals with these matters is called psychiatry'

The neuter form *det* 'it' exhibits the same functional variation and in addition may also function as expletive subject and object.[42]

4.1.2 Definiteness

In section 3.3 we examined semantic definiteness and discussed criteria for definiteness, including identifiability and more discourse-pragmatic notions related to the cognitive status of an element. The Scandinavian languages mark

[40] There are alternative genitive constructions which are remniscient of the genitive alternation found e.g. in English. For instance, in Norwegian *jentas bror* 'the girl's brother' and *broren til jenta* 'the brother of the girl'. The alternation is less common in standard Swedish which typically uses the genitive suffix in these cases. In both languages, however, there is the possibility of expressing part-whole relations in terms of prepositional modification with the preposition *på* 'on': *taket på huset* 'the roof on the house'. We will have more to say about differential properties of genitive constructions in chapter 6 and 9.

[41] The examples in (37)–(38) are taken from the Swedish treebank Talbanken05, see section 5.1.1.

[42] See section 8.3.1 and examples therein for a corpus study of the different argument relations in Swedish, including formal subjects.

definiteness morphologically, but formal definiteness is not completely isomorphic with semantic definiteness.

Nouns are marked for definiteness by a definite suffix, e.g. *bil-en* 'car-DEF', *hus-et* 'house-DEF'.[43] There is agreement for definiteness within the noun phrase, governed by the nominal head:

(39) *det gamla året*
the old-DEF year-DEF
'the old year'

The definite suffix is not, however, necessary, nor sufficient for semantic definiteness. Noun phrases may be semantically definite without definite marking on the noun when rendered definite by properties of the construction, e.g., by a definite determiner. For instance, genitive determiners, as in (40), some definite determiners, as in (41), as well as the universal quantifier may combine with an indefinite noun to form a semantically definite noun phrase:[44]

(40) *Gabriels* **bil**
Gabriel-GEN car
'Gabriel's car'

(41) *Den* **bil** *som Gabriel äger*
the car which Gabriel owns
'the car that Gabriel owns'

There are also nouns with definite marking which are not semantically definite. In particular, definite nouns may be employed with generic reference, to refer to instances as a type, rather than a particular instance:

(42) **Lejonet** *är Afrikas största köttätare*
lion-DEF is Africa-GEN largest carnivore
'The lion is Africa's largest carnivore'

4.2 Word order

The classical descriptive model for Scandinavian word order is based around organization into so-called *topological fields* (Diderichsen 1957). The topological fields approach separates the clause into, roughly speaking, three parts: the

[43] The particular definite suffix is determined by the *gender* of the noun.

[44] Scandinavian noun phrases and definiteness is an intriguing subject which we will not aim to cover in the current context. For instance, nouns with definite marking may occur as a bare noun phrase, e.g, *bilen* 'car-DEF', but may also be specified by a definite determiner, exhibiting so-called "double definiteness", e.g. *den bilen* 'that car-DEF'. See Börjars 1998 for an in-depth study and analysis of Scandinavian noun phrases.

initial field, the *mid field* and the *end field* (Teleman, Hellberg and Andersson 1999):

(43)

	Initial	**Mid**	**End**
MAIN	*I morgon*	*kan hon inte*	*vara med vid sammanträdet*
	tomorrow	can she not	be with at meeting-DEF
SUBORD	*eftersom*	*hon inte kan*	*vara med vid sammanträdet*
	since	she not can	be with at meeting-DEF

Note that the topological fields are not constituents in the phrase structural sense, and do not pass standard constituency tests such as topicalization. For syntactic theories which propose a separation between functional structure and linearization,[45] however, topological fields provide a natural extension for expressing linearization in Germanic languages.[46] The separation into topological fields enables generalization over the word order patterns in the Scandinavian languages, capturing some key properties regarding the positioning of the verb, as well as positioning of arguments and adverbials across various clause types. Some relevant properties of Scandinavian syntactic structure captured in the fields approach are summarized below.

4.2.1 Initial variation

The initial position is characterized by a great deal of variation. It has been claimed to mark the syntactic-semantic type of the clause and is closely related to the speech act expressed by the clause (Platzack 1987). Moreover, the initial constituent is often topical, in the sense that it links the sentence to the preceding context.[47] Most clausal constituents may occupy initial position in declarative main clauses, e.g., subjects (44), direct objects (45) and adverbials (46). Constituent questions contain a *wh*-word in initial position, as in (47).

(44) *Statsministern* *håller* *talet* *i* *morgon*
primeminister-DEF holds speech-DEF in tomorrow
'The primeminister gives the speech tomorrow'

[45] This is true of LFG (Bresnan 2001), most flavours of dependency grammar (Sgall, Hajičová and Panevová 1986; Mel'čuk 1988; Hudson 1990), as well as some versions of HPSG, e.g., Pollard and Sag 1994.

[46] See Ahrenberg (1990) for an early formalization employing regular expressions over a constituent-based analysis constituting a separate level (*t*-structure) within an LFG grammar, and Bröker 1998 for an implementation of a dependency grammar with ordering by topological fields introduced as so-called *metacategories*.

[47] The realization of an argument in initial position is referred to as *topicalization* and is thought of as movement to clause-initial position in transformational theories.

(45) *Talet* *håller* *statsministern* *i* *morgon*
speech-DEF holds primeminister-DEF in tomorrow
'The speech, the primeminister gives tomorrow'

(46) *I* *morgon* *håller* *statsministern* *talet*
in tomorrow holds primeminister-DEF speech-DEF
'Tomorrow, the primeminister gives the speech'

(47) *När* *håller* *statsministern* *talet?*
when holds primeminister-DEF speech-DEF
'When does the primeminister give the speech?'

The initial position may also be empty. Imperative clauses and yes/no-questions are verb-initial in Scandinavian, cf. (48)–(49).

(48) *Håll* *talet* *i* *morgon!*
hold speech-DEF in tomorrow
'Give the speech tomorrow!'

(49) *Håller* *statsministern* *talet* *i* *morgon?*
holds primeminister-DEF speech-DEF in tomorrow
'Does the prime minister give the speech tomorrow?'

4.2.2 Rigid verb placement

Like the majority of Germanic languages, but unlike English, the Scandinavian languages are *verb second (V2)*; the finite verb is the second constituent in declarative main clauses, see (44)–(47) above. Subordinate clauses, however, are not V2:

(50) ...*eftersom* *statsministern* *nog* *inte* **håller** *talet* *i* *morgon*
since primeminister probably not holds speech-DEF in tomorrow
'...since the prime minister probably will not give the speech tomorrow'

Non-finite verbs follow the finite verb, but precede their complements.[48] Neither of the elements in the end field, i.e., the non-finite verb, followed by various objects and adverbials, are obligatory. In fact, the only obligatory element

[48] In this respect Scandinavian differs from German, which positions non-finite verbs in clause final position.

in the clause is the finite verb, and, with a few exceptions, the subject. However, the presence of a non-finite verb introduces a greater rigidity in terms of positioning and interpretation of the clausal constituents.[49] With respect to arguments, only subjects may intervene between a finite and non-finite verb, as in (52), and, as mentioned already, only objects may follow the non-finite verb, as in (51):

(51) *Statsministern* **ska hålla** *talet*
primeminister-DEF shall hold speech-DEF
'The primeminister will give the speech'

(52) *Talet* **ska** *statsministern* **hålla**
speech-DEF shall primeminister hold
'The speech, the primeminister will give'

Main clauses consisting of a finite, transitive verb along with its arguments are structurally ambiguous, as in (53), whereas the placement of a non-finite verb in the same clause clearly indicates syntactic functions, as in (54)–(55):

(53) *Vem såg Ida?*
who saw Ida
'Who saw Ida / Who did Ida see?'

(54) *Vem har sett Ida?*
who has seen Ida
SUBJ OBJ
'Who has seen Ida?'

(55) *Vem har Ida sett?*
who has Ida seen
OBJ SUBJ
'Who has Ida seen?'

These rigid placement constraints extend also to particles and prepositional modifiers of the finite verb:

(56) *Vem kom ihåg Ida?*
who came in-memory Ida
SUBJ OBJ
'Who remembered Ida?'

[49] This fact has been taken as an indication that Swedish exhibits evidence for a VP only in clauses with a non-finite lexical verb (Andréasson 2007). Such an analysis clearly questions the status of Scandinavian as a strictly configurational language.

(57) *Vem kom Ida ihåg?*
who came Ida in-memory
OBJ SUBJ
'Who did Ida remember?'

4.2.3 Variable argument placement

The generalization that most constituents may occupy sentence-initial position entails that they have two alternative positions – initial position and a non-initial position. A schematized version of the predictions of the fields analysis with respect to the linearization of verbs and (non-initial) arguments in main clauses is provided in (58) below (Engdahl, Andréasson and Börjars 2004):[50]

(58) Linearization of grammatical functions in declarative, main clauses:
XP | V_{fin} SUBJ S-ADV | $V_{non-fin}$ OBJ_{ind} OBJ_{dir} ADV

The subject, for instance, may occupy either the initial position or the position immediately following the verb. Note that the fields analysis does not capture the generalization that the subject is the most common initial constituent. The basic word order of a language is "typically identified with the order that occurs in stylistically neutral, independent, indicative clauses [...], it is the ordering of constituents in prototypical transitive clauses". (Siewierska 1988: 8). In this respect, the Scandinavian languages must be said to be SVO languages.

Subordinate clauses differ from the schema in (58) in that they have a different ordering of the arguments with respect to the finite verb in the mid field:

(59) Linearization of grammatical functions in subordinate clauses
subj | SUBJ S-ADV V_{fin} | $V_{non-fin}$ OBJ_{ind} OBJ_{dir} ADV

A uniform analysis of main- and subordinate clauses has been proposed under the assumption that the subjunction and the finite verb are instances of the same category (C_0), which expresses the finiteness of the clause (Platzack 1987).

[50] Note the similarity with the hierarchy of grammatical functions presented in section 3.1. There we made a distinction between primary and secondary objects, following Bresnan 2001. The more traditional terms 'direct' and 'indirect' object will however be employed in the following. As we remember, primary objects denote indirect objects and objects of monotransitive verbs and secondary objects denote the direct objects of ditransitive verbs. On this mapping, the ordering in the schema in 58 corresponds directly to the one proposed in the hierarchy. Note however, that the original hierarchy in Keenan and Comrie 1977 propose the reverse ordering (OBJ<IOBJ) based on accessibility for relativization.

4.2.4 More variation

The topological fields model is first and foremost a descriptive model and its predictions are not formally explicit. In particular, constituent optionality results from variation in the schemas in (58) and (59). It has been noted several places that absence of a non-finite verb, or a "verbal frame" (Rahkonen 2006), causes a greater variation in the positioning of constituents. For instance, the phenomenon known as 'object shift', whereby an unstressed pronominal object may precede a sentential adverbial, does not take place when there is a non-finite verb (Holmberg 1986).

There is also a greater variation between the constituents in the mid field than has earlier been acknowledged and entailed by the fields analysis (Börjars, Engdahl and Andréasson 2003; Engdahl, Andréasson and Börjars 2004; Andréasson 2007). Andréasson (2007) proposes an analysis with free variation in the mid field, which is structurally external to a VP containing non-finite verbs and complements. The ordering of subject and adverbials is determined by the interaction of a set of syntactic, semantic and pragmatic constraints.

5 RESOURCES

This chapter introduces the resources employed in the following two parts of the thesis. We examine corpora, machine learning algorithms, as well as external software employed in the experiments described in Parts II and III. More detailed overviews of lexical acquisition, which is the topic of Part II, and syntactic parsing, which is the topic of Part III, are presented in the introduction to the respective parts.

5.1 Corpora

As stated earlier in chapter 1, this thesis has a largely empirical focus. The data employed both for linguistic investigations and as basis for automatic induction are corpora. The corpora differ in language, size and levels of annotation and will be introduced in the following section. Talbanken05 is a treebank, hence contains syntactic annotation which has been manually checked (Abeillé 2003). As a practical consequence, it is also quite small and we have therefore chosen to supplement it with two considerably larger, automatically annotated corpora – the Swedish Parole corpus and the Norwegian Oslo Corpus.

5.1.1 Talbanken05

Talbanken05 is a Swedish treebank containing approximately 300,000 tokens[51] of both written and spoken language. It was created in the 1970s (Einarsson 1976a, b) and converted to dependency representation in 2005 (Nilsson, Hall and Nivre 2005; Nivre, Nilsson and Hall 2006). Talbanken05 is freely available,[52] and as table 5.1 shows, the treebank contains material from various sources and of varying modality.

[51] The counts for running tokens include punctuation.

[52] http://w3.msi.vxu.se/ nivre/research/Talbanken05.html

Talbanken05		
Modality	Type	Tokens
Written	professional prose	97335
	student essays	99788
Spoken	conversation/debate	58341
	interviews	86725
Total		342209

Table 5.1: Material in Talbanken05 by modality and type.

The annotation in the original Talbanken, the MAMBA scheme, is described in detail in Teleman 1974. It consists of a column-based markup, where two main *layers* may be distinguished – a lexical and a syntactic one. The annotation in Talbanken05 is a result of a conversion from the original annotation, which is a mixture of constituent- and dependency based analysis, to a pure dependency analysis. The conversion has been performed by way of a flat phrase structure representation, see Nilsson and Hall 2005 for more detail.[53]

5.1.1.1 A note on dependency grammar

As Nivre (2006) notes, there is no uniform theory of dependency grammar. Common to all dependency-based grammar theories, however, is the notion of *dependency* – a binary, asymmetrical relation between lexical items or words. Each word in a sentence has a head or governor of which it is a dependent (Mel'čuk 1988).[54] The dependency relation which holds between two words may or may not be labeled and its participants, the head and dependent, may or may not be ordered. Many of the theoretical proposals of dependency grammar separate dependency structure from word order (Mel'čuk 1988; Sgall, Hajicŏvá and Panevová 1986).

A notion of syntactic *head* is central to most syntactic theories, and in particular the lexicalist theories mentioned initially in chapter 3. The head, it is assumed, has lexically specified subcategorization requirements which must

[53]The conversion relies on head-finding rules in cases when locating the head is non-trivial based on the phrase structure representation, e.g., in locating the head of main clauses. Note however, that, as opposed to conversion from a phrase-structure representation like the one in the Penn treebank (Marcus, Santorini and Marcinkiewicz 1993), the original annotation in Talbanken contains information on head-status for a majority of constituents, making conversion considerably more reliable.

[54]There are however, dependency representations which allow for more than one head per dependent, see e.g., Hudson 1990.

be fulfilled for sentence wellformedness. Further criteria for syntactic head status mention its possibility to replace the head and its dependent, its obligatoriness, government of agreement and that dependents often are ordered with respect to the head. It is clear, however, that none of these criteria apply to all heads and they feature a mixture of morphological, syntactic and semantic criteria (Nivre 2006). For instance, the criterion of replacement applies only at a phrasal level, and not to clauses.

The *projectivity* of the dependency tree is another issue where proposals for dependency-based analysis differ. This difference obtains mainly between the largely theoretical and the formal or computational approaches to dependency analysis. Projectivity obtains between two words A and B, where A depends on B, if "all words between A and B are also subordinate to B" (Covington 2001: 3). In short, this amounts to disallowing crossing branches in the dependency tree.[55] The dependency analysis in Talbanken05 consists of *projective dependency graphs* (Nivre 2006).[56] Projective dependency graphs are labeled directed acyclic graphs with the following properties:

Root The dependency graphs have a designated root node.

Connected The dependency graph is (weakly) connected.[57]

Single head Dependents have exactly one head.

Projective The dependency graph is projective.

Figure 2 shows the labeled dependency graph of example (60), taken from Talbanken05.

(60) *Därefter* *betalar* *patienten* *avgift* *med* *10* *kronor* *om*
thereafter pays patient-DEF fee with 10 kronas in
dagen
day-DEF
'Thereafter, the patient pays a fee of 10 kronas a day'

For each token, Talbanken05 contains information on word form, in row 1 in figure 2, a coarse and more fine-grained part-of-speech tag, in rows 2-3 in figure 2, head and dependency relation, as well as various morphosyntactic and lexical semantic categories, presented in row 4 in figure 2. The root of the dependency tree occupies position 0 and is denoted by '_'.

[55] To be precise, projectivity amounts to disallowing crossing branches only under the assumption that there is a single, artificial root node.

[56] Note that Talbanken05 contains some non-projective structures – approximately 1% of all dependencies are non-projective and 9.8% of all sentences contain a non-projective structure.

[57] Weak connectivity for directed graphs indicate that the underlying, undirected graph is connected, i.e. every pair of nodes are connected by an (undirected) path.

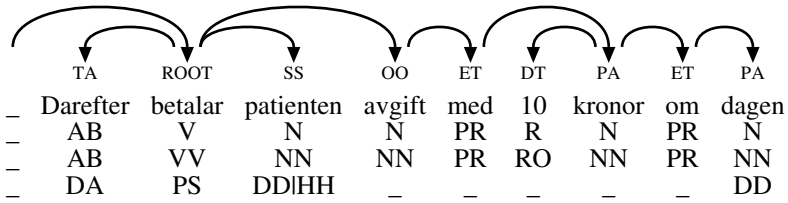

Figure 2: Dependency representation of example from Talbanken05.

Part-of-speech	Annotation
noun (N):	*definiteness, person reference, case*
pronoun (PO):	*pronoun type, person reference, case*
adjective (AJ):	grade of *comparison, person reference, case*
verb:(V)	*tense, voice*
participle (TP/SP):	*person reference, voice*
adverb (AB):	*semantic class*, e.g., definite, temporal, interrogative
preposition (PR):	grade of *comparison*
conjunction (++):	*semantic class*, e.g., disjunctive, explanatory
subjunction (UK):	*semantic class*, e.g., temporal, causal

Table 5.2: Overview of lexical semantic and morphosyntactic annotation in Talbanken05 by part-of-speech.

5.1.1.2 Lexical categories

The nature of the morphosyntactic and lexical semantic information in the treebank varies depending on the part-of-speech of the tokens, as illustrated by the overview in table 5.2. For the lexical semantic and morphological categories, lack of annotation in Talbanken05 is iconic in the sense that it conveys lack of the property in question, e.g., indefinites lack definiteness, non-genitive nouns are morphologically unmarked for case etc. In figure (2), we see that the person-denoting, definite noun *patienten* 'patient-DEF' is annotated explicitly as such with the tags DD (definite) and HH (person referring), whereas the noun *avgift* 'fee' on the other hand bears a null value, since it does not refer to a person, and is in furthermore indefinite and unmarked for case.

5.1.1.3 Syntactic categories

The syntactic annotation in Talbanken05 contains a rich set of dependency relations, expressing a range of distinctions regarding argumenthood, various types of adverbials and modification etc. A dependency grammar analysis, expressing relations between lexical elements only, does not distinguish between clausal arguments and non-arguments structurally, since both types of relations are verbal dependents. They differ, theoretically, in terms of subcategorization by the verb, see the discussion in section 3.1. In the dependency representation in Talbanken05, arguments and non-arguments are distinguished primarily by dependency label. Table 5.3 presents the dependency relations found in Talbanken05. We distinguish the following groups of dependency relations in table 5.3:

Arguments Relations pertaining to elements that are subcategorized for and/or thematically entailed by a predicate.

Non-Arguments Relations pertaining to elements that are not subcategorized for by a predicate, hence are optional.

Verbal relations Relations pertaining to verbs and verb groups.

Coordination Relations pertaining to coordination or subordination of elements

Other Relations pertaining to miscellaneous other types of elements, e.g. determiners, infinitive markers, punctuation

The dependency relations which are not grouped as arguments in table 5.3 denote relations which are not subcategorized for by the predicate of a clause.[58] Coordination is a special case in this respect, since coordination takes place between most types of constituents, whereby the resulting coordinated structure may occupy a range of different relations. Coordination is thus a metaoperation which should be treated separately. See Nivre 2006 for an overview of the treatment of coordination within dependency grammar. In the analysis of coordination found in Talbanken, the first conjunct is the head of the coordination, whereby other conjuncts (CC) are dependents of the first conjunct.

Syntactic theories differ in the way they treat so-called functional categories. In the analysis in Talbanken05, we find a set of dependency relations

[58] Object adverbials (OA) are adverbials which are closely related to the verb, much like objects, without necessarily being subcategorized for by the verb (Teleman 1974). Since this category contains a mix of subcategorized and non-subcategorized elements, we group these with non-arguments. See (132) in section 8.2.3 for an example of the OA-relation.

Class	DepRel	Explanation
Arguments	AG	demoted passive agent
	EO	logical object
	ES	logical subject
	FO	formal object
	FS	formal subject
	IO	indirect object
	OO	direct object
	OP	object predicative
	SP	subject predicative
	SS	subject
	VO	object-with-infinitive (small clause)
	VS	subject-with-infinitive (small clause)
Non-arguments	AA	adverbial
	AN	apposition
	AT	nominal (adjectival) pre-modifier
	CA	contrastive adverbial
	EF	relative clause in cleft
	ET	nominal post-modifier
	KA	comparative adverbial
	MA	modal adverbial
	NA	negation adverbial
	OA	object adverbial
	PT	participial attribute
	RA	place adverbial
	TA	time adverbial
Coordination	++	coordinating conjunction
	CC	second conjunct
	+A	conjunctional adverbial
	+F, MS	main clause coordination
	VA	dual coordination adverbial
Verbal relations	VG	non-finite verb in verb group
	PL	verb particle
Other	DB	doubled function
	DT	determiner
	PA	complement of preposition
	I{C,G,K,P,Q,R,S,T,U}	misc punctuation
	ID	part of multiword unit
	IM	infinitive marker
	J{C,G,R,T}	misc punctuation
	ST	paragraph
	UK	subordinating conjunction
	X{A,F,T,X}	misc discourse units

Table 5.3: Overview of the dependency relations in Talbanken05.

which in other theories would be known as functional heads, e.g. complementizers, determiners and infinitival markers. In the dependency analysis, functional categories are dependents of their lexical heads. For instance, with respect to the analysis of subordinate clauses, the verb is assigned status as head of the clause, providing a uniform analysis of main and subordinate clauses.[59] Functional elements are thus dependents of the verbal head in the dependency analysis. Figure 3 shows the dependency analysis for example (61) where a subordinate clauses functions as a direct object in the main clause:

(61) *Kontrollera att defrostern fungerar bra*
control that defroster-DEF works well
'Check that the defroster works well'

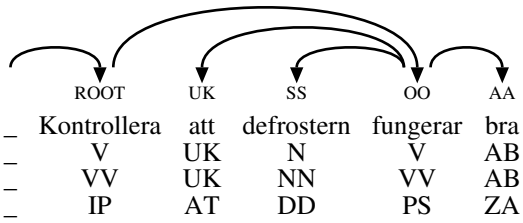

Figure 3: Dependency representation of example with subordinate clause from Talbanken05.

The artificial root node takes as dependent the head of the sentence, usually the finite verb of a matrix clause, which is assigned the relation ROOT. For subordinate clauses, the main verb is assigned the dependency relation which the subordinate clause holds with respect to its matrix clause. For instance, in the example in figure 3, the subordinate clause headed by the finite verb *fungerar* 'works' functions as a direct object (OO).

5.1.2 Parole

The Swedish Parole corpus was collected within the context of the EU project Parole which ended in 1997 and the corpus is freely available for research purposes.[60] It contains approximately 21.6 million tokens, including punctu-

[59] The annotation of clausal complements in Talbanken05 shows that these may exhibit a range of different argument and non-argument functions, e.g. subject, object, temporal adverbial, unlike e.g. LFG where clausal arguments (COMP, XCOMP) are distinguished functionally from clausal adjuncts (XADJ).

[60] http://spraakbanken.gu.se

ation, taken from different genres: novels (22.7%), newspaper texts (70.1%), magazines (2.1%), as well as miscellaneous web texts (5.1%).

5.1.3 The Oslo Corpus

The Oslo Corpus is a corpus of Norwegian texts of approximately 18.5 million words. It consists of texts of three main genres: fiction (1.7 million words), newspapers/magazines (9.6 million words) and non-fictional prose (7.1 million words), and has been automatically annotated using the Oslo-Bergen tagger (Hagen, Johannessen and Nøklestad 2000), a morphosyntactic tagger assigning a Constraint Grammar (CG) analysis (Karlsson et al. 1995).[61]

Constraint Grammar is characterised by an eliminative approach and in the syntactic analysis, which follows a morphological disambiguation, the tagger starts out by administering all possible syntactic functions for each word. Unlikely candidates are then removed from each word, according to a set of CG rules. (63) gives the Constraint Grammar analysis of the sentence in (5.1.3), where syntactic tags are distinguished from morphology by the @-symbol:

(62) *Brevet* *med* *det* *pussige* *innholdet* *skrev* *jenta.*
letter-DEF with the strange content-DEF wrote girl-DEF
'The letter with the strange content, the girl wrote'

(63)
```
"<Brevet>"
        "brev" noun common sing def neuter @obj @subj
"<med>"
        "med" preposition @adv
"<det>"
        "det" determiner demonstrative sing neuter @det>
"<pussige>"
        "pussig" adjective sing def @adj>
"<innholdet>"
        "innhold" noun common sing def neuter @<p-utfyll
"<skrev>"
        "skrive" verb past tr1 i1 tr11 pa1 d1 pa5 pa3 @fv
"<jenta>"
        "jente" noun common sing def fem @obj @subj
```

As we can see, only the noun of a noun phrase receives the syntactic function tags `@subj` and/or `@obj`, whereas other modifying elements in the phrase will receive modifier-tags, relating them to the noun. Unlike the dependency analysis of Talbanken05, see section 5.1.1, the dependency relation is underspecified with respect to *which* element is head of a syntactic label or tag. It

[61] The Oslo Corpus is available for research purposes, see http://www.hf.uio.no/tekstlab

does not, for instance, make explicit that the subjects and object are dependents of the main (and only) verb. Notice also that the subject and object have not been disambiguated, both readings are still present in the output. The above example illustrates a *containment of ambiguity* which follows directly from the eliminative approach of Constraint Grammar. Since the rules have not been able to remove all but one analysis, both remain in the output.

5.2 Machine Learning

This section presents the machine learning algorithms and software which are employed in Part II of this thesis dealing with lexical acquisition of animacy information for common nouns in Scandinavian. Machine learning may be defined as follows (Mitchell 1997: 2):

Definition: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E

Properties of the training experience give rise to the distinction between *supervised* and *unsupervised* learning, where the former is characterized by direct evidence and the latter by indirect evidence. In supervised learning, the training experience consists of input-output pairs, whereas unsupervised learning involves learning without output values. The input to learning is commonly represented as a *feature vector*, a tuple of features with corresponding values $\langle f_1 = v_1, \ldots, f_n = v_n \rangle$ which defines an n-dimensional vector or feature space. We will primarily be concerned with supervised learning in the following and present two supervised machine learning systems, C5.0 presented in section 5.2.1, and TiMBL presented in section 5.2.2, based on decision-tree learning and memory-based learning, respectively. However, we will also employ the unsupervised technique of clustering as a method for data exploration in section 6.7.1 and we briefly present the clustering software Cluto in section 5.2.3.

Machine learning is based on inductive reasoning, hence improvement of the performance measure is usually defined through generalization to unseen instances. Most machine learning tasks may furthermore be reduced to learning a target function and therefore rely on an algorithm for locating the function that best fits the training data. The way in which the search for the best hypothesis is performed is part of the inductive bias of the machine learning algorithm. We experiment with two quite different machine learning algorithms instantiating the general distinction between *eager* and *lazy* learning algorithms. Eager learning algorithms generalize over the data prior to the application to unseen instances, whereas lazy algorithms postpone generaliza-

tion until the application to a new instance. The main difference between the two is thus found in the fact that lazy algorithms may consider the unseen instance when deciding how to generalize, whereas the eager algorithm may not (Mitchell 1997: 244f). The c4.5-algorithm employed for decision-tree learning is an eager algorithm, whereas the *k*-nearest neighbor algorithm employed in memory-based learning, is a lazy learning algorithm.

5.2.1 Decision trees (C5.0)

A decision tree is a classification model which relates a set of predefined classes with properties of the instances to be classified. Classification using a decision tree proceeds by means of a set of weighted, disjunctive tests which at each step, or node, in the decision tree assigns an appropriate test to an input, and which proceeds along one of its branches, representing possible outcomes of the test.

The software package employed for decision tree learning is C5.0 (Quinlan 1993).[62] Decision trees may be learned inductively by examining a set of training data and based on properties of these, constructing a classification tree. An initial tree is constructed from the training data by means of a *splitting criterion* and a *stopping criterion* (Manning and Schütze 1999). The splitting criterion grows the tree by dividing the training data into increasingly smaller subsets, whereas the stopping criterion tells the learner when to stop splitting. Following the c4.5 algorithm for decision-tree learning (Quinlan 1993), the splitting of a training set T into subsets $T_i, .., T_n$ in accordance with a test X with n outcomes is determined by a measure of *information gain*, i.e., the information gained by applying the test X to the training data T. The information gain of a particular test X is the difference between the amount of information needed to identify the class of a case in T on average and the information gained by partitioning the data in accordance with a particular test X.

$$gain(X) = info(T) - info_X(T)$$

The first term ($info(T)$) is obtained by summing over the information resulting from choosing each class $C_j, ..., C_k$, weighted by the frequency of the class in the training set T:[63]

$$info(T) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} \times \log_2 \left(\frac{freq(C_j, T)}{|T|} \right)$$

[62] The C5.0 software package may be downloaded from http://www.rulequest.com/.

[63] On a more general note $infoX) = H(X)$, the entropy for a single random variable.

The information measure ($info_X(T)$) for a certain test X which partitions the training data into n subsets, is obtained by summing over the information contained within each subset, as weighted by the frequency of the subset cases in the training set as a whole:

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i)$$

At each node in the decision tree, the test is chosen which maximises the information gain.[64] The splitting process is terminated when (i) all the subsets contain cases of the same class, or (ii) no further tests improve the results further.

The decision tree resulting from the initial splitting phase usually has the disadvantage of *overfitting* the data, i.e. it places too much significance on the observations made in the training data and may induce generalizations from mere coincidental properties of these. As a second stage in constructing a decision tree, a stage of *pruning* is vital to performance on unseen test cases. In the C5.0 system, the pruning of a decision tree is based on the predicted error rate of all the subtrees (Quinlan 1993).[65]

5.2.2 Memory-Based Learning (TiMBL)

Memory-Based Learning (Daelemans 1999; Daelemans and van den Bosch 2005) is a machine learning approach which is characterized by a notion of *analogy*, rather than abstraction. Training instances which constitute the training experience are simply stored in memory. At classification, some definition of similarity is employed in order to locate the instance(s) most similar to the new, unseen instance to be classified. Classification of the new instance is based directly on the knowledge of the previous assignment to these similar examples and learning is therefore supervised. Memory-based learning furthermore employs a lazy learning algorithm, the *k*-nearest neighbor algorithm, which postpones learning until classification time.

In our experiments we make use of the Tilburg Memory-Based Learner (TiMBL) (Daelemans et al. 2004).[66] TiMBL has a range of parameters which may be set to affect the learning process in various ways. The most important

[64] In fact, the measure of information gain has the disadvantage of favoring tests with many outcomes, with a worse-case scenario of one case per leaf node. Therefore the C5.0 system employs a refined measure of information gain - the *gain ratio* (Quinlan 1993: 23)

[65] The predicted error rate is calculated directly from the training data and thus requires no held-out data for pruning. This is a clear advantage when data are sparse.

[66] TiMBL is freely available at http://ilk.uvt.nl/software.html

of these parameters relate to either the selection of the nearest neighbors, i.e. which instances in memory should be allowed to affect the classification of a new instance, or the influence which each of these neighbors may exert over the final classification.

With regard to locating the nearest neighbors, various similarity metrics may be employed, which calculate the distance in vector space between the instance to be classified and various candidate neighbors. With regard to distance metrics, the default setting is the Overlap metric, where the distance is calculated as the sum of differing values of features. The *k* option allows the user to specify the region within which the *k*-nearest neighbors are found, where *k* is the number of distances considered (Daelemans et al. 2004). The parameter settings for feature weighting provide methods for assigning differing importance to individual features. One may also choose to give all features equal weight. The Information Gain setting takes the informativity of each feature into account when assigning weights, and the Gain Ratio (default) setting in addition normalizes for the number of values each feature may take on, based on the training data.

Finally in classification, the *k*-nearest neighbors determine the class of the unseen instance. The manner in which this decision is made can be influenced by the class voting weights. Either all neighbors have equal weight and the majority determines the class, so-called "majority voting", or the votes of closer neighbors are given more importance than more distant ones.

5.2.3 Clustering (Cluto)

Clustering is one of the primary methods of *unsupervised* machine learning where elements are grouped together based on their level of similarity. It is an unsupervised method since there is no use of manually annotated training data. Similarity is usually defined by distance in a high-dimensional vector space.

The clustering experiments presented in section 6.7 are performed using the Cluto clustering software (Karypis 2002), which is freely available.[67] Clustering algorithms are commonly classified as either bottom-up, so-called *agglomerative* clustering, or top-down, also known as *partitive* or divisive clustering. Cluto supports a range of different clustering algorithms of both types, as well as a range of parameters specifying a *criterion function*. The criterion function is defined over the instances of the clusters and provides a value expressing the level of similarity within the set of clusters, between the individual clusters or a combination of these. The main goal of clustering can then be reduced to locating the cluster solution which optimizes a certain criterion function.

[67] The Cluto software may be obtained from http://glaros.dtc.umn.edu/gkhome/views/cluto

A key problem in automatic clustering resides in locating the optimal number of clusters given a data set and no further information regarding the number of assumed categories. In Cluto this problem is bypassed by requiring the user to define the desired number of clusters in a clustering solution with the *k*-parameter. Section 6.7 provides more detail on the specific algorithm, criterion function and *k*-values employed, as well as discussing evaluation of the obtained cluster solutions.

5.3 Parsing

In Part III we present experiments in data-driven dependency parsing. We primarily employ the MaltParser system, however, a contrastive study is also performed with the MSTParser system. An introduction to data-driven dependency parsing is provided in section 8.1, as well as a more detailed introduction to the MaltParser system.

5.3.1 MaltParser

The freely available MaltParser[68] is a language-independent system for data-driven dependency parsing. MaltParser is based on a deterministic parsing strategy, first proposed by Nivre (2003) and extended to labeled dependency graphs in Nivre, Hall and Nilsson 2004, in combination with treebank-induced classifiers for predicting the next parsing action. See section 8.1.3 for more detail.

5.3.2 MSTParser

MSTParser is freely available[69] and is a language-independent system for data-driven dependency parsing. It searches for the maximum spanning tree over directed graphs, and employs large-margin discriminative training for the induction of scoring functions (McDonald, Crammer and Pereira 2005; McDonald et al. 2005). See section 9.3.1.2 for more detail and a comparison with MaltParser.

[68] http://w3.msi.vxu.se/users/nivre/research/MaltParser.html

[69] MSTparser is freely available from http://mstparser.sourceforge.net

Part II

Lexical Acquisition

6 ACQUIRING ANIMACY – EXPERIMENTAL EXPLORATION

Lexical acquisition deals with the automatic induction of lexical information. This area of computational linguistics has gained an increasingly important role as large linguistic corpora have become more easily available. The methods employed are per definition data-driven and rely on statistical inference over language data in some form. Whereas lexical information can refer to a wide variety of properties, most recent work has focused on the acquisition of lexical semantics. The basic approaches can be summarized as follows (Baldwin 2006):

Lexical similarity identify "near-matches" (synonyms, near-synonyms, associated word etc.) to the given lexical item, and "inherit" their semantic properties

Lexico-syntactic patterns identify the lexico-syntactic patterns associated with a given phenomenon, and look for corpus occurrences thereof

Resource mining mine pre-existing lexical resource(s) for relevant information

The view of lexical semantics which underlies the two first approaches to lexical acquisition is what is often referred to as the *distributional hypothesis*, stating that words which have similar distributions in language will also have similar meanings. The context of usage is thus defining for the meaning of lexical items and the context is usually represented as a high-dimensional vector. The way in which the context of usage (or the dimensions of the vector) is defined, varies, ranging from the use of simple word forms (Schütze 1998; Sahlgren 2006), lemmas, parts-of-speech to syntactic relations (Lin 1998). The second approach which relies on lexico-syntactic patterns in addition assumes that the syntactic distribution of lexical items constitutes a reliable predictor of semantics or meaning.

There are, generally speaking, two main methods for the evaluation of the acquired information in lexical acquisition – *internal* and *external* evaluation.

In an internal evaluation the acquired information is evaluated against a gold standard of some kind, e.g. a manually annotated corpus or a lexical resource (thesaurus, ontology etc.). An external evaluation evaluates the effect of the added information in terms of performance on a separate NLP task, such as parsing or semantic role labeling.

In chapter 3, we discussed the lexical semantic property of animacy and its relation to the realization and interpretation of syntactic arguments. We saw that animacy is a central factor in argument differentiation since arguments differ in their degree of correlation with the dimension of animacy, leading to observable frequency effects in a range of languages. Explanations for these correlations appeal to notions of accessibility, cognitive status and prominence, see section 3.6. We also distinguished animacy as a denotational property from a referential property, claiming that animacy is largely a denotational property of nouns.

This chapter aims at investigating the theoretical claims from chapter 3 further.[70] In particular, we argue that the task of *animacy classification*, a subtask of the general problem of lexical acquisition of animacy information, provides us with a methodology for evaluating the theoretical proposals. We employ data-driven methods which highlight the correlation between syntax and semantics and enable us to quantify the strength of this association. In particular, we investigate the extent to which the syntactic distribution of a noun is indicative of its animacy, and to what extent it is possible to generalize from syntactic behaviour to semantic animacy for unseen nouns. The assumption that animacy is a denotational property of nouns entails that the animacy of a noun is fairly stable across different contexts. We may consequently test whether animacy can be obtained as a class property at the level of lemmas or types, i.e. whether animacy may be acquired based on information regarding lemmas. Furthermore, the dimension of animacy and its delimitation and gradience may be tested by examining different types of nouns.

To be more precise, we will investigate the following issues:

Viability Can animacy be acquired through morphosyntactic distributional features for noun lemmas?

Features Can we approximate the linguistic generalizations regarding animacy as empirical features taken from a corpus? Which features may be employed to acquire animacy? Which of these features are most important?

Generalizability – robustess How is the classification affected by sparse data? How much data is needed to acquire stable generalizations?

[70] Parts of this chapter builds on work presented earlier in Øvrelid 2005, 2006.

Generalizability – machine learning algorithm Does the choice of machine learning algorithm affect the results? Are the results generalizable across machine learning algorithms with different properties? We examine the following distinctions, see section 5.2 for more detail:

1. eager versus lazy learning algorithms
2. supervised versus unsupervised learning

Class granularity Do the results say anything about the delimitation of the animacy dimension and its division into classes? Can we test the hypotheses regarding gradience in a principled manner?

6.1 Previous work

6.1.1 Animacy

Lexical acquisition of animacy information constitutes a relatively unexplored field of study in computational linguistics. It bears some resemblance to the task of named entity recognition (NER) (Tjong Kim Sang 2002b) which usually makes reference to a 'person' class, (see, e.g., Chinchor et al. 1999). However, whereas most NER systems make extensive use of orthographical, morphological or contextual clues (titles, suffixes) and gazetteers, animacy for nouns is not usually signaled overtly in the same way. The work which has been done on acquisition of animacy information falls into the third category of lexical acquisition tasks mentioned above, namely that of 'resource mining'. The lexical resource employed is that of the English WordNet (Fellbaum 1998). WordNet can be seen to represent the animacy distinction by having very general hypernyms (so-called 'unique beginners'), such as 'person' and 'artifact', and the idea is that the hyponyms of these general concepts inherit their animacy. However, direct extraction from WordNet is not completely trivial; the animacy of the unique beginners do not unequivocally distribute to their hyponyms and WordNet, as is well-known, contains extensive polysemy, i.e. words may belong to several senses. Orăsan and Evans (2001, 2007) present a study where animacy information is inherited from hypernyms (starting with unique beginners) in WordNet and where polysemy is distributed evenly across the various senses. A threshold for animacy is then determined empirically. An extended approach in addition makes use of a few contextual clues in the annotation of a corpus for animacy.[71] Orăsan and Evans (2007) show that the

[71] Orăsan and Evans (2007) employ machine learning in order to annotate the animacy of a given noun in a corpus. In addition to the dominant animacy of the noun, taken from WordNet,

acquired animacy information can be beneficial for anaphora resolution, hence evaluate the classification externally.

It is clear from the above that the methods employed in previous work on acquisition of animacy information are restricted to languages for which large scale lexical resources expressing this distinction, such as WordNet, are available.

6.1.2 Verb frames and classes

Whereas acquisition of information regarding nouns in general has not been given much attention in work on lexical acquisition, various syntactic and semantic properties of verbs have been widely studied.[72] In particular, the induction of various types of *verb frames* and *verb classes* has been the subject of a range of studies.

The subcategorization frame of a verb describes the types of arguments a verb takes and is usually assumed to be a gradient property in computational work, hence describes the differing propensities of verbs for different syntactic frames or sets of arguments. Lexical acquisition of subcategorization information (Manning 2003; Carroll and Rooth 1998) usually relies on a parsed corpus and differ with respect to the number of frames and the information (parts-of-speech, lexicalization) provided for the frames (Schulte im Walde 2007).

Acquisition of selectional restrictions (Resnik 1996; Erk 2007), see sections 3.5 and 9.2.8, can be seen as an extension of subcategorization frame induction, where the syntactic classes of arguments are specified for semantic class.[73] This information is usually taken from a lexical resource, typically WordNet, however, lexical similarity has also been employed in order to generalize frames to unknown instances (Erk 2007).

Finally, induction of verbal, semantic classes has been extensively studied from the perspective of lexical acquisition. The availability of a lexicon of verb classes (Levin 1993) has enabled a common platform for evaluation. It is

---

they make use of the "animacy of the verb" if the noun in question is subject and the proportion of animate/inanimate pronouns in the text. It is unclear, however, how the former feature is calculated – whether the information is taken from the WordNet resource in some way or from the gold standard corpus. The latter feature provides a measure of the frequency of animate entities on a whole for the text in question.

[72] See Schulte im Walde 2007 for a comprehensive overview of work on acquisition of verbal frames and classes.

[73] Selectional restrictions as an extension of subcategorization is an oversimplification, due to the fact that these provide information from different linguistic levels. The latter is strictly syntactic, whereas the former is semantic. It is therefore in principle possible to separate the two distinctions, i.e. the selectional restrictions need not necessarily entail syntactic subcategorization. More on this in section 9.2.8 below.

an underlying assumption in the work on acquisition of verb classes that the syntactic distribution of a verb, in particular with respect to so-called *alternations*, is largely determined by its semantic class.[74] This assumption makes it possible to acquire semantic classes largely based on syntactic, distributional features. Merlo and Stevenson (2001) study the acquisition of verb class information as a classification problem, focusing on three classes of optionally intransitive verbs in English - unergative verbs, e.g., *race*, unaccusative verbs, e.g., *melt* and object-drop verbs, e.g., *play*. They make use of a small set of linguistically motivated features in a 3-way classification task and show a considerable improvement over a random baseline (69.8% accuracy with a baseline of 33.9%). Joanis and Stevenson (2003) extend this work to a larger feature set with comparable performance. The verb classification in Levin 1993, however, only provides one out of many possible ways of classifying verbs and it is also only available for English. Unsupervised approaches to acquisition of verb classes partially address this problem by clustering verbs without the necessity of a resource like Levin for English (Stevenson and Joanis 2003) and for languages where such resources are not available, e.g. German (Schulte im Walde 2006).[75]

The close relation between syntax and semantics which is highlighted in particular in the work on verb classes bears a strong resemblance to our problem of animacy classification, as defined above, and we will use it as inspiration in the following.

6.2 Data preliminaries

In the following we formulate the task of acquiring animacy information as a classification problem where the learning task consists of classifying nouns as being either animate or inanimate. The goal of learning is thus to find the approximated target function, V, which best performs this task. That is, given a set of noun lemmas, we want to locate the function which gives as output the greatest number of correctly assigned classification values. This is a discrete-valued function from noun instances to a class from a predefined set of classes, $V : NounLemma \rightarrow c \in \{Anim, Inan\}$.

In order to train a classifier to distinguish between animate and inanimate

[74] A syntactic alternation is a variation in terms of the syntactic realization of arguments with respect to a particular verb. Examples are the dative alternation, as in examples (10)–(11) in section 3.2.1 and spray/load alternations, e.g. *spray the wall with paint* vs *spray paint onto the wall* . See Levin 1993 for an extensive overview.

[75] Note, however, that a gold standard is still assumed for evaluation purposes, although it is not employed for training.

nouns, we have to decide on the appropriate training experience for this task. In particular, we must select a set of representative instances and decide on a relevant representation of these instances. Section 6.2.1 discusses the choice of language and corpus resource for the present classification study, and 6.2.2 will detail the selection of nouns for classification.

In section 3.2 above we discussed the role of animacy in argument differentiation as a cross-linguistic tendency which exhibits clear frequency effects in language. We also discussed linguistic dimensions closely related to animacy and argument differentiation, such as agentivity, individuation and accessibility, see section 3.6.2. In section 6.2.3 we formulate a set of theoretically motivated features which exploit the close relation between animacy and distinctions in syntactic argumenthood, as well as related notions. It is important to note that these features only provide practical *approximations* of more theoretical notions of argumenthood, agentivity and accessibility. These theoretical notions are approximated by emprical features which may be extracted from an automatically annotated corpus.

6.2.1 Language and corpus resource

All experiments in this chapter are performed on Norwegian. In chapter 7, we will investigate the application of the methods developed in this chapter to another Scandinavian language, namely Swedish.

Since we wish to employ the morphosyntactic distribution of a noun as an indicator of its animacy, we need a corpus with morphological and syntactic annotation. Also, the corpus should be as large as possible, since we rely on inductive inference from frequencies in language use for classification. For the extraction of morphosyntactic distributional information, we choose to employ the Oslo Corpus, a corpus of Norwegian texts of approximately 18.5 million words. The corpus is morphosyntactically annotated and assigns an underspecified dependency-style analysis to each sentence, see section 5.1.3 for more detail.[76] The containment of ambiguity which is a property of Constraint Grammar can be seen to be an advantage in the approximation of features, since it enables the exclusion of instances which were deemed ambiguous by the grammar.

[76] All examples in the current chapter are taken from the Oslo Corpus, unless otherwise stated.

Animate	Inanimate
barn 'child', *direktør* 'director', *far* 'father', *flyktning* 'refugee', *forfatter* 'author', *gutt* 'boy', *jente* 'girl', *kvinne* 'woman', *leder* 'leader', *lege* 'doctor', *lærer* 'teacher', *mann* 'man', *medlem* 'member', *mor* 'mother', *person* 'person', *president* 'president', *sjef* 'boss', *soldat* 'soldier', *trener* 'coach', *venn* 'friend'	*aksje* 'stock', *artikkel* 'article', *bil* 'car', *bok* 'book', *brev* 'letter', *dag* 'day', *eiendom* 'property', *fly* 'airplane', *hus* 'house', *informasjon* 'information', *natt* 'night', *oppgave* 'task', *opplysning* '(piece of) information', *penge* 'coin/money', *pris* 'price', *produkt* 'product', *spørsmål* 'question', *svar* 'answer', *ting* 'thing', *vare* 'merchandise'

Table 6.1: Highly frequent ($>$ 1000) animate and inanimate nouns; Norwegian

6.2.2 Noun selection

As training data for the classifier, a set of 40 nouns were manually selected – 20 animate and 20 inanimate nouns, see table 6.1. The nouns were chosen based on two criteria: i) they are all Norwegian translations of nouns taken from the English WordNet, all of which are hyponyms of concepts distinguishing animacy, and ii) they are all highly frequent in the corpus.[77]

The animate nouns that were chosen were all hyponyms of the *person*-relation, which is itself a hyponym of *animate thing/ living thing*. A corpus study of Norwegian simple transitives in a sample of the Oslo Corpus showed that nouns expressing reference to animals, i.e. animate beings aside from humans, are very infrequent in the corpus (Øvrelid 2004), and these types of nouns will therefore not be included in the following.[78] There is no single category in WordNet that expresses the property of inanimateness, so the inanimate nouns were taken from two main hypernyms which ensure a spread in terms of the abstractness of the noun, namely the concepts *artifact*, e.g. *bil* 'car', *bok* 'book' and *abstraction*, e.g. *pris* 'price', *informasjon* 'information'.

The choice of highly frequent nouns was made in order to ensure a sufficient amount of data to test our various features on. The threshold for these nouns

[77] There is no Norwegian WordNet resource, but there is a small semantic lexicon, SIMPLE, available which contains semantic information for approximately 10,000 nouns structured according to the notion of *qualia*-roles (Pustejovsky 1991). Unfortunately however, the choice of nouns in this lexicon is rather limited. For instance, there are entries for *bildekk* 'car tyre' and *bilradio* 'car radio', but no entry for *bil* 'car'! Due to our initial frequency threshold, the information available in the SIMPLE lexicon was deemed too specialized.

[78] Øvrelid (2004) found that only 0.6% (5/889) of the common nouns in the sample refer to animals.

was set at > 1000 occurrences. In section 6.4 below, we examine the effect of alternative threshold assignments on animacy classification.

6.2.3 Features of animacy

The nouns listed in table 6.1 above are represented by a set of features which express properties of their morphosyntactic distribution. For each noun *w*, relative frequencies of various morphosyntactic features f_i are calculated from the corpus:

$$\frac{freq(f_i, w)}{freq(w)}$$

The features chosen to represent the nouns are presented below. Note that the features are all assumed to be noisy, as they are based on automatic annotation and simple regular expressions for extraction. We will test empirically in the experimental section whether these features capture the relevant distinctions with respect to animacy despite the noise caused by input data and feature approximation.

Subject and object

Subjects and objects tend to differ with respect to animacy and this tendency has been observed as frequency effects in a range of languages. We have, in particular, discussed the effect of relative animacy in transitive constructions. The proportion of subject and object occurrences for each noun is therefore recorded. For transitive subjects (SUBJ), we extract the number of instances where the noun in question is unambiguously tagged as subject and followed by a finite verb and an unambiguously tagged object.[79] The frequency of direct objects (OBJ) for a given noun was approximated to the number of instances where the noun in question was unambiguously tagged as object. We here assume that an unambiguously tagged object implies an unambiguously tagged subject. However, by not explicitly demanding that the object is preceded by a subject, we also capture objects with a "missing" subject, such as relative clauses and infinitival clauses.

[79] The tagger works in an eliminative fashion, so tokens may bear two or more tags when they have not been fully disambiguated.

Genitive

Genitive marking typically signals a semantic relation of possession, a relation which has been shown to favour animate possessors (Rosenbach 2002; Dahl and Fraurud 1996). However, this requirement is certainly not an absolute constraint on the construction; semantic relationships figuring inanimate entities such as a part-whole relations, e.g. *bilens hjul* 'the car's wheel' also occur commonly.[80] The feature extraction for the genitive feature (GEN) counts the number of times each noun occurs with genitive case marking, i.e. the suffix *-s*.

Passive

Agentivity is also related to animacy, see section 3.5. Animate entities are inherently sentient, capable of acting volitionally and causing an event to take place - all properties of the prototypical agent (Dowty 1991). The passive construction, or rather the property of being expressed as the demoted agent in a passive construction, is a possible approximator of agentivity. Transitive constructions tend to passivize better (hence more frequently) if the demoted subject bears a prominent thematic role, preferably agent. Norwegian has two ways of expressing the passive, a morphological passive (verb + *s*) and a periphrastic passive (*bli* + past participle). The counts for the passive feature (PASS) include both types of passives preceding the *by*-phrase containing the noun lemma in question.

Anaphoric reference

In section 3.6, we discussed the idea that animate entities tend to be more *individuated* and more cognitively *accessible*. An entity which is highly individuated and accessible is also more likely to be referred to again later on in discourse. Anaphoric reference is a phenomenon where the animacy of a referent is clearly expressed. The personal pronouns distinguish their antecedents along the animacy dimension - animate *han/hun* 'he/she' vs. inanimate *den/det* 'it-MASC/NEUT'. This is one reason why information regarding the animacy

[80] An alternative construction to the s-genitive in Norwegian is constructed by inserting the possessive pronoun *sin* between the possessor and the possessed, as in *mannen sin bil* 'the man's car'. The *sin*-genitive is to be preferred when the relation is one of possession (Faarlund, Lie and Vannebo 1997), hence often involving an animate possessor. However, data on this construction was far too sparse and yielded zero occurrences for a large number of the nouns (both animate and inanimate), and was hence abandoned.

of a noun can be helpful in the task of coreference resolution (Orăsan and Evans 2007). Coreference resolution is a complex problem, and certainly not one that we shall attempt to solve in the present context. However, we might attempt to come up with a metric that approximates the coreference relation in a manner adequate for our purposes. Hale and Charniak (1998) describe a method for extracting gender statistics for English nouns by making use of coreference approximations. Their most simple method is the "last noun seen" method, where an anaphoric link is established between the last noun of one sentence and an initial pronoun in the next. This method is reported to account for approximately 43% of all anaphoric coreferences in a hand-tagged subset of the Wall Street Journal corpus. They also make use of the Hobbs algorithm (Hobbs 1976), which relies on a phrase-structure parse of the sentence in question as well as the preceding text, and exploits syntactic cues for coreference. This strategy alone is reported to achieve an accuracy of 65.3% on the same WSJ subset.

In our attempt to approximate coreference relations between a common noun and a subsequent personal pronoun, we make use of the fact that a personal pronoun usually refers to a discourse salient element which is fairly recent in the discourse. Now, if a sentence only contains one core argument (i.e. an intransitive subject) and it is followed by a sentence initiated by a personal pronoun, it seems reasonable to assume that that these are likely to be coreferent (Hale and Charniak 1998). (64) below shows an authentic example from the results for the noun *mann* 'man' taken from the Oslo Corpus:

(64) **Mannen**$_i$ ble pågrepet etter tre kvarters dramatisk biljakt. **Han**$_i$ var beruset og satt med den ladde haglen over knærne.
The man$_i$ was apprehended after a three-quarter long car chase. He$_i$ was intoxicated and sat with the loaded shot gun across his knees.

For each of the nouns in table 6.1, we count the number of times it occurs as a subject with no subsequent object and an immediately following sentence initiated by (i) an animate personal pronoun (ANAAN) –*han* 'he', *hun* 'she' or *de* 'they', and (ii) an inanimate personal pronoun (ANAIN) – *den* 'it-MASC' or *det* 'it-NEUT'. The 3rd person plural pronoun *de* 'they' is not a clear indicator of animacy since it may refer to both animate and inanimate referents, as in English. Merlo and Stevenson (2001) show that, in English, this plural pronoun exhibits a preference for animate reference and in a selection of 100 occurrences of this pronoun, they found that 76% of these had an animate antecedent.[81] We therefore make the same assumption for Norwegian. Although

[81] Merlo and Stevenson (2001) make use of personal pronouns as indicators of argument structure for a verb. If the verb often occurs with an animate pronominal subject, they assume that it assigns an agentive role to its subject.

	SUBJ	OBJ	GEN	PASS	ANAAN	ANAIN	REFL
forfatter	0.1734	0.0809	0.0639	0.0020	0.0109	0.0034	0.0075
artikkel	0.0799	0.1091	0.0032	0.0032	0.0013	0.0032	0.0006

Figure 4: Example feature vectors.

this is a possible source for mistakes in the counts, we assume that the general distribution of instances will still make the relevant distinction with regards to animacy.

For the inanimate pronouns, the neuter form *det* 'it-NEUT' is problematic as this is also the expletive subject form, hence this pronoun often initiates a sentence, but has a clearly non-referential function. However, as there is no obvious way of automatically distinguishing between the pronominal and expletive use, we count all occurrences of this pronoun when it initiates a following sentence. Another possibility would have been to exclude all occurrences of *det* 'it-NEUT' from the counts, with the consequence that this test would be inapplicable for the set of neuter nouns in our training set (8 nouns).

Reflexive

Reflexive pronouns represent another form of anaphoric reference which, contrary to the personal pronouns, locate their antecedent locally, i.e. within the same clause. The third person reflexive pronoun *seg* 'him/her/itself' does not, however, position its antecedent along the animacy dimension. In the reflexive construction the subject and the reflexive object are, typically, coreferent and it describes an action directed at oneself. Although the reflexive pronoun in Norwegian does not distinguish for animacy, the agentive semantics of the construction might favour an animate subject.

The feature of reflexive coreference (REFL) is more straightforward to approximate, as this coreference takes place within the same clause. For each noun, the number of occurrences as a subject followed by a verb and the 3rd person reflexive pronoun *seg* 'him-/her-/itself' are counted.

6.2.3.1 Data overview

For classification, each noun is represented as a feature vector of distributional features and is labeled with its *class* – animate or inanimate. Figure 4 shows the individual feature vectors representing the animate noun *forfatter* 'writer' and the inanimate noun *artikkel* 'article'.

	Animate		Inanimate		
	Mean	SD	Mean	SD	#
SUBJ	0.14	0.05	0.07	0.03	16813
OBJ	0.11	0.03	0.23	0.10	24128
GEN	0.04	0.02	0.02	0.03	7830
PASS	0.006	0.005	0.002	0.002	577
ANAAN	0.009	0.006	0.003	0.002	989
ANAIN	0.003	0.003	0.006	0.003	944
REFL	0.005	0.0008	0.001	0.0008	558

Table 6.2: Mean relative frequencies and standard deviation for each class (20 animate, 20 inanimate nouns) and feature, as well as total data points for each feature (#).

The mean relative frequencies with standard deviations for each class – animate and inanimate – and feature are presented in table 6.2. The total data points for each feature following the data collection are also presented in the last column of table 6.2. As we can see, quite a few of the features express morphosyntactic cues that are rather infrequent. This is in particular true for the passive feature (PASS) and the anaphoric features ANAAN, ANAIN and REFL. When examining the features in table 6.2, however, these features still express the relevant distinctions, and all differences between the means of the two groups are significant.[82]

Another point is that the values for the features that one would expect to be quite frequent, e.g. SUBJ and OBJ only range from about 3% to 14% of all occurrences. The reason for this is that the regular expressions designed to extract the counts require the subjects and objects in question to be *unambiguously tagged*. This means that the transitive subjects and objects that *are* counted are only those that occur in a syntactic environment which clearly disambiguates them functionally.[83]

[82] Statistical significance was calculated with an unpaired *t*-test. We compared the mean of means between the group of animate and inanimate nouns, and found that all differences were significant – SUBJ,OBJ,REFL:$p<.0001$; ANAAN:$p<.0005$; PASS:$p<.005$; ANAIN:$p<.01$; GEN:$p<.05$.

[83] In practice this includes transitive complex VPs (due to the V2-property of Norwegian), i.e. VPs containing auxiliary or modal verbs, sentences where something other than the subject or object occupies sentence initial position, or subjects or objects appearing in subordinate clauses of different types, see section 4.2

6.2.3.2 Other features

In addition to the features presented in table 6.2, several other features were extracted, which did not exhibit a significant distinction. This was partly due to errors in the automatic analysis. For instance, indirect objects in ditransitive constructions, turned out to yield a result that was contrary to the expected results. The mean result for the animate class was 0.007%, whereas the inanimate class had the higher count of 0.008%. However, a quick look at some of the extracted sentences shows that the tagger's analysis of indirect objects is inaccurate. Other features that proved not to differ significantly for the two classes include morphological definiteness and the 'last noun seen' anaphoric reference approximation (Hale and Charniak 1998).

6.3 Method viability

We start out by testing the viability of the method as such, i.e. whether unseen nouns may be classified for animacy based on a small set of linguistically motivated distributional features. We also test the effect of the various features individually and in combination.

6.3.1 Experimental methodology

The experimental methodology chosen for the classification experiments is similar to the one described in Merlo and Stevenson 2001 for verb classification. We employ decision tree learning for construction of classifiers, see section 5.2.1 and leave-one-out training and testing of the classifiers. In leave-one-out cross validation, each noun is used as test data exactly once, whereby the $n-1$ other instances are used for training the classifier. This is a good option when the set of training data is small, as in the present context. In addition, all our classifiers employ the *boosting* option for constructing classifiers (Quinlan 1993).[84] For calculation of the statistical significance of differences in the performance of classifiers tested on the same data set, McNemar's test (Dietterich 1998) is employed. Note however, that due to the small data set, the test provides a very strict criterion by which to determine difference. We therefore report results even though they are not statistically significant, but remark on significance explicitly wherever relevant. The baseline we employ is a random baseline of 50% accuracy.

[84] In boosting, several classifiers are constructed during training and applied to each test instance, whereby the classification is determined by majority voting.

Feature	Accuracy (%)
SUBJ	85.0
REFL	82.5
OBJ	72.5
GEN	72.5
ANAAN	67.5
PASS	62.5
ANAIN	50.0

Table 6.3: Accuracy for classifiers trained with individual features.

Used	Not Used	Accuracy (%)
1. SUBJ OBJ GEN PASS ANAAN ANAIN REFL		87.5
2. OBJ GEN PASS ANAAN ANAIN REFL	SUBJ	85.0
3. SUBJ GEN PASS ANAAN ANAIN REFL	OBJ	87.5
4. SUBJ OBJ PASS ANAAN ANAIN REFL	GEN	85.0
5. SUBJ OBJ GEN ANAAN ANAIN REFL	PASS	82.5
6. SUBJ OBJ GEN PASS ANAIN REFL	ANAAN	82.5
7. SUBJ OBJ GEN PASS ANAAN REFL	ANAIN	87.5
8. SUBJ OBJ GEN PASS ANAAN ANAIN	REFL	75.0

Table 6.4: Accuracy for classifiers trained with all features and 'all minus one'.

6.3.2 Experiment 1

Table 6.3 shows the performance of each individual feature in the classification of animacy. As we can see, the performance of the features differ quite a bit, ranging from mere baseline performance (ANAIN) to a 70% error reduction compared to the baseline (SUBJ). The first line of table 6.4 shows the performance using all the seven features collectively where we achieve an accuracy of 87.5%, an error reduction of 75%. The SUBJ, REFL, OBJ and GEN features employed individually are the best performing individual features and their classification performance do not differ significantly from the performance of the combined classifier, whereas the rest of the individual features do ($p<.05$).

The subsequent lines (2-8) of table 6.4 show the accuracy results for classification using all features except one at a time. This provides an indication of the contribution of each feature to the classification task. In general, the removal of a feature causes a 0%-12.5% deterioration of results, however, only the difference in performance caused by the removal of the REFL feature is significant ($p<.05$). Since this feature is one of the best performing features

individually, it is not surprising that its removal causes a notable difference in performance. The removal of the ANAIN feature, on the other hand, does not have any effect on accuracy whatsoever. This feature was the poorest performing feature with a baseline, or mere chance, performance.

6.3.2.1 Discussion

The above experiments have shown that the classification of animacy for common nouns is achievable using morphosyntactic distributional data from a corpus. The results of the experiments are encouraging, and due to the fact that the features are linguistically motivated, hopefully also generalisable to a larger set of nouns. However, several questions remain unanswered following these initial experiments.

We have chosen to classify along a binary dimension (animate vs. inanimate) with a small set of nouns. Two related objections may be put forward at this point. Firstly, it might be argued that a binary dimension such as this is artificial and that there should be a finer subdivision of nouns. Zaenen et al. (2004) describe an encoding scheme for the manual encoding of animacy information in part of the English Switchboard corpus. They make a three-way distinction between human, other animates, and inanimates, and also provide further subdivisions of these. The 'other animates' category describe a rather heterogeneous group of entities: organizations, animals, intelligent machines and vehicles. What these have in common is that they may all be construed linguistically as animate beings, even though they, in the real world, are not. Interestingly, the two misclassified inanimate nouns in our experiments were *bil* 'car' and *fly* 'airplane', both vehicles.[85] They exhibited a more agentive pattern which showed up in the transitive subject feature, the passive feature and the reflexive feature, in particular. However, they did not pattern completely with the animate nouns, they had a high object count and behaved like the inanimate nouns when it came to anaphoric pronouns. Secondly and related to the above, the choice of nouns in the experiment might be considered too limited. Had we chosen to include, for instance, nouns that have a metonymic use, e.g., organizations, the classification into only two classes might have been less successful. However, we chose to start out with a binary classification in order to test the viability of the method and its suitability for the classification task.

The features represent linguistic dimensions which have been claimed to correlate with animacy, such as syntactic functions and thematic roles. One

[85] Inanimate subjects have been claimed to be ungrammatical in Japanese. However, sentence production experiments have employed examples like *A taxi picked up a traveler*, which were deemed acceptable (Branigan, Pickering and Tanaka 2008).

might ask whether the chosen features represent sufficient information to base classification on. One of the misclassified animate nouns was *venn* 'friend', a clearly animate noun. However, according to our seven chosen features, this noun largely patterns with the inanimate nouns. When considering it, this probably also makes sense, as we are basing our classification of a real world property only on our linguistic depiction of it. A friend is probably more like a physical object in the sense that it is someone one likes/hates/loves or otherwise reacts *to*, rather than being an agent that acts upon its surroundings. This is reflected in a low proportion of subject occurrences (.076), as well as reflexive reference (.0012) and a high proportion of direct object occurrences (0.19), see table 6.2.

In conclusion then, we have seen that the method yields promising results for classification of animacy when applied to Norwegian common nouns using a set of seven linguistically motivated features of animacy. The features, which capture syntactic distributional properties of the nouns where animacy has been shown to cause frequency effects, proved important in classification.

6.4 Robustness

The classification experiments reported above impose a frequency constraint (absolute frequencies $>$1000) on the nouns used for training and testing in order to study the interaction of the different features without the effects of sparse data. In the light of the results from these experiments, however, it might be interesting to further test the performance of our features in classification as the frequency constraint is gradually relaxed. To this end, three sets of common nouns each counting 40 nouns (20 animate and 20 inanimate nouns) were randomly selected from groups of nouns with approximately the same frequency in the corpus. The first set included nouns with an absolute frequency of 100 ± 20 ($\sim$100), the second of 50 ± 5 ($\sim$50) and the third of 10 ± 2 ($\sim$10). Feature extraction followed the same procedure as in experiment 1, relative frequencies for all seven features were computed and assembled into feature vectors, one for each noun.

6.4.1 Experiment 2: Effect of sparse data on classification

In order to establish how much of the generalizing power of the classifier is lost when the frequency threshold for the extraction of nouns is lowered, an experiment was conducted which tested the performance of the earlier classifier, i.e. the classifier trained on the more frequent nouns, as applied to the

Freq	All	SUBJ	OBJ	GEN	PASS	ANAAN	ANAIN	REFL
>1000	87.5	85.0	72.5	72.5	62.5	67.5	50.0	82.5
$\sim$100	70.0	75.0	80.0	72.5	65.0	52.5	50.0	60.0
$\sim$50	57.5	75.0	62.5	77.5	62.5	57.5	50.0	55.0
$\sim$10	52.5	52.5	65.0	50.0	57.5	50.0	50.0	50.0

Table 6.5: Accuracy obtained when applying the high-frequency classifiers trained with all and individual features to the lower-frequency nouns ($\sim$100,$\sim$50,$\sim$10).

three groups of less frequent nouns. As we can see from the first column in table 6.5, we observe a clear deterioration of results, from our earlier accuracy of 87.5% to new accuracies ranging from 70% to 52.5%, barely above the baseline. Not surprisingly, the results decline steadily as the absolute frequency of the classified noun is lowered.

Accuracy results provide an indication that the classification is problematic. However, it does not indicate what the damage is to each class as such. A confusion matrix is in this respect more informative. Confusion matrices for the classification of the three groups of nouns, $\sim$100, $\sim$50 and $\sim$10, are provided in table 6.6. These clearly indicate that it is the animate class which suffers when data becomes more sparse. The percentage of misclassified animate nouns increases drastically from 50% at $\sim$100 to 80% at $\sim$50 and finally 95% at $\sim$10. The classification of the inanimate class remains pretty stable throughout. The fact that a majority of our features (SUBJ, GEN, PASS, ANAAN and REFL) target animacy, in the sense that a higher proportion of animate than inanimate nouns exhibit the feature, gives a possible explanation for this. As data gets more limited, this distinction becomes harder to make, and the animate feature profiles come to increasingly resemble the inanimate. Because the inanimate nouns are expected to have low proportions (compared to the animate) for all these features, the data sparseness is not as damaging.

In order to examine the effect of the lowering of the frequency threshold on each individual feature , we also ran classifiers trained on the high frequency nouns with only individual features on the three groups of new nouns. These results are depicted in columns 3-9 in table 6.5. As the frequency threshold is lowered, the performance of the classifiers employing all features and those trained only on individual features become more similar. For the $\sim$100 nouns, only the two anaphoric features ANAAN and the reflexive feature REFL, have a performance that differs significantly ($p<.05$) from the classifier employing all features. For the $\sim$50 and $\sim$10 nouns, there are no significant differences between the classifiers employing individual features only and the classifiers

∼100 nouns		
(a)	(b)	← classified as
10	10	(a) class animate
2	18	(b) class inanimate

∼50 nouns		
(a)	(b)	← classified as
4	16	(a) class animate
1	19	(b) class inanimate

∼10 nouns		
(a)	(b)	← classified as
1	19	(a) class animate
	20	(b) class inanimate

Table 6.6: Confusion matrices for classification of lower frequency nouns with the high-frequency classifier.

trained on the feature set as a whole. This indicates that the combined classifiers no longer exhibit properties that are not predictable from the individual features alone and they do not generalize over the data based on the combinations of features.

In terms of accuracy, a few of the individual features even outperform the collective result. On average, the three most frequent features, the SUBJ, OBJ and GEN features, cause a 9.5% and 24.6% reduction of the error rate for the ∼100 and ∼50 nouns, respectively. For the lowest frequency nouns (∼10) we see that the OBJ feature alone reduces the errors by almost 24%, from 52.5% to 65 % accuracy. In fact, the OBJ feature seems to be the most stable feature of all the features. When examining the means of the results extracted for the different features, the OBJ feature is the feature which maintains the largest difference between the two classes as the frequency threshold is lowered. The second most stable feature in this respect is the SUBJ feature. Figure 5 clearly illustrates the effect of sparse data on classification accuracy. The group of experiments reported above shows that the lowering of the frequency threshold for the classified nouns causes a clear deterioration of results in general, and most gravely when all the features are employed together.

6.4.2 Experiment 3: Back-off features

The three most frequent features, the SUBJ, OBJ and GEN features, were the most stable in the two experiments reported above and had a performance which did not differ significantly from the combined classifiers throughout. In light of this we ran some experiments where all combinations of these more frequent features were employed. The results for each of the three groups of nouns is presented in table 6.7. The exclusion of the less frequent features has

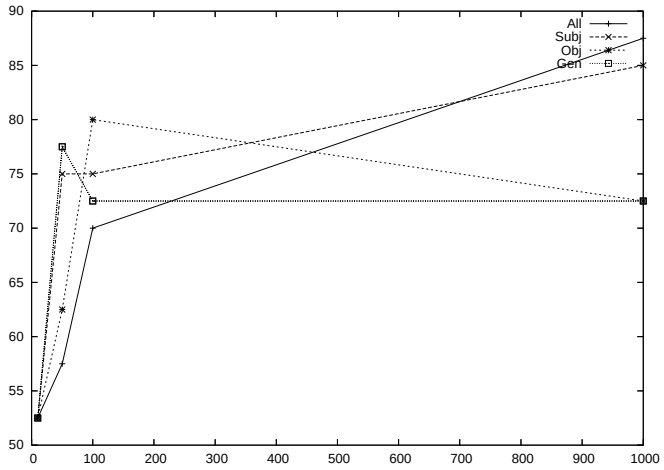

Figure 5: Accuracy as a function of absolute noun frequencies for classifiers employing all features, as well as the individual SUBJ, OBJ and GEN classifiers.

Freq	SUBJ&OBJ&GEN	SUBJ&OBJ	SUBJ&GEN	OBJ&GEN
∼100	87.5	87.5	77.5	85.0
∼50	82.5	90.0	70.0	77.5
∼10	57.5	50.0	50.0	47.5

Table 6.7: Accuracy obtained when applying classifiers trained with combinations of the most frequent features to the lower-frequency nouns.

a clear positive effect on the accuracy results. For the ∼100 and ∼50 nouns, the performance has improved compared to the classifier trained with the full set of features, as well as the classifiers trained with individual features. The classification performance for these nouns is now identical or only slightly worse than the performance for the high-frequency nouns in experiment 1. For the ∼10 group of nouns, the performance is, at best, the same as for all the features and at worse fluctuating around baseline.

In general, the best performing feature combinations are SUBJ&OBJ&GEN and SUBJ&OBJ. These two differ significantly ($p<.05$) from the results obtained by employing all the features collectively for both the ∼100 and the ∼50 nouns, hence indicate a clear improvement. The feature combinations both contain the two most stable features – one feature which targets the animate class (SUBJ) and another which target the inanimate class (OBJ) – a property which facilitates distinction even as the general differences between the

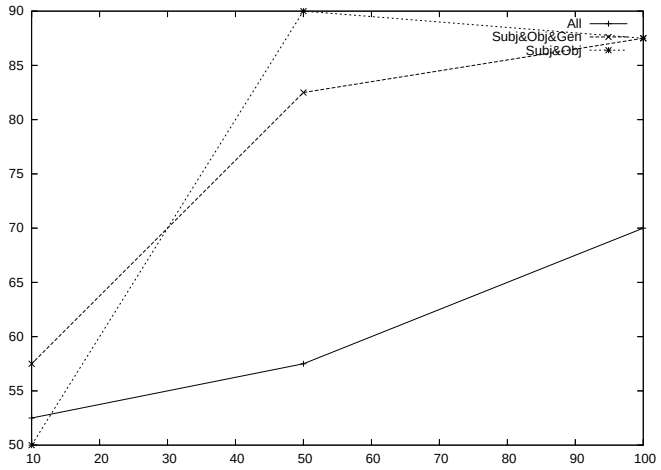

Figure 6: Accuracy as a function of absolute noun frequencies for classifiers employing all features, as well as the backed off SUBJ&OBJ&GEN and SUBJ&OBJ classifiers.

two decrease. Figure 6 illustrates the clear improvement of feature back-off compared to the full set of features.

It seems, then, that backing off to the most frequent features might constitute a partial remedy for the problems induced by data sparseness in the classification. The feature combinations SUBJ&OBJ&GEN and SUBJ&OBJ both significantly improve the classification performance and enable us to maintain the same accuracy for the $\sim$100 and $\sim$50 nouns as for the higher frequency nouns, reported in experiment 1.

6.4.3 Experiment 4: Back-off classifiers

Another option, besides a back-off to more frequent features in classification, is to back off to another classifier, i.e. a classifier trained on nouns with a similar frequency. An approach of this kind attempts to exploit any group similarities that these nouns may have in contrast to the mores frequent ones.

In this set of experiments, classifiers were trained and tested using leave-one-out cross-validation on the three groups of lower frequency nouns and employing individual, as well as various other, feature combinations. The results for all features as well as individual features are summarized in table 6.8.

As we can see, the result for the classifier employing all the features has improved somewhat compared to the corresponding classifiers in experiment 3

Freq	All	SUBJ	OBJ	GEN	PASS	ANAAN	ANAIN	REFL
>1000	87.5	85.0	72.5	72.5	62.5	67.5	50.0	82.5
$\sim$100	85.0	52.5	87.5	65.0	70.0	50.0	57.5	50.0
$\sim$50	77.5	77.5	75.0	75.0	50.0	50.0	50.0	50.0
$\sim$10	52.5	50.0	62.5	50.0	50.0	50.0	50.0	50.0

Table 6.8: Accuracy obtained when applying lower-frequency classifiers trained with all and individual features to new lower-frequency nouns. Performance with the high-frequency classifier (>1000) is provided for comparison.

Freq	SUBJ&OBJ&GEN	SUBJ&OBJ	SUBJ&GEN	OBJ&GEN
$\sim$100	85.0	85.0	67.5	82.5
$\sim$50	75.0	80.0	75.0	70.0
$\sim$10	62.5	62.5	50.0	62.5

Table 6.9: Accuracy obtained when applying lower-frequency classifiers trained with combinations of the most frequent features to new lower-frequency nouns.

(as reported above in table 6.5) for all our three groups of nouns. This indicates that there is a certain group similarity for nouns of similar frequency that is captured in the combination of the seven features. However, backing off to a classifier trained on nouns that are more similar frequency-wise does not cause an improvement in classification accuracy. Apart from the SUBJ feature for the $\sim$100 nouns, none of the other classifiers trained on individual or all features for the three different groups differ significantly ($p<.05$) from their counterparts in experiment 3.

As before, combinations of the most frequent features were employed in the new classifiers trained and tested on each of the three frequency-sorted groups of nouns. In the terminology employed above, this amounts to a backing off both classifier- and feature-wise. The accuracy measures obtained for these experiments are summarized in table 6.9. For these classifiers, the backed off feature combinations do not differ significantly from their counterparts in experiment 3, where the classifiers were trained on the more frequent nouns with feature back-off.

6.4.4 Summary

Experiments 1–4 have shown that the classification of animacy for Norwegian common nouns is achievable using distributional data from a morphosyntactically annotated corpus. The chosen morphosyntactic features of animacy have proven to distinguish well between the two classes. As we have seen, the transitive subject, direct object and morphological genitive provide stable features for animacy even when the data is sparse(r). Four groups of experiments have been reported which indicate that a reasonable remedy for sparse data in animacy classification consists of backing off to a smaller feature set in classification. These experiments indicate that a classifier trained on a small set of highly frequent nouns (experiment 1) backed off to the most frequent features (experiment 3) sufficiently capture generalizations which pertain to nouns with absolute frequencies down to approximately fifty occurrences and enables an unchanged performance approaching 90% accuracy.

6.5 Machine learning algorithm

Decision-tree learning represents an *eager* machine learning algorithm; generalization over the training data is constructed in the form of a decision tree prior to the observation of unseen test instances. In the following we investigate whether the animacy classification task generalizes to a *lazy* machine learning algorithm: Memory-Based Learning. See section 5.2 for more on machine learning and the eager-lazy distinction.

Memory-based learning, a class of instance-based learning algorithms, has been applied successfully to a range of NLP tasks, such as named-entity recognition (Tjong Kim Sang 2002a), parsing (Kübler 2004) and semantic role labeling (Morante and Busser 2007). In the following we experiment with the application of memory-based learning (MBL) to the animacy data described above. We compare the performance of the MBL classifiers to the corresponding decision-tree classifiers and experiment with feature weighting for classification of lower frequency nouns.

6.5.1 Experimental methodology

All experiments make use of the TiMBL software package for Memory-Based Learning (Daelemans et al. 2004), see section 5.2.2 for more detail. As before, we employ leave-one-out training and testing, as well as McNemar's test for statistical significance.

6.5.2 Experiment 5: High frequency nouns

The first set of experiments using TiMBL was conducted on the set of 40 high frequency nouns presented in table 6.1 which have an absolute frequency in the corpus of a thousand occurrences or more. We also experiment with different parameter settings available in TiMBL, as well as combinations of these, in order to locate the optimal setting for classification.

Classification in Memory-Based Learning is performed by comparison of new instances to the set of training instances. The determination of relevant examples, the *k*-nearest neighbors, is therefore an important component of learning. In this respect, we may experiment with different sized neighborhoods. We may also vary the influence of the neighbors on classification, which is either performed by majority voting or a weighted voting, where closer neighbors are given more weight in determining the class of a new instance. With inverse linear scaling, the weights of neighbor instances in classification are scaled linearly to reflect distance in vector space.[86] In experiments 1-4 we examined the influence of the various features in classification. TiMBL supports *feature weighting*, where features are given differing weights during classification, or rather, in determining the *k*-nearest neighbors for a given instance. During parameter optimization we experiment with information-based feature weighting schemes Information Gain and Gain Ratio. These approximate the information contained in a feature during training and weight accordingly at classification.

We perform a set of experiments where we test all possible variations over the following parameters:

Feature weighting No feature weighting (NO), Information Gain (InfoGain), Gain Ratio

Nearest neighbors $k = \{1, 5, 7, 17\}$

Class voting Majority, Inverse Linear

The various parameter settings do not have a great effect on the results, and, in fact, the best accuracy of 95% is achieved using no feature weighting, $k = 1$ and the default, majority class voting. This indicates that all the features contribute positively to classification of the high frequency nouns. This corroborates the findings in the decision-tree experiments for this group of nouns, where the classifier employing the total set of features was the best performing. Varying the *k* parameter increases the set of nearest neighbors, but does not cause an

[86] The simpler, inverse linear scaling outperforms other class voting variants (inverse distance weight and exponential decreasing weights) in a majority of cases (Daelemans et al. 2004: 25).

Nouns	NO	InfoGain	GainRatio	ChiSquare	SharedVar
∼100	80.0	80.0	80.0	80.0	80.0
∼50	77.5	75.0	72.5	77.5	77.5
∼10	50.0	50.0	50.0	50.0	50.0

Table 6.10: Accuracy for classifiers with all features and only high frequency nouns and tested on lower frequency nouns, employing different sorts of feature weighting - no weighting (NO), Information Gain, Gain Ratio, Chi Square and Shared Variance.

improvement of results. This is explained by the fact that the training and testing examples are so few (only 40), so enlarging the neighborhood only serves to introduce errors. None of the results for the different experiments differ significantly from the result obtained using a decision-tree classifier on this group of nouns.

6.5.3 Experiment 6: Lower frequency nouns

In section 6.4 we observed that the performance of the decision-tree classifiers deteriorated notably when moving from high frequency to lower frequency nouns, ranging from 70.0% to around baseline accuracy of 52.5% for the set of lowest frequency nouns (∼10). Furthermore we found that by employing only a subset of the initial features, we were able to maintain a performance similar to that of the high frequency nouns also for the ∼100 and ∼50 nouns. This set of experiments examine the performance of classifiers trained using memory-based learning on the same data sets. We will in particular look at possibilities for replacing the feature back-off by various schemes for feature weighting.

6.5.3.1 No feature weighting

The first column in table 6.10 shows the results for the three groups of nouns when employing the best feature setting from the first group of experiments (no feature weighting, k=1, majority voting). As expected, these results are lower than the performance obtained when classifying the high frequency nouns (95%). The accuracy for the ∼100 nouns (80%) is better, however, not significantly better, than the corresponding result employing a decision-tree classifier (70%). The result for the ∼50 nouns (77.5%), however, is significantly better than the corresponding decision-tree result (57.5%), and differs only slightly

from the result for the ∼100 nouns. This is an interesting result because it indicates that MBL deals better with sparse data in this case.

A problem which was identified above is that it is the animate class in particular that suffers as data gets sparser. A majority of the features (SUBJ, GEN, PASS, ANAAN and REFL) target animacy in the sense that a higher proportion of animate than inanimate nouns exhibit the feature. As data gets more limited, the animate feature profiles become increasingly similar to the inanimate profiles, hence are frequently misclassified. If we examine the MBL results for the ∼50 nouns we find that the percentage of misclassified animate nouns has dropped dramatically compared to the decision-tree result, from 80% to 35%. A closer look at the most similar data point, i.e. the nearest neighbor, for each of the animate instances that were misclassified by the decision-tree classifier, but correctly classified by MBL, reveals that two instances are recurrent as singleton nearest neighbor in a clear majority of these cases (78%). These two instances are not, however, a nearest neighbor to any of the high frequency nouns from Experiment 1, hence are in that respect *outliers*.

The above observations are very much in line with those made in Daelemans, van den Bosch and Zavrel 1999, who argue against the editing away of exceptions. In their terminology, the two examples above would have a class prediction strength of zero for the high frequency nouns, making them very bad class predictors and candidates for editing.[87] However, these are exactly the examples responsible for the significant improvement of results for the animate ∼50 nouns. This highlights an important difference between decision tree learning and memory-based learning which resides in the fact that the former employs an eager machine learning algorithm, whereas the latter a lazy one. It is inherent in decision-trees that they represent some sort of generalization over the data. The C5.0 algorithm (Quinlan 1993) abstracts away from exceptions through pruning of the tree and always prefers smaller trees to larger ones. It is then highly likely that the properties of the two exceptions were pruned away in the decision-tree approach, leaving more of the ∼50 nouns for misclassification. MBL, on the other hand, conserves all examples and does not in this sense generalize over the data. This property proved to be beneficial in the classification of the lower frequency nouns and points to a possible advantage of lazy learning.

[87] The class prediction strength of an instance is the ratio of the number of times the instance is a nearest neighbor of another instance with the same class and the number of times that the instance is the nearest neighbor of another instance regardless of the class (Daelemans, van den Bosch and Zavrel 1999: 12).

SUBJ	OBJ	GEN	ANAAN	ANAIN	PASS	REFL
0.32	0.47	0.15	0.02	0.02	0.01	0.01

Table 6.11: Feature weights representing relative frequency in the training set of high frequency nouns.

6.5.3.2 TiMBL's feature weighting

TiMBL offers a range of different feature weighting schemes and the remaining columns in table 6.10 show the results from employing different feature weighting settings to the three different groups of nouns. As mentioned, the InfoGain and GainRatio settings are entropy-based measures (see section 5.2.1), whereas ChiSquare and SharedVar(iance) employ the χ^2-test of statistical significance to compute differences in the distributions of features in the training data.[88] These do not have any significant effect, and in a majority of cases actually have no effect at all. This is not so surprising, as the feature weights are calculated based on the training data, the high frequency nouns in this case. The measures are based on the informativity of the feature in this data set, and quite correctly points out the REFL feature as one of the most informative features. However, the exclusion of features in the earlier decision-tree experiments was not done on the basis of informativity, but rather on the basis of frequency, under the assumption that features more frequent in the data also provide more stable class predictors. In fact, the REFL feature is one of the rarest features in the feature set and does not hold up well against sparse data.

6.5.3.3 Frequency-based feature weighting

Based on the above considerations, we formulate an alternative, frequency-based feature weighting scheme. Here, we employ the conditional relative frequency of a feature f – the number of data points covered by a feature relative to all data points in the training data – as weights:

$$freqweight(f) = \frac{\sum_i freq(f, w_i)}{\sum_i \sum_j freq(f_i, w_j)}$$

The weights for each of the features are presented in table 6.11 below. As in the decision-tree experiments, the OBJ, SUBJ and GEN features are the top

[88] With numeric features, as in the present study, the feature values are discretized prior to application of the ChiSquare and SharedVar weighting schemes. SharedVar extends ChiSquare by correcting for degrees of freedom (Daelemans et al. 2004: 22).

Nouns	FreqWeights	SUBJ&OBJ&GEN	SUBJ&OBJ
∼100	90.0	95.0	90.0
∼50	77.5	75.0	87.5
∼10	67.5	67.5	62.5

Table 6.12: Accuracy obtained when applying classifiers with frequency-based feature weighting or combinations of the most frequent features to lower frequency nouns.

three most frequent, hence with the highest weights. The results from these experiments are shown in the first column of table 6.12. The feature weighting results in an improvement of the accuracy for the ∼100 and ∼10 nouns, however, this does not constitute a significant improvement compared to the results with no feature weighting (as reported in table 6.10).

Finally, experiments ignoring all features except the top three (SUBJ, OBJ, GEN) and top two (SUBJ, OBJ) most frequent features were performed. These results are reported in the remaining columns in table 6.12. Only the result for the ∼100 nouns (95%) employing the three most frequent features improve significantly on the result obtained with no feature weighting. However, none of these results differ significantly from the ones obtained in the corresponding decision-tree experiments.

6.5.4 Summary

The aim of this section was dual. First, we applied Memory-Based Learning to the classification of animacy for Norwegian common nouns. Second, these results were compared with corresponding results achieved when employing decision-trees to the same task in earlier experiments. For the set of highly frequent nouns, we achieved a classification accuracy of 95%. The different parameter settings in TiMBL did not affect the results notably and the best accuracy was achieved employing the most "basic" settings - no feature weighting, $k = 1$, and majority class voting. When this classifier was applied to the three groups of lower frequency nouns, the results deteriorated somewhat to 80% for the ∼100 nouns and 77.5% for the ∼50 nouns and quite drastically, a mere baseline performance, for the ∼10 nouns. Based on the previous results from the decision-tree classification, the results for the lower frequency nouns were to be expected.

In general, the performance of the Memory-Based learner did not differ significantly from that of the decision-tree classifiers, hence it is difficult to draw firm conclusions with respect to superiority of one machine learning algorithm

over another. As mentioned already, the lack of significance in differences may be partially due to the size of the data set, a question to which we shall return in chapter 7. We may conclude, however, that the results generalize across machine-learning algorithms.

6.6 Class granularity: classifying organizations

The earlier experiments have shown a binary classification of animacy to be worthwhile, with best accuracies approaching 95%. Zaenen et al. (2004) propose that nouns denoting *organizations* inhabit an intermediate position with respect to the two other main categories in an animacy hierarchy: *animate* and *inanimate*.[89] This section describes the automatic classification of organization nouns along the animacy dimension already established in the previous sections. In doing so, we assess distributional evidence for such a distinction and examine how a more fine-grained notion of animacy affects our earlier results in classification. Under the assumption that the set of features outlined above serve to capture important aspects of the property of animacy, we may in turn examine their generalization to a new set of nouns, namely nouns which denote organizations.

6.6.1 Data

The set of organization nouns employed in this study was collected while annotating sentences for a corpus study of animacy in Norwegian (Øvrelid 2004). The nouns consist primarily of *collective* nouns or nouns that have a regular *metonymic* usage where they are employed to refer to organizations, see section 3.2.4. Following Garretson et al. 2004, we annotate as organizations nouns which denote collectivities of humans which display group identity. The implicational hierarchy in (65) illustrates the distinction between the human class and the class of organizations.

(65) Implicational hierarchy distinguishing humans and organizations (Garretson et al. 2004):

```
chartered/official > temporally stable >
collective action, voice or purpose > collective
```

The hierarchy in (65) states that anything that is chartered or official is also temporally stable, has a collective voice etc., but not vice versa. The cut-off

[89] See section 3.2 for more on gradience in the animacy dimension.

	Animate		Inanimate		Organizations		
	Mean	SD	Mean	SD	Mean	SD	#
SUBJ	0.14	0.05	0.07	0.03	0.20	0.10	31537
OBJ	0.11	0.03	0.23	0.10	0.06	0.03	28046
GEN	0.04	0.02	0.02	0.03	0.12	0.06	16419
PASS	0.006	0.005	0.002	0.002	0.012	0.015	1203
ANAAN	0.009	0.006	0.003	0.002	0.0009	0.001	1047
ANAIN	0.003	0.003	0.006	0.003	0.005	0.003	1253
REFL	0.005	0.0008	0.001	0.0008	0.004	0.0017	840

Table 6.13: Mean relative frequencies and standard deviations for each class (20 animate, 20 inanimate, 20 organization nouns) and feature, as well as total data points for each feature (#).

point between human and organization is here set at being *temporally stable*. Both organizations and groups of humans (e.g. a crowd or a mob) can be collective and have a collective action, voice or purpose, however only organizations are in addition temporally stable and possibly also chartered or official.

In order to control for frequency effects, only the organization nouns that occurred more than 800 times in the corpus were included. This is close to the restriction imposed on the animate and inanimate nouns, hence make the groups directly comparable. As before, we ensure a uniform distribution in the training data and employ 20 organization nouns in the study. These nouns are presented in (66):

(66) *administrasjon* 'administration', *bank* 'bank', *bedrift* 'company', *bystyre* 'city council', *departement* 'ministry', *forening* 'association', *fylkeskommune* 'county', *komité* 'committee', *kommisjon* 'commission', *kommune* 'municipality', *kommunestyre* 'municipality board', *lag* 'team', *myndighet* 'authority', *organisasjon* 'organization', *parti* 'party', *regjering* 'government', *byrett* 'city court', *stat* 'state', *styre* 'board', *utvalg* 'committee'

Feature extraction for these nouns is performed in the same manner as for the animate and inanimate nouns. The mean relative frequencies obtained for each of these features is represented in the rightmost columns of table 6.13, where we also provide total data points for the data set consisting of all three classes. The total data points covered by each feature for the organization nouns only are as follows: SUBJ: 14724, OBJ: 3918, GEN: 8589, PASS: 626, ANAAN: 58, ANAIN: 309, REFL: 282.

It is worth noting that the relative frequencies of the various features differ from the animate and inanimate classes in several ways. We see that the

Feature	$\frac{\#\,\text{Org}_{anim}}{\#\,\text{Org}_{all}}$
ALL	1.00
SUBJ	0.95
OBJ	1.00
GEN	1.00
PASS	0.85
ANAAN	0.00
ANAIN	0.45
REFL	0.75

Table 6.14: Proportions of organization nouns classified as animate when classifying along the binary animacy dimension, employing all and individual features.

proportion of subjects and genitives is notably high for these nouns, significantly higher, in fact, than the frequencies observed for the animate nouns.[90] We also find that there is quite a bit of variation, as represented by the standard deviation. With respect to anaphoric reference, the organization nouns on average differ most markedly from the animate class, and are in this respect more similar to the inanimate nouns.

6.6.2 Experiment 7: Granularity

We now proceed to investigate further properties of the organization nouns by examining a three-way classification task, based on the same feature set as earlier. The experimental methodology is identical to the one employed in section 6.5.2 above, employing an MBL learner with leave-one-out cross validation.[91]

6.6.2.1 Animate or inanimate?

The first experiment involved testing the classifier trained only on the binary classified nouns on the new set of organization nouns. The main point of this experiment was to study how the classifier deals with these new nouns, whether they are classified as being animate or inanimate in any systematic way.

[90] As in section 6.2.3, statistical significance was calculated with an unpaired *t*-test.

[91] For all experiments with more than one feature we employed the most basic settings ($k = 1$ and no feature weighting), following the parameter optimization in section 6.5.2. For the experiments where we test only one feature at a time, we increased the number of nearest

Table 6.14 shows the proportion of organization nouns classified as animate by the old binary classifier. The results indicate that the organization nouns exhibit overall distributional properties which are more similar to the animate nouns than to the inanimate nouns. All of the organization nouns were classified as animate when all seven features were employed. By varying the features used during classification we obtain a clearer picture of where the animate characteristics of the organization nouns surface. Among the individual features the OBJ and GEN features classify all the organization nouns (100%) as being animate when employed individually. The subject (95.0%) and passive (85.0%) features are also strong indicators of animateness for the organization nouns. Another case worth noticing is the feature ANAAN, which classifies all the organization nouns as inanimate (hence 0.0% as animate). All of these results corroborate the proportional relationships observed in the relative frequencies for each feature. Just like the animate nouns, organizations have few object occurrences and a higher proportion of genitive forms, compared to the inanimate group of nouns. The results should not, however, be compared to the performance of the features as predictors of the classes animate and inanimate. Rather, the performance of the features in this context is neither good nor poor, but simply indicators of the degree of animacy these nouns exhibit in the various morphosyntactic constructions covered by the different features.

6.6.2.2 Three-way classification

The earlier classification experiment showed that the organization nouns had more in common with animate than inanimate nouns when it came to the morphosyntactic distributional properties measured by our seven features. Whereas the earlier experiment tested which other class of nouns the organization nouns are more alike, this experiment tests whether they are different enough to enable a three-way classification. We investigate whether the organization nouns might be better captured by an intermediate category 'organization'. This would indicate that these nouns constitute a natural group, which share a set of properties disjoint from the animate and inanimate.

In order to test the validity of a new category, a new data set was constructed by concatenating the data of high-frequency animate and inanimate nouns with the data set consisting of organization nouns. It thus contains a uniform distribution of classes, i.e. an equal number of cases from each of the three categories - animate, organization and inanimate. This new data set consists of distributional data, as summarized in table 6.13, for 60 nouns, 20 from

neighbors to three ($k = 3$) in order to achieve somewhat more informed similarities and control for the influence of outliers in the classification space.

Feature	Accuracy (%)
SUBJ	75.0
OBJ	68.3
GEN	71.7
PASS	43.3
ANAAN	50.0
ANAIN	26.7
REFL	36.7

Table 6.15: Accuracy for classifiers with individual features in 3-way animacy classification.

Used	Not Used	Accuracy (%)
1. SUBJ OBJ GEN PASS ANAAN ANAIN REFL		88.3
2. OBJ GEN PASS ANAAN ANAIN REFL	SUBJ	86.7
3. SUBJ GEN PASS ANAAN ANAIN REFL	OBJ	81.7
4. SUBJ OBJ PASS ANAAN ANAIN REFL	GEN	85.0
5. SUBJ OBJ GEN ANAAN ANAIN REFL	PASS	90.0
6. SUBJ OBJ GEN PASS ANAIN REFL	ANAAN	81.7
7. SUBJ OBJ GEN PASS ANAAN REFL	ANAIN	78.3
8. SUBJ OBJ GEN PASS ANAAN ANAIN	REFL	80.0

Table 6.16: Accuracy for classifiers with all features and 'all minus one' in 3-way animacy classification.

each class. A classifier was constructed and evaluated by means of leave-one-out cross validation. Since this is a three-way classification task, we assume a random baseline of 33.3%. The results are summarized in tables 6.15-6.16. The classifier constructed by means of all the seven features receive an accuracy of 88.3%, which constitutes a clear improvement in comparison with the 33.3% baseline. When it comes to the individual performance of the different features, shown in table 6.15), we see that the best performing features are the subject (75.0%), genitive (71.7%) and object (68.3%) features. As in the earlier experiments, these features stand out with respect to the others as stable class predictors.

The results indicate that organization nouns are distributionally different when it comes to their realization as subject, object and/or genitive modifier. We earlier remarked on the fact that the organization nouns were most alike the animate nouns with respect to these features in particular. How then is it possible that these also help distinguish them from the animate class? The cor-

pus data show that the organization nouns have higher proportions of subjects and genitives, and lower proportions of objects than the animate nouns. Rather than indicating an intermediate animacy status, this in a sense makes them more animate than the animate nouns themselves. The next section examines the distribution of organization nouns over these three features in a bit more detail.

6.6.3 The distribution of organizations

As mentioned earlier, organizations have a distribution which sets them apart from the strictly animate or inanimate nouns. The fact that these have been argued to occupy a middle ground with respect to animacy indicates that they have properties in common with both animate and inanimate nouns, see section 2.4 on gradience. As we shall see, this dual status is clearly reflected in their linguistic behaviour with regard to the chosen set of features.

6.6.3.1 Possessive relations

Let us start by examining closer the linguistic behaviour captured in the genitive feature and how the three classes differ. The organization nouns have a significantly higher proportion of genitive case marking than the animate class (GEN:$p<.0001$). It has been claimed that animate and inanimate nouns differ in the types of possessive relations expressed by the genitive construction (Rosenbach 2003). Animate nouns prototypically express ownership, e.g, *guttens bil* 'boy's car', body parts, e.g., *guttens arm* 'the boy's arm' and kinship terms, e.g., *guttens far* 'the boy's father'. Less prototypical relations for animate nouns are states (*guttens tilstand* 'the boy's condition') and abstract possession (*guttens liv* 'the boy's life'). Inanimate nouns have a more limited ranges of possessive relations available of which the part-whole relation (*husets tak* 'the house's roof') is the most prototypical. A hypothesis which would support the intermediate status of organizations, is that they have available the full range of possessive relations for both animate and inanimate expressions, and as a consequence are more frequent in the genitive.[92]

In order get a more detailed picture of the genitive relations the organization nouns occur in, we performed a corpus study of 50 genitive occurrences for three nouns from each of the three classes, alltogether 450 genitive instances.

[92]The purpose of Rosenbach's study is to investigate the genitive alternation in English. She excludes from her corpus study all kinds of collective nouns, thereby creating a strictly binary animacy opposition.

Tag	Description	Genitive relation
`ProtoAnim`	Prototypical for animate	body parts , kinship terms and permanent/legal ownership
`ProtoInan`	Prototypical for inanimate	part-whole
`Non-Proto`	Non-prototypical for animate and inanimate	states , abstract possession , non-part whole
`Nom`	Nominalizations	subject of nominalized verb

Table 6.17: Overview of annotation classes in the corpus study of genitive organization nouns.

The nouns are listed below:

(67) *gutt* 'boy', *kvinne* 'woman', *president* 'president'

(68) *bank* 'bank', *bedrift* 'company', *kommisjon* 'commission'

(69) *bil* 'car', *fly* 'plane', *hus* 'house'

The nouns and sampled corpus for the animate and organization classes were chosen randomly. However, since many of the inanimate nouns are rare in the genitive, we chose nouns which denote internally structured entities, which in principle can express the part-whole relation claimed to be the prototypical relation for this class (Rosenbach 2003).[93]

We annotated the resulting corpus of genitive constructions according to the distinctions made in Rosenbach 2003, but collapsed the non-prototypical categories for the animate and inanimate classes, covering states, abstract possession and non-part-whole relations.[94] We also distinguished nominalizations as a separate class.[95] The classes we annotated the genitives for are presented in table 6.17.

Nominalizations are closely linked to the argument structure of a verb (Grimshaw 1990), a dimension where animacy is clearly an important factor, see sections 3.5 and 3.6.2.[96] We annotated as nominalizations all constructions

[93]Not all of the inanimate nouns had as many as 50 genitive occurrences, which is why the results are given as relative frequencies rather than absolute ones.

[94]Non-prototypical genitive relations for inanimate nouns are all relations which are non part-whole, e.g. *husets skjebne* 'the house's destiny'.

[95]Rosenbach (2003) excludes genitive constructions where the head is a nominalization from her study, however, we chose to examine these as well.

[96]Grimshaw (1990) distinguishes between two types of nominalizations – result nominalizations and complex event nominalizations, where only the latter actually has an argument

	ProtoAnim		ProtoInan		Non-Proto		Nom	
Noun class	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Animate	42.7	5.7	n/a	n/a	38.7	9.3	17.3	5.0
Organization	10.7	4.1	21.3	3.7	22.0	7.5	44.7	10.0
Inanimate	n/a	n/a	60.6	16.4	38.7	17.2	0.0	0.0

Table 6.18: Mean percentages and standard deviations for the noun classes in different genitive relations from the corpus study of 450 genitive constructions.

where the head noun was clearly derived from a verb and where the genitive expressed the subject of such a verb. These may or may not express any additional arguments overtly, see (70) and (71) respectively:

(70) . . . *bystyrets* **vedtak** *om å slippe C02-avgiftene*
. . . city-council-GEN decision about to pass CO2-taxes-DEF
'The city council's decision to pass the C02 taxes'

(71) . . . *anvendt i forhold til bystyrets* **prioriteringer**
. . . used in relation to city-council-DEF.GEN priorities
'. . . used in relation to the city-council's priorities'

Nominalizations may also take the form of compounds, where the non-head expresses the object of the nominalized verb:[97]

(72) . . . *bystyrets* **boikottvedtak**
. . . city-council's boycott-decision
'the city-council's decision to boycott'

The results from the small corpus study are presented in table 6.18 where we present the aggregated means of the three nouns from each class for a total of 450 genitives. We find that the animate and inanimate nouns follow the patterns predicted in Rosenbach 2003, with a greater percentage of prototypical class

structure. The genitive in the former case is simply a modifer whereas in the latter case, it is a suppressed subject and the other arguments (if any) are obligatory. Many nominalizations are ambiguous between the two readings, and Grimshaw posits seveal tests which elucidate the difference. For instance, adverbials which relate to event structure, such as *constant* and *frequent*, force an event reading, and hence require the internal arguments to be expressed:

(i) The constant assignment *(of unsolvable problems) is to be avoided.

[97] We did not attempt any systematic disambiguation with respect to result vs. event nominalizations (Grimshaw 1990). This distinction seems difficult to apply in practice and it is not clear that the tests proposed hold also for Norwegian.

usages than non-prototypical. However, for the animate class this difference is not significant and we see a large proportion of more abstract possessive relations, e.g. *gutters liv* 'boys' life', *kvinners status* 'women's status', *presidentens moral* 'the president's morale'. It is also apparent that the organization nouns in the study do in fact occur with all the possible possessive relations, i.e. the relations typically associated with both animate and inanimate nouns:

- Possession:

  (73) …*bankens sedler og mynter* 'the bank's bills and coins'

- Abstract/state:

  (74) …*bedriftens lønnsomhet* 'the company's equity'

- Nominalization:

  (75) …*bankens vedtak* 'the bank's decision …'

- Part-whole:

  (76) …*kommisjonens leder* 'the commission's leader'

It is interesting to note that it is not the prototypical animate relations which dominate the organization usage, but rather the nominalizations. This observation fits nicely in with the fact that organizations function so frequently as transitive, main clause subjects. However, it is also clear that the prototypical inanimate part-whole relation accounts for a fair portion (21.3%) of the genitive usages of organizations.

Our initial hypothesis is thus supported; we find that organizations may occur with the posseessive relations associated with both animate and inanimate nouns, hence have a higher frequency of genitives. This confirms their intermediate status.

6.6.3.2 Subjecthood

Organizations occur significantly more often as transitive subjects than animate nouns, and subjecthood turned out to be a successful predictor in our three-way classification. This is not that surprising, as organizations are per definition decision-making entities, hence would be expected to exhibit a high level of agentivity. If we examine the lexical verbs which occur most often in

our material with organization subjects we find the following list, sorted by frequency:[98]

(77) *ha* 'have', *gi* 'give', *være* 'be', *fastsette* 'determine', *foreslå* 'suggest', *mene* 'mean/think', *få* 'get', *legge* 'lay', *gjøre* 'do', *ta* 'take', *vurdere* 'assess', *se* 'see', *finne* 'find', *styre* 'govern', *bestemme* 'decide', *foreta* 'perform', *understreke* 'underline', *kreve* 'demand', *anta* 'assume', *bruke* 'use', *bli* 'become', *anbefale* 'recommend', *si* 'say', *betale* 'pay', *kunne* 'know', *ønske* 'wish', *stille* 'ask', *utarbeide* 'develop', *sette* 'set'

The majority of the verbs in (77) are clearly agentive verbs which denote events of decision-making and opinionating. Furthermore, we find that the verb *ha* 'have' is the most frequent verb for the organization nouns. In a study of Swedish, Dahl and Fraurud (1996) find that inanimate transitive subject are to a large extent subjects of the verb *ha* 'have'. Some examples from our study of organization nouns are provided below:

(78) *Banken* *har* *innskudd* *i* *Rogaland*
bank-DEF has deposits in Rogaland
'The bank has deposits in Rogaland'

(79) *Administrasjonen* *har* *ansvaret* *for* ...
adminstration-DEF has responsibility-DEF for ...
'The administration are responsible for ...'

(80) *Bedriften* *har* *for* *svake* *eiere*
company-DEF has too weak owners
'The company has too weak owners'

(81) *Foreningen* *har* *100* *medlemmer*
association-DEF has 100 members
'The association has 100 members'

We see that the construction is clearly compatible with inanimate nouns, as it expresses a part-whole relationship as in (80)–(81), as well as possessive ownership as in (78) and abstract state in (79).[99]

The distribution of organizations as subjects is compatible with the findings of the corpus investigation for the genitive construction. Organizations are frequent in this construction as they may occur in both animate and inanimate guise. They may occur with clearly agentive verbs, as well as with the possessive verb *ha* 'have' to mark a part-whole reading.

[98] These are *lexical head* verbs only, i.e. verbs that occur either as the single finite verb in a clause or as non-finite participle along with a finite auxiliary.

[99] In the example in (79) *ha* 'have' is actually more like a light verb which together with its complement forms the complex predicate 'be-responsible'.

6.6.3.3 Objecthood

Organization nouns occur infrequently on average as direct objects, in fact, significantly less often than regular animate nouns ($p<.0001$). In order to get a better idea of the differences in direct object realization between the three classes, we performed a corpus study. From each of the three classes three randomly chosen nouns were selected and we extracted 80 random examples of object occurrences for each of these.[100] The three nouns are presented in (82)–(84):

(82) *jente* 'girl', *lege* 'doctor', *president* 'president'

(83) *departement* 'ministry', *kommisjon* 'commission', *regjering* 'government'

(84) *hus* 'house', *opplysning* '(piece of) information', *penge* 'coin'

The corpus studies examining the distribution of organizations in genitive and subject constructions examined more fine-grained semantic relationships expressed by these syntactic constructions and we examined differences in frequency distributions between the three classes. Objects are typically patients or themes, generally expressing an affected participant in the event denoted by the verb (Dowty 1991). When annotating the data set, we distinguish between regular direct objects and objects which have a dual status in that they are also logical subjects of a following active subordinate clause. We earlier assumed that the organizations occupied an intermediate position on the animacy hierarchy which gave them a flexibility with respect to the genitive construction. When functioning as objects, this duality or intermediacy may be reflected in the types of structures where an argument at one and the same time stands in a thematic relationship with two different verbs, i.e. functioning as subject and object at the same time. In the annotation we adopted the following definitions and annotated for two classes:

Regular object the noun is a direct object of a transitive verb and is not a subject of a following subordinate clause

Dual object the noun is a direct object but is also the logical subject of an ensuing subordinate clause

[100] We chose to extract 80 examples as there turned out to be quite a bit of noise in the tagger's analysis of objects and we wanted to ensure a reasonable amount of data. Simple errors in the automatic analysis were manually filtered out from the annotation, leaving only the real direct objects for analysis. This amounted to a total of 406 object instances.

Note that we are not requiring the dual objects to be logical objects of the verb, only structural ones. This allows us to include so-called ECM or raising-to-object constructions, where the object is not a thematically entailed argument of the matrix verb, as well as cleft constructions where the structural object is the logical subject of both the matrix and subordinate clause. Examples of the types of constructions included in this annotation category are given in (85)–(89) below. (86) and (85) provide examples of an ECM and control construction respectively, and in (87) we have an example of a cleft construction with a copula verb, an expletive subject and an object which is the subject participant of the obligatory relative clause. In (88) the object argument is modified by a subject-relative clause, whereas it in (89) forms a small clause construction.[101]

(85) *De anklagde kommisjonen for å utnytte sin posisjon*
they accused commission-DEF for to abuse their/its position
'They accused the commission of taking advantage of their/its position'

(86) *De fikk kommisjonen til å utnytte sin posisjon*
they got commission-DEF to to abuse their/its position
'They made the commission take advantage of their/its position'

(87) *Det er kommisjonen som utnytter sin posisjon*
it is commission-DEF that abuses their/its position
'It is the commission that takes advantage of their/its position'

(88) *De tilhørte kommisjonen som utnyttet sin posisjon*
they belong commission-DEF that exploited their/its position
'They belong to the commission that took advantage of their/its position'

(89) *De anså kommisjonen som suspekt*
they considered commission-DEF as suspicious
'They considered the commission to be suspicious'

Table 6.19 shows the mean frequencies, along with standard deviations, of regular direct objects and objects which are logical subjects of following subordinate clauses in the data collected for the nouns from each class. The organization nouns are clearly more frequent in the dual object position. This relates to the fact that they occur so frequently as subjects, and most often as agents, as we saw in the above section. The subordinate verbs in the dual object constructions are overwhelmingly agentive:

[101] The examples in (85)–(89) are constructed.

	Regular object		Dual object	
Noun class	Mean	SD	Mean	SD
Animate	79.4	6.4	20.6	6.4
Organization	53.3	6.7	46.7	6.7
Inanimate	98.3	1.5	1.7	1.5

Table 6.19: Mean percentages and standard deviations for the noun classes in different object relations from the corpus study of 406 objects.

(90) *Det er dette departementet som skal* **ta** *avgjørelsen*
it is this department-DEF that shall take decision-DEF
'It is this department that will make the decision'

(91) *Kina hadde tidligere bedt den danske regjeringen om*
China had earlier asked the Danish government-DEF
å **avlyse** *besøket*
to cancel visit-DEF
'China had earlier asked the Danish government to cancel the visit'

In comparison, the animate nouns clearly more frequently occur as regular transitive objects, with almost 80% of the animate object occurrences in the corpus data. The inanimate nouns hardly have any dual object constructions at all (3 instances, 1.7%) and the ones we find consist of a predicative subordinate clause, as in (92), or are examples of metaphorical extensions through anthropomorphization, as in (93):

(92) *De har plikt til å gi alle opplysninger som er*
they have duty to give all information that is
nødvendig
necessary
'They are obliged to provide all information necessary'

(93) *Men det finnes et og annet hus som eier sin eier*
but there exists one and another house that owns its owner
'But there are houses that own their owners'

In section 3.6.2 above we discussed the notion of *individuation*, which is often mentioned as one which either intersects with animacy (Dahl and Fraurud 1996; Yamamoto 1999) or which subsumes animacy along with other properties such as definiteness and referentiality. The level of inidviduation relates to the degree to which we view an entity in the discourse as being a "clearly delimited and identifiable indiviual" (Dahl and Fraurud 1996). Animates will

tend to be high in individuation, hence making them well-suited as objects (Hopper and Thompson 1980).[102] Organizations are referentially mass-like and abstract. They do not point out a clearly delimited individual, however are often in definite form, e.g., *regjeringen, byretten, kommisjonen* 'the government, city-court, commission', because there is only one of them with respect to particular time and place.

One point, which we have mentioned only in passing, but which relates to the dimension of individuation, is pronominal reference. Animate nouns have a strong tendency for reference by a personal pronoun, whereas this is lower for inanimates (Dahl and Fraurud 1996). Our data on the pronominal features (ANAAN, ANAIN) have measured the pronominal reference by personal pronouns, i.e., the pronouns which clearly show the animacy of their referent. An interesting property of the organizations is that they, like the inanimate nouns, are very infrequently referred to by means of an animate personal pronoun. That is, even though the type of sequence in (94) is perfectly grammatical, it is hardly encountered:[103]

(94) *Komiteen$_i$* *ankom i formiddag. De$_i$ ville* ...
committee-DEF arrived this morning. They wanted ...
'The committee arrived yesterday. They wanted ...'

In this respect, the organization nouns differ distinctly from the animate nouns ($p<0.0005$) and behave more like the inanimate nouns. This indicates that they are not individuated enough to merit pronominal reference.

6.6.4 Conclusion

The results from a set of classification experiments indicated that organizations behave linguistically in a manner which clearly sets them apart from regular animate or inanimate common nouns, a behaviour which can be exploited in automatic classification. With reference to the morphosyntactic SUBJ, OBJ and GEN features, we investigated the nature of this difference in terms of more fine-grained syntactic and semantic distinctions. In particular we found empirical evidence for the intermediacy of organizations, where these nouns may take on both animate and inanimate readings. There is also a clear agentive pattern which emerges in the distribution of these nouns; they occur frequently

[102] Note that the predictions of Hopper and Thompson (1980) with respect to direct objects are contrary to those predicted by prominence-conserving theories of argument realization, such as Aissen 2003. On this view, animate entities are marked objects precisely because they are prominent.

[103] The example in 94 is constructed.

as transitive subjects, have a large proportion of nominalizations in the genitive and even as objects, often dually hold an agentive role in a following subordinate clause. It is in the very semantic nature of organizations that they are action-taking and decision-making entities, so it should not be surprising that this shows up in their linguistic behaviour. What makes them interesting from the point of view of animacy, is that they exhibit a duality compatible with both extremes of the animacy dimension – highly individuated and agentive or less individuated, mass-like and internally structured.

6.7 Unsupervised learning as class exploration

We have seen in the above sections that the chosen set of linguistically motivated distributional features approximate the property of animacy well. Section 6.6 investigated how the addition of organization nouns to the data set influenced the classification results and also investigated the proposal that these nouns constitute an intermediate animacy category empirically, both through machine learning and corpus studies. In the following we will apply the unsupervised machine learning technique of clustering, see section 5.2.3. We employ clustering primarily as a technique for *data exploration* (Boleda, Badia and Batlle 2004; Boleda 2007) and thereby provide an additional perspective on the task of animacy classification. The main goal is to examine the categories which, under a distributional view of the nouns and based on our selected features, emerge when we do not classify according to a predefined set of classes. We also assess whether the unsupervised categories correspond in any systematic manner with the categories employed in our earlier, supervised experiments.

6.7.1 Experiment 8: Clustering

Experiment 8 investigates how the individual nouns cluster based on their distributional properties and examine the two levels of granularity discussed earlier. The advantage of employing an unsupervised technique is that there is no bias towards a predefined set of classes, but rather a direct focus on the properties of the nouns.

6.7.1.1 Experimental methodology

We employed the same data sets as in Experiment 1 and 7, i.e. the set of high-frequency nouns with binary – animate, inanimate – classification, as well as

the data set with three-way – animate, inanimate and organization – classification. For clustering we employed the Cluto software with default settings, see section 5.2.3.[104] Clustering is *partitive*, whereby a clustering solution is obtained by partitioning the data set into an increasingly larger set of clusters until the desired *k* number of clusters is obtained. At each partitioning of the vector space, a *criterion function* is optimized. We employ an internal criterion function which maximizes the inter-cluster similarity of each cluster, where similarity is computed with the cosine function.[105] The parameter which is varied in the experiments is the *k*-parameter which specifies the desired number of clusters.

6.7.1.2 Overview and evaluation of cluster solutions

When a clustering solution has been obtained for a data set, it must also be presented in a manner which provides an overview of the content of each cluster. There are several different cluster properties which in various ways provide a summary of a cluster solution. We focus on the following:[106]

Internal quality: The *tightness* of a cluster can be obtained by looking at the average internal similarities of the nouns contained in the cluster and *overlap* by looking at the average similarity of the elements in the cluster with the rest of the elements in the data set. These are internal quality measures, as they do not make use of class information which is external to the clustering solution.

Features: A cluster may also be summarized by the features which were most important in obtaining the particular cluster solution: the *descriptors* are the features which contribute the most to the similarities of the instances in the cluster and the *discriminators* are features which contribute the most in distinguishing the cluster elements from the total set of instances.

The above measures do not take into account any predefined classification of the instances. However, since our data set does contain classified instances, we may also take these into account when evaluating the cluster solution. We

[104]The default settings in Cluto are: clustering by repeated bysections (rb) with the criterion function I2.

[105]To be precise, the criterion function maximizes the similarity between each member of a cluster and the *centroid* vector of the cluster, which is obtained by averaging over the vectors in the cluster.

[106]The tightness and overlap of a cluster solution corresponds to the ISim and ESim measures of Karypis (2002).

	Size	Tightness	Overlap	Anim	Inan
1	1	1.0	0.5	0	1
2	21	0.97	0.77	2	19
3	18	0.97	0.77	18	0

Table 6.20: Cluster solution with best internal quality for the high-frequency animate-inanimate data; ordered by decreasing tightness−overlap.

thus examine information that was not employed during clustering, hence is *external* to the clustering as such. The *purity* of a clustering solution measures the degree to which a proposed cluster contains instances of the same class, and gives the proportion of cluster elements which are of the majority class. For a given cluster S_r of size n_r, we have that (Zhao and Karypis 2003):

$$Purity(S_r) = \frac{1}{n_r} \max_i(n_i^r)$$

where n_i^r is the number of instances assigned to the ith class, assigned to the rth cluster. The purity of the entire cluster solution is computed as the weighted sum of the purities of the individual clusters.

6.7.1.3 Anim-Inan data

A clustering experiment was performed on the data set consisting only of high-frequency animate and inanimate nouns, where the number of clusters was varied: $k = \{2,3,4,5,6\}$. A 3-way clustering solution obtained the best internal quality, i.e. in terms of average tightness and overlap of the clusters. Table 6.20 shows the clustering solution, where each row represents a cluster. The purity of the solution is 0.95, which is high. We find that the cluster solution with the best internal quality has two clusters which roughly correspond to our classes of animate (cluster 3) and inanimate (cluster 2), and an additional cluster which consists of only one element, namely the noun *dag* 'day' (cluster 1). We find that the genitive feature was the descriptive and discriminating feature for this cluster. Temporal expressions are often mentioned in work on genitive constructions (Rosenbach 2002), because they behave unlike other inanimate nouns in this respect. The noun *dag* 'day' has an unusual high proportion of genitive occurrences (0.15) compared with the average inanimate noun (0.02).[107] Cluster 2 consists primarily of inanimate nouns, as well as the

[107]The noun *dag* 'day' has a high proportion of genitive instances also compared to another temporal noun in the data set, namely *night* 'natt'. In the data for these two we find the expected

	Size	Tightness	Overlap	Anim	Inan
1	62	0.92	0.62	56	6
2	58	0.91	0.62	4	54

Table 6.21: 2-way cluster solution for the high-frequency, ∼100 and ∼50 animate-inanimate data; ordered by decreasing tightness−overlap.

two animate nouns *barn* 'child' and *venn* 'friend'. We noted already that the useful features of subject and objecthood do not give sufficient distributional evidence for nouns of this type. Children and friends are typically entities that we possess, and are not that frequently agentive. We find that for this cluster, the highest ranked descriptive features are the object, subject and genitive features.

In order to evaluate the effect of sparse data on the clustering, an identical experiment was performed on the entire data set of animate and inanimate nouns, i.e. the concatenation of the data on high-frequency nouns of absolute frequencies ∼100 and ∼50. The cluster solution is presented in table 6.21 and we find that the results largely corroborate the ones obtained in the supervised classification experiments – feature back-off enables a good distinction between the two classes. A two-way clustering yields a total purity of 91.7, where the SUBJ and OBJ features are the primary features employed.

6.7.1.4 Anim-Org-Inan data

A similar clustering experiment with $k = \{2,3,4,5,6\}$ was performed on the data set containing nouns from the three classes of animate, organization and inanimate. Whereas the optimal clustering solution according to the internal quality metrics for the binary classification also obtained the highest external quality, i.e. purity, this is not the case for the ternary (Anim-Org-Inan) data. The clustering solution with $k = 2$ in fact achieves the best internal quality, but the lowest purity. Table 6.22 shows the clustering solution, where each row once again represents a cluster. The purity of the solution is 0.65.

The cluster solution once again supports the main distinction between animate and inanimate nouns, but also gives an indication of gradience. Cluster

construction where the genitive expresses the temporal situation of the head noun, e.g. *dagens møte* 'the day's meeting', *nattens match* 'the night's match'. However, *dag* 'day' also occurs in a more general, fixed expression meaning something like 'the current state of', as in *dagens samfunn/skole/generasjon* 'the current state of society/schools/generation'. This usage is dominant in the data for this noun and contributes to explain the difference in distribution between these two, otherwise semantically similar nouns.

	Size	Tightness	Overlap	Anim	Org	Inan
1	23	0.94	0.66	4	0	19
2	37	0.91	0.66	16	20	1

Table 6.22: 2-way cluster solution with best internal quality for the anim-org-inan data; ordered by decreasing tightness−overlap.

	Size	Tightness	Overlap	Anim	Org	Inan
1	23	0.94	0.66	4	0	19
2	18	0.93	0.69	0	17	1
3	19	0.97	0.82	16	3	0

Table 6.23: 3-way cluster solution for the anim-org-inan data; ordered by decreasing tightness−overlap.

1 contains all the inanimate nouns excluding the aforementioned *dag* 'day'. In addition, four animate nouns are clustered with the inanimate nouns. The feature descriptors for this cluster lists the OBJ feature as being of particular importance, and, indeed, the cluster contains the aforementioned *barn* 'child' and *venn* 'friend'.

We find that all the organization nouns and a majority of the animate nouns have been assigned to one cluster (cluster 2). The most important features for the creation of the clusters were primarily the subject and object features. It is clear that the linguistic behaviour of the organization nouns captured in these features is more distinct from the inanimate nouns than the animate nouns themselves. It is not surprising then, that the organization nouns form the basis for a cluster along with a majority of the animate nouns.

The cluster solution for $k = 3$ is presented in table 6.23. This solution has a sligthtly lower internal quality, but better purity (0.87). We find that cluster 1 is identical to cluster 1 in the $k = 2$ experiment presented in table 6.22. The two additional clusters have been created by splitting the initial cluster 2 into two separate clusters (clusters 2 and 3 in table 6.22). We find that the classes of organization and animate correspond fairly well with this partitioning of the instances; cluster 2 consists primarily of organization nouns, whereas cluster 3 consists in majority of animate nouns. Even so, the internal quality measure of overlap indicates that these two clusters are very similar, hence the internal quality of the cluster solution deteriorates.

We stated initially that the main goal of this section was to employ unsupervised machine learning for data exploration. The results have clearly indicated

the existence of a distinction between animate and inanimate nouns in the data itself. This shows that the two classes are natural and not simply superimposed on the data. Just like the supervised experiments and the corpus studies, however, the clustering experiments emphasize the gradience of the animacy dimension. The granularity of the animacy dimension has been addressed by looking at a set of organization nouns. The clustering experiments pick up on the same trend as shown in the preceding sections; the organization nouns are more similar to the animate nouns than the inanimate nouns but constitute a group in the sense that they have a common set of distributional properties. As the corpus studies in 6.6.3 showed, however, our features are morphosyntactic approximators of more fine-grained syntactic and semantic distinctions, where the organization nouns further confirm their intermediate status.

6.8 Summary of main results

At the beginning of this chapter we formulated a set of research question to be addressed. The general *viability* of a method for animacy classification has been addressed throughout the work described above. We have seen that animacy may be acquired through a set of morphosyntactic features capturing the morphosyntactic distribution of nouns and emphasizing the clear correlation between syntax and semantics with regard to animacy. We formulated a set of *features* which approximate linguistic correlations between animacy and distinctions in argumenthood, agentivity and individuation. We also tested the importance of the various features in classification and obtained results that show the importance of animacy in argument differentiation. We found that the SUBJ and OBJ features were central predictors of animacy throughout the above sections. With respect to the generalizability of the method, we examined the *robustness* of classification in the face of sparse data. It is not surprising that sparse data affects a method which relies on distributional features negatively, and this was established for our method as well. However, we found that the classification accuracy obtained for high frequency nouns (with absolute frequencies $>$1000) can be maintained for nouns with considerably lower frequencies ($\sim$50) by backing off to a smaller set of features at classification. We also examined the generalizability of the method across *machine learning algorithm*. The switch from decision-trees to memory-based learning gave slight improvements and highlighted general differences between eager and lazy learners. We also looked at unsupervised learning and found that the same class distinctions were made without a set of supervised training examples. Finally the issue of gradience in the animacy dimension was approached through experiments with *class granularity* and a more fine-grained, three-way

classification was tested empirically. The results underline the main opposition between animate and inanimate, however also show that finer distinctions may be approached empirically through data-driven animacy classification.

The experiments described above were geared primarily towards the use of machine learning to evaluate the theoretical proposals regarding animacy and argumenthood in particular. In order to make a detailed study feasible, a small set of manually selected nouns were employed. A uniform distribution of classes was maintained in the data in order to ensure sufficient data for classification. However, both of these assumptions present clear idealizations. A natural next step would be to test the method developed in the current chapter further, by applying it to a larger set of nouns and evaluate the extent to which it is scalable. This is the topic of chapter 7.

7 ACQUIRING ANIMACY – SCALING UP

The experiments reported in chapter 6 allowed us to explore several topics related to animacy classification, such as feature selection, data sparseness and class granularity with a manually selected set of nouns. This chapter reports on experiments dealing with the scaling up of animacy acquisition and in doing so assessing the generalizability of the methods described in the previous chapter. In order to apply the supervised learning methods tested in the previous chapter, we need a set of nouns annotated for animacy. In section 7.1, we will examine and assess annotation schemes for animacy and discuss the annotation for person reference found in a Swedish treebank. Section 7.2 presents the resulting data set and discusses data representation in terms of features. In section 7.3, we will describe a set of classification experiments on the resulting data set.

We address the following questions:

Animacy annotation Which criteria may be employed for annotation of animacy? Which properties should animacy annotation have in order to support lexical acquisition?

Transfer of method Will the general classification method and features transfer to another data set and to a different, although closely related, language?

Robustness revisited To what extent is classification robust to data sparseness?

Class distribution How will a non-uniform distribution affect the results?

Feature importance Which features are important in the scaling up of animacy classification?

Machine learning algorithm revisited Do we observe any significant differences between eager and lazy machine learning algorithms in animacy classification?

Class granularity Can we find evidence for gradience of animacy in human annotation for this property? Can we find evidence for gradience of animacy in the experimental results?

7.1 Obtaining animacy data

This section examines methods for obtaining data on animacy. We will start out by discussing criteria for animacy annotation with basis in previous annotation schemes and move on to present the manually annotated data in the Swedish treebank Talbanken05.[108]

7.1.1 Animacy annotation

Annotation for animacy is not a common component of corpora or treebanks. However, following from the theoretical interest in the property of animacy, as discussed in chapter 3, there have been some initiatives directed at animacy annotation of corpus data. In the following, we present an annotation scheme developed for English and a small annotation study aimed at testing the scheme for Swedish.

7.1.1.1 Annotation schemes

Corpus studies of animacy (Yamamoto 1999; Dahl and Fraurud 1996) have made use of annotated data, however they differ in the extent to which the annotation has been explicitly formulated as an *annotation scheme*. The annotation study presented in Zaenen et al. 2004 makes use of a coding manual designed for a project studying genitive modification (Garretson et al. 2004) and presents an annotation scheme for animacy, illustrated by figure 7.[109] The main class distinction for animacy is three-way, with subclasses under two of the main classes:

- Human (HUM)
- Other animate: Organizations (ORG), Non-Human Animates or Animals (ANIM)

[108] All examples in this section are taken from the Talbanken05 treebank.

[109] The fact that the study focuses on genitival modification has clearly influenced the categories distinguished, as these are all distinctions which have been claimed to influence the choice of genitive construction. For instance, as mentioned earlier in section 6.7.1, temporal nouns are frequent in genitive constructions, unlike the other inanimate nouns.

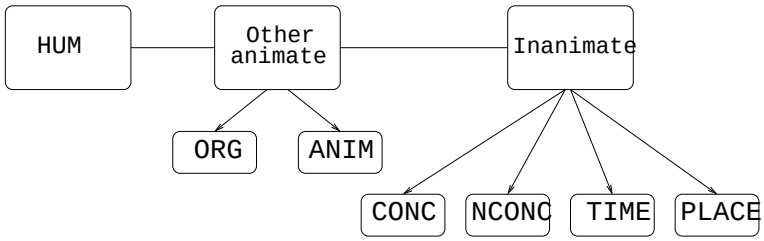

Figure 7: Animacy classification scheme.

- Inanimate: Concrete (CONC), Non-Concrete (NCONC), Time (TIME), Place (PLACE)

The 'Other animate' class further distinguishes Organizations and Animals. Within the group of inanimates, a distinction is made between concrete and non-concrete inanimate. The concrete class is employed when the markable refers to "'prototypical' concrete objects or substances. Excluded are things like air, voice, wind and other intangibles. Body parts are concrete" (Zaenen et al. 2004: 4). The non-concrete class is the default class, and is employed for markables that refer to entities that are not prototypically concrete but clearly inanimate. This includes events, abstract concepts or generalizations. Place and time expressions are also distinguished within the main category of inanimate.

7.1.1.2 Annotation study

A small annotation study for Swedish was performed on the Talbanken05 material in order to test the scheme proposed in Garretson et al. 2004 and get an overview of the distribution of the different classes. In order to do so, we annotated a semi-random sample from the prose section of Talbanken05 consisting of 108 sentences, and 383 markables.[110] The markables in the study include all common nouns in the sample, with a few exceptions.[111]

The resulting distribution of annotated markables over the classes is presented in table 7.1. The 'non-concrete' category is in clear majority (61.4%), followed by the 'Human' class (16.2%). Due to this, the main category 'Inanimate' is also in clear majority, accounting for 78.6% of the markables. We find

[110] To be precise, every 60th sentence was extracted from the prose section of Talbanken05 and annotated.

[111] All common nouns were annotated with the following exceptions: first conjuncts in abbreviated compound coordination constructions of the type 'N- och NN', e.g. **familje**- *och statsbudgeten* 'family- and state budget', parts of functional multiword units, e.g. *på* **grund** *av* 'for reasons of' and quantifying nouns, e.g. *en* **rad**$_{DET}$ *förmåner* 'a row (of) benefits'

Class	#	%	Sub-class	#	%
HUM	62	16.2	HUM	62	16.2
Other animate	20	5.2	ORG	13	3.4
			ANIM	7	1.8
Inanimate	301	78.6	CONC	40	10.4
			NCONC	235	61.4
			TIME	13	3.4
			PLACE	13	3.4
Tot	383	100.0		383	100.0

Table 7.1: The distribution of markables over the different classes and sub-classes in the annotation study.

that the intermediate category 'other animate' is quite infrequent, accounting for only 5.2% of the nouns in the sample, with animals at 1.8% and organizations at 3.4%.

7.1.1.3 Reference as annotation criterion

In the pilot annotation study we followed the annotation scheme described in Garretson et al. (2004): to annotate the markables according to the animacy of their *referent* in the particular context. However, using reference as a criterion can be problematic. First of all, by doing so one implicitly assumes that all markables refer and hence have a determinable referent. Secondly, by taking a context-dependent view of animacy, there is a danger that the resulting annotation does not deal with animacy at all, but rather a context-dependent notion of individuation or accessibility. We will examine these issues in turn below.

Garretson et al. (2004) state that "when coding for animacy [...] we are not considering the nominal per se (e.g., the word 'church'), but rather the entity that is the referent of that nominal (e.g. some particular thing in the real world)". This indicates that for all possible markables, a referent should be determinable. In the annotation of the Swedish sample, however, it became clear that this assumption is problematic. In (95) below, we find an example of person-denoting nouns with generic readings.

(95) *Hyressättningen* *grundas* *på* *avtal* *mellan* **hyresvärd** *och*
rent-setting-DEF built on agreement between landlord and
hyresgäst
tenant
'The rent is based on an agreement between landlord and tenant'

The referent of a generic reading differs from a specific one in being a 'reference to kinds' (Carlson 1980). In a very narrow interpretation of reference, one may want to exclude generic readings completely. However, it is not the case for these that the animacy of the markables may not be determined.

Another problematic area with regard to reference deals with noun phrases which incur a predicational reading, e.g. (96)–(97) below:

(96) *Det är en uppöver öronen förälskad* **flicka**
it is a up-over ears in-love girl
'That is an utterly infatuated girl'

(97) *Han är* **representant** *för Svenska Kyrkan*
he is representative for Swedish Church-DEF
'He is a representative for the Swedish Church'

Both of the examples in (96)–(97) are descriptive predicatives, which serve to classify or characterize the predicated argument further. These types of predicatives may be employed clearly referentially in Swedish with an indefinite article and often with a deictic argument (Teleman, Hellberg and Andersson 1999), as in (96). It is clear that the lack of indefinite article "dereferences" the predicative:

(98) *Det där är *flicka/*representant.*
that there is girl/representative

For these classificational predicatives, the referent is rather a generic role, as in (97) above. One might claim that rather than being referential, these express a predication which concerns the subject and hence are propositional in nature.

The above discussion of generics and predicatives illustrates that relying on reference as a criterion for annotation can be problematic. This brings us to our second problem with the annotation principle of reference. If one assumes that a reference may be determined for all markables, there is risk that the notion of animacy becomes diluted. In particular, such an approach confuses with animacy a range of related factors such as definiteness or individuation. Information to this end is present in the choice of NP-type, the formal definiteness of the NP, its abstractness and accessibility in the discourse. Additional annotations expressing these types of information are possibilities which might be explored, but, should possibly be kept separate from the animacy dimension. The above discussion ties in with the proposal in chapter 3 that animacy is largely a denotational property of nouns. Whereas reference may vary with the linguistic context, denotational properties are stable across contexts. In chapter 6, this assumption lead us to the hypothesis that aggregated frequency data, i.e.

data collected at the level of lemmas, could be exploited in animacy classification.

7.1.2 Person reference in Talbanken05

The Swedish treebank Talbanken05, see section 5.1.1, expresses a distinction for nominal elements between reference to person and non-person in the lexical layer of its annotation. The annotation manual (Teleman 1974) states that a markable should be tagged as person if it may be replaced by the interrogative pronoun *vem* 'who' and be referred to by the personal pronouns *han* 'he' or *hon* 'she'. This goes for singular markables, whereas for their plural counterparts, the instruction is to annotate them as one would their singular forms. The following describes the annotation in a bit more detail.

7.1.2.1 Annotation scheme

As mentioned earlier, the annotation in the original Talbanken (the MAMBA scheme) consists of a column-based markup, where two main *layers* may be distinguished - a lexical and a syntactic one (Teleman 1974). The annotation for the distinction between person and non-person reference is found in the lexical layer, along with information about part-of-speech and varying other types of semantic information, depending on the part-of-speech in question, see section 5.1.1.

The person/non-person distinction is marked for the following parts-of-speech:

- Nouns: common (NN), proper (PN), meta nouns (MN), adjectival (AN) and verbal (VN)
- Pronouns (PO)
- Adjectives (AJ)
- Participles: present and perfect (SP/TP)
- Others: indefinite article (EN), numerals (RO)

The analysis found in the lexical layer ideally represents the type of information that is inherent for the word in question and hence non-contextual (Teleman 1974). For instance, the part-of-speech category of pronouns (PO)

does not distinguish determiners from nominal heads in the lexical layer.[112] With respect to the annotation for person reference, it is clear that the syntactic environment has been taken into account. Nouns are marked for personhood regardless of syntactic context, whereas pronouns, adjectives, participles, indefinite articles and numerals are marked for personhood only when they are heads in nominal phrases (Teleman 1974). (99)–(103) exemplify nominal elements with person annotation (HH) of various parts of speech – pronoun in (99), adjective in (100), participle in (101), indefinite article in (102) and numeral in (103):

(99) **De** *som tagits ut till underofficersutbildning* …
they who taken out to under-officer-education …
'Those who have been chosen for the subordinate officer education …'

(100) …*att flytta den* **unge** *från hemmet*
…to move the young from home-DEF
'…to move the young one from his/her home'

(101) *Antalet* **skadade** *var 140 000*
number injured was 140 000
'…The number of injured was 140 000'

(102) *En ropar det rytmiska ga-ga*
one calls the rythmical ga-ga
'One calls out the rythmical ga-ga'

(103) *År 1970 hade ungefär* **700 000** *förvärvsarbete*
year 1970 had approximately 700 000 gainful-employment
'In the year 1970 approximately 700 000 had gainful employment'

Even though the annotation manual clearly states that only when functioning as a nominal head should a pronoun, adjective or participle be annotated for personhood, we find examples where adjectives and participles functioning as genitive modifiers are annotated as persons:

(104) …*de* **försäkrades** *egna sjukförsäkringsavgifter*…
…the insured-GEN own health-insurance-fees…
'…the health insurance fees of the insured'

[112] A pronoun like *de* 'the-PL/they', for instance, is annotated with the part-of-speech PO regardless of whether it bears a nominal syntactic function, e.g. subject/object, or functions as a determiner, e.g. *de inkomster* 'the incomes'. These are distinguished only in the syntactic annotation in terms of dependency relation. See section 4.1 for more on pronouns in Scandinavian.

Since genitives have a reference which is independent of the nominal which it modifies, this decision seems reasonable.

As mentioned earlier, all pronominal heads are annotated for personhood. The relative pronoun or marker *som* 'who' is analysed as a core argument in the relative clause (either subject or object) in Talbanken05 and always "inherits" the animacy from the head argument which it modifies.

The manual treats collective nouns as non-persons, including examples like *personalen* 'staff-DEF', *polisen* 'police-DEF', *domarkåren* 'judge corps', *folket* 'people-DEF' (Teleman 1974). Animals are in general not treated as person referring, except in contexts where they are "anthromorphised" and may be referred to by the pronouns *han, hon* 'he, she' (Teleman 1974).

7.1.2.2 Person reference and animacy

In section 7.1.1 above we discussed the annotation scheme employed in Zaenen et al. 2004. There are clear similarities between the annotation for person reference found in Talbanken05 and the annotation for animacy. Regardless of annotation scheme, the person/non-person distinction can be viewed as forming the outer perimeters of the animacy dimension and, in this respect, the annotation schemes do not conflict. Following the above overview of the annotation found in Talbanken05, we may compare it with annotation schemes for animacy, in particular the ones found in Garretson et al. 2004, and hence Zaenen et al. 2004), as well as Yamamoto 1999. We find that the schemes differ primarily in the granularity of classes distinguished and the types of markables which are annotated:

- **Classes:** There is a partial overlap in classes between the person reference annotation in Talbanken05 and the approaches that explicitly annotate for the property of animacy. Garretson et al. 2004 contains the category Human, as well as Inanimate (at the top-level of annotation), which must be assumed to correspond to the person/non-person distinction. The main source of variation in class distinctions consists in the annotation of collective nouns, including organizations, as well as animals. Animals and organizations are treated as inanimate in the Talbanken05 scheme, whereas they form an intermediate category in Garretson et al. 2004. The Talbanken05 scheme is similar to Yamamoto 1999 in treating organizations as inanimate, but differs in not providing a more detailed treatment for animals.

- **Markables:** Talbanken05 annotates slightly more markables than Yamamoto (1999) in also annotating for adjectives and participles as nom-

inal heads. Like Yamamoto, the Talbanken05 includes relative, interrogative and indefinite pronouns. Garretson et al. 2004 is not comparable in this respect as it only annotates genitive constructions, and Zaenen et al. (2004) do not state explicitly what the exact markables of their study are.

We may conclude that the person/non-person distinction in Talbanken05 provides a valuable source of data on animacy. First of all, it makes the main distinction which is common to all approaches to animacy and animacy annotation – the distinction between human and inanimate. As the annotation study in 7.1.1 showed, organizations and animals are infrequent classes, hence we may assume that these do not disrupt generalizations regarding the class of inanimates in any significant way. Second, the annotation in Talbanken05 provides information regarding a wide range of markables including common and proper nouns, as well as pronouns.

7.1.2.3 The distribution of person reference in Talbanken05

In chapter 6 we examined distinctions in the distribution of animacy with respect to a set of theoretically motivated morphosyntactic features. In this section we approach this matter empirically and examine the general distribution of person versus non-person referring nominals in Talbanken05. We focus largely on syntactic distribution and examine distinctions within the groups of argument, as well as non-argument, relations.

A note on counts

As explained above, for several parts-of-speech personhood is expressed only when these function as nominal heads. When comparing the distributions of persons and non-persons for parts-of-speech other than nouns, our population should hence only consist of nominal heads. However, ascertaining when a pronoun or an adjective is head of a nominal phrase is not completely straightforward in a dependency annotation where there is no direct concept of phrases. In the following section, we approximate the notion of nominal head to head of a *nominal dependency relation* and define these to be the argument functions defined in section 5.1.1.[113] As mentioned earlier, person reference is also rel-

[113]This is admittedly a simplification, nominal elements may certainly also have other functions, however, it is fair to assume that the argument functions are the functions which are predominantly nominal.

Part-of-speech		Person #
N	noun	7066
PO	pronoun	9809
AJ	adjective	280
P	participle	57
R	numeral	33
EN	indef. pronoun	12
Total		17257

Table 7.2: Total number of tokens annotated as persons in the written sections of Talbanken05, broken down by part-of-speech.

evant for genitival modifiers since these have an independent reference, hence we include these also in our overview.

Overview

Table 7.2 shows the absolute number of tokens annotated as person in the written sections of Talbanken05, broken down by part-of-speech. We find that person reference is most common for pronouns and nouns, which together account for 97.8% of the total person instances.

Table 7.3 shows the distribution of person/non-person over nominal heads, also broken down by part-of-speech. In general, without discerning individual NP-types, we see that non-persons are more frequent in the corpus than persons. It is also clear, however, that the personhood or animacy dimension influences referentiality and more specifically, the part-of-speech employed. As mentioned in section 3.4, persons are often referred to by a pronoun and we find that the percentage of persons is high for pronominal arguments (51.8%).

Table 7.4 presents the distribution of person and non-person nouns and pronouns across various dependency relations in Talbanken, see table 5.3 in section 5.1.1.[114][115] There are some clear tendencies towards differences in distribution between the two categories (person/non-person) and we can ascertain that person and non-person referring nouns and pronouns differ significantly in their general syntactic distribution ($p<.0000, df = 19$).[116]

[114]By limiting the overview to nouns and pronouns we ensure a comparison of nominal functions where person reference is possible. For instance, adjectival predicatives are not nominal and referential. Also, clausal complements are annotated as objects, however are not nominal and referential and should not be employed to compare the distribution of person vs. non-person in direct objects.

[115]Table 7.4 includes dependency relations which have more than 10 person instances.

[116]Pearson's Chi-Squared test with Yates' continuity correction with 19 degrees of freedom over a 2x20 matrix with rows=dependency relations and columns=person,non-person.

	Person		Non-Person		Total	
	#	%	#	%	#	%
Noun	5187	15.4	28421	84.6	33608	100.0
Pronoun	7596	51.8	7067	48.2	14663	100.0
Adj	206	6.7	2871	93.3	3077	100.0
Part	39	5.5	665	94.5	704	100.0
Num	30	7.4	375	92.6	405	100.0
Indef pro	10	16.7	50	83.3	60	100.0

Table 7.3: Absolute (#) and relative frequencies (%) of person/non-person nominal heads in the written sections of Talbanken05, broken down by part-of-speech.

In section 3.1 we established a set of distinctions within the group of argument relations and argued that animacy is a dimension by which arguments are differentiated. We may now test empirically whether different types of arguments differ with respect to person reference. The argument relations for which we find person referring elements are the subject (SS), indirect and direct object (IO, OO), subject and object predicative (SP, OP), as well as the logical subject (ES) relations.[117] We find that indirect objects and subjects exhibit the highest percentages of person referring nominals: 87.5% and 44.8%, respectively.[118] The percentage of person referring direct objects is clearly lower (21.2%). We have noted several places that subjects and direct objects tend to differ with respect to animacy. Dahl and Fraurud (1996) show that person NPs are more likely to occur as subjects of a transitive clauses than non-persons. The counts for subjects in our case contains all subjects, not only subjects of transitive verbs, however, we clearly see the same trend. In the Talbanken05 data, we find that the person reference of subjects and direct objects differ significantly ($p<.0000$).[119] The core argument functions are subjects, objects and indirect objects and non-core are the rest of the argument functions, in this case: the group of predicative relations (SP, OP). We find that the core and non-core arguments also differ significantly with respect to the property of person

[117]The logical subject is a relation employed in conjunction with an expletive or formal subject and denotes for instances demoted agents in presentational constructions or impersonal passives. See section 8.3.1 for more on the argument distinctions expressed in Talbanken05.

[118]The high percentage of person referring formal objects is a result of a "quirk" of the annotation, where the reciprocal pronoun *själv* 'him/herself' has been annotated as a formal object in examples like *Jag tror själv att* … 'I, myself, think that …'

[119]Pearson's Chi-Squared test with Yates' continuity correction with 1 degree of freedom over a 2x2 matrix with rows=person,non-person and columns=binary argument distinctions; e.g. subject,object.

		Person		Non-person		Total	
		#	%	#	%	#	%
SS	subject	8385	44.8	10349	55.2	18734	100.0
PA	prep. compl.	2355	14.4	14035	85.6	16390	100.0
DT	determiner	2139	12.2	15330	87.8	17469	100.0
OO	dir. obj.	1963	21.2	7281	78.8	9244	100.0
CC	conjunct	733	17.9	3373	82.1	4106	100.0
IO	indir. obj.	365	87.3	53	12.7	418	100.0
SP	subj. pred.	235	11.3	1849	88.7	2084	100.0
AN	apposition	130	17.9	596	82.1	726	100.0
HD	head of idiom	121	26.8	330	73.2	451	100.0
ET	post-nom. mod.	86	27.6	226	72.4	312	100.0
ROOT	root	72	11.7	542	88.3	614	100.0
XX	unclass.	66	25.0	198	75.0	264	100.0
FO	formal obj.	64	44.4	80	55.6	144	100.0
ES	logical subj.	60	16.0	315	84.0	375	100.0
KA	comp. adv.	58	32.4	121	67.6	179	100.0
+F	coord. clause	41	19.7	167	80.3	208	100.0
AA	adv.	13	9.3	127	90.7	140	100.0
OA	obj. adv.	13	27.1	35	72.9	48	100.0
OP	obj. pred.	13	23.6	42	76.4	55	100.0

Table 7.4: Absolute (#) and relative frequencies (%) of nouns and pronouns annotated as persons and non-persons in the written sections of Talbanken05, broken down by dependency relation.

reference ($p<.0000$). The core arguments include the indirect object function. Whereas for the subject and object functions some variation is to be expected, indirect objects have been noted to exhibit a strong preference for animate realization (Bresnan et al. 2005) and one would expect non-persons to be virtually non-occurring in this relation. However, a closer look at the data indicate that animals, as in (105), collective nouns, as in (106), and organization nouns, as in (107), are in majority among the elements annotated as non-persons in indirect object position. There are also some clearly inanimate indirect objects, as in (108) and (109).

(105) …*att man ofta ger* **hunden** *ett mål mat per dag*
…that one often gives dog-DEF a portion food per day
'…that one often gives the dog one meal per day'

(106) *TV gav* **familjen** *en ny samlingspunkt*
tv gave family-DEF a new gatheringpoint
'Television provided the family with a point of union'

(107) *Forsmark kraftstation kommer att tillföra* **kommunen** *ett*
Forsmark powerstation will to supply municipality-DEF a
avsevärt tillskott
considerable increase
'Forsmark power station will supply the municipality with a considerable increase'

(108) *På gatsten och betong kan ett fulldubbat däck ge*
on cobble-stone and concrete can a studded tire give
bilen *helt livsfarliga egenskaper*
car-DEF totally life-dangerous properties
'A studded tire can, on cobble stone or concrete, provide the car with life-threatening properties'

(109) *Idag försöker man i regel ge det här* **argumentet** *en*
today tries on in rule give this here argument-DEF a
positiv formulering
positive expression
'Nowadays one usually tries to give this argument a positive expression'

The examples in (105)–(109) illustrate the flexibility of the ditransitive construction with respect to its possible arguments. The example in (105) is typical for what one might call the prototypical ditransitive 'giving'-situation, where an animate agent transfers a concrete object to another animate participant. In (106) the subject is not an agent but rather expresses a causing event (the acquiring of a television set). The examples in (108) and (109) also show a more abstract instantiation of the prototypical giving involving inanimate recipients and no sense of transfer at all. Differentiating properties of the arguments may vary with the sense of the dative verb in question, e.g. whether *give* is employed in a transfer sense, communication sense or abstract sense (Bresnan et al. 2005). However, we may also establish a general trend with respect to properties of the arguments such as animacy, in line with the findings for English.

We see from table 7.4 that there are a range of other, non-argument syntactic relations in which person referring nominals occur. In general, person reference is less frequent in these relations and we find that the argument and

non-argument relation differ significantly ($p<.0000$) with respect to person reference.

Even so, we do find a number of the person referring nominals in non-argument functions, most notably, functioning as determiners (DT) or prepositional complements (PA). The person referring determiners turn out to be almost exclusively genitive modifiers, hence corroborate the correlation between animacy and genitive expression studied in chapter 6 for Norwegian. The fact that these account for such a small proportion of the nominal determiners is that, as noted earlier, nominal pronouns and determiners are assigned the same part-of-speech in Talbanken05.

As table 7.4 indicates, prepositional complements (PA) show a clear preference for non-person reference. Although this tendency is clear in itself, a more detailed and informative picture emerges if we consider the type of prepositional head the nominals in question are governed by. Table 7.5 shows the percentage of person/non-person referring elements among the nominal prepositional complements in Talbanken05, broken down by the most frequent governing prepositions.[120]

We find that for a majority of the prepositions, there is a strong tendency for non-person complements and some of these take almost exclusively (100-99%) non-person complements: *vid* 'by/next-to', *före* 'before', *utanför* 'outside', *sedan* 'since', *i* 'in', *efter* 'after', *inom* 'inside', *under* 'under'. These are all prepositions which position their complement spatially or temporally. The converse situation is much more rare, only two prepositions, *mellan* 'among' and *hos* 'at' show a stronger tendency for person complements than non-person ones. The fact that some prepositions show a stronger preference for person-denoting complements is somewhat surprising. For instance, the preposition *hos* 'at somebody's' is typically used to position a person, but in the Talbanken data it has only 58.8% person complements. However, if we examine the data a little closer, we find that most of the complements are actually not typically inanimate even though they are annotated as non-persons. Of the complements of *hos* which are tagged as non-person, 61.5% denote animals, as in (110), and 23.1% organizations, as in (111). In fact, only 15.4% are actually inanimate, as in (112).

(110) *Liknande* *förhållanden* *finner* *man* *hos* *häckande* **måsar**
similar circumstances finds one at hatching seagulls
'One finds similar circumstances among hatching sea gulls'

(111) *Hos* **försäkringskassan** *finns* *särskild* *broschyr*
at insurance-company-DEF exists special brochure
'At the insurance company they have a special brochure'

[120]Table 7.5 presents only prepositions with an absolute frequency of more than 100 occurrences.

	Person		Non-person		Total	
Preposition	#	%	#	%	#	%
i 'in'	31	0.8	3804	99.2	3835	100.0
av 'by/of'	479	21.7	1732	78.3	2211	100.0
på 'on'	184	9.7	1714	90.3	1898	100.0
för 'for'	463	31.8	991	68.2	1454	100.0
med 'with '	346	26.0	984	74.0	1330	100.0
till 'to'	182	14.3	1091	85.7	1273	100.0
om 'about'	60	10.1	535	89.9	595	100.0
från 'from'	40	8.9	410	91.1	450	100.0
vid 'beside'	14	4.1	325	95.9	339	100.0
under 'under'	2	0.6	312	99.3	314	100.0
mellan 'between'	154	58.3	110	41.7	264	100.0
mot 'against'	48	20.6	185	79.4	233	100.0
inom 'within'	3	1.4	213	98.6	216	100.0
efter 'after'	3	1.4	212	98.6	215	100.0
enligt 'following'	50	23.3	165	76.7	215	100.0
genom 'through'	9	4.6	188	95.4	197	100.0
hos 'at/among'	90	58.8	63	41.2	153	100.0
utan 'without'	3	2.3	127	97.7	130	100.0
ur 'out-of'	4	3.3	118	96.7	122	100.0

Table 7.5: Absolute (#) and relative frequencies (%) of person and non-person noun complements for prepositions; ranked by total, absolute frequency in the written sections of Talbanken05.

(112) *Därför* *lägger* *man* *vikt* *vid* *andra* *egenskaper* *hos*
therefore lays one weight on other properties at
bilen
car-DEF
'This is why one emphasizes other properties of the car'

The above examples of indirect objects in (105)–(109) and prepositional complements in (110)–(112), illustrate the fact that the person/non-person distinction as such does not incorporate the more fine-grained distinctions often represented in an animacy hierarchy. The general tendency for PA dependency relation is thus a strong preference for non-person reference.

Other non-argument relations in which we find person referring nominals include the adverbial relations of comparative adverbials (KA), as in (113), and object adverbials (OA), with 32.4% and 27.1%, respectively. We also find person referring elements (17.9%) among the appositions (AN), as in (114).

(113) *Utlänningar betalar skatt som* **svenskar**
foreigners pay taxes as Swedes
'Foreigners pay taxes just like Swedes'

(114) *Föräldrarna, särskilt* **fäderna**, *ägnar alltför* ...
parents-DEF, especially fathers-DEF, devote too ...
'The parents, especially the fathers, devote too much ...'

To summarize, we find that the distribution of person/non-person in the Swedish data instantiates the general pattern of animacy with respect to argument differentiation, as discussed in chapter 3. We also find that person referring nominals are not limited to the argument relations, but also occur as nominal heads of some non-argument relations.

Annotation consistency

In light of our earlier discussion on issues in annotation for animacy, it might be interesting to examine a few properties of the annotation a bit closer. With respect to annotation of nouns, we may differentiate between a purely *denotational* (type level) annotation strategy and a purely *referential* (token level) one. A denotational strategy entails that an element consistently be assigned to only one class. A referential strategy, in contrast, does not impose this restriction on the annotation, hence class assignment may vary depending on the specific context. The brief instruction given in the annotation manual for Talbanken05 (Teleman 1974: 223) gives leeway for interpretation in the annotation.

With the general aim of obtaining animacy data for supervised animacy classification, an extraction of person information from Talbanken at the level of noun lemmas will clearly be problematic if there is a lot of variation in class assignment at the level of tokens. We may thus examine the intersection of the two classes for noun lemmas in the written sections of Talbanken, i.e. the set of nouns which have been assigned both classes. It contains 82 noun lemmas, which corresponds to 1.1% of the total number of noun lemmas in Talbanken (7554 lemmas all together). This is clearly such a small proportion that it should not be problematic to employ the annotation at the lemma level. After an inspection of the intersective elements, we may group the nouns which were assigned to both classes, roughly into the following categories:[121]

Abstract nouns These are nouns with underspecified or vague denotational (type-level) properties with respect to animacy, such as quantifying nouns

[121] Recall that 'HH' is the tag for person referring, whereas the lack of such a tag, '_', denotes a non-person referring element.

whose reference is determined mostly by context, e.g. *hälft* 'half', *miljon* 'million', *nästa* 'next', as well as other nouns which may be employed with varying animacy, e.g. *element* 'element', *fiende* 'enemy', *part* 'party', as in (115) and (116):

(115) *men det förutsätter att också den andra* **parten**$_{HH}$ *står utanför*
but that presupposes that also the other party-DEF stands outside
'but that presupposes that the other party is also left outside'

(116) *I ett förhållande är aldrig bägge* **parter**_ *lika starka*
in a relationship are never both parties same strong
'In a relationship, both parties are never equally strong'

We also find that nouns which denote abstract concepts regarding humans show variable annotation, e.g. *individ* 'individual', *adressat* 'addressee', *medlem* 'member', *kandidat* 'candidate', *representant* 'representative', *auktoritet* 'authority'

Reference shifting contexts These are nouns whose denotational animacy is quite clear but which are employed in a specific context which shifts their reference. Examples include metonymic usage of nouns, as in (117) and nouns occurring in dereferencing constructions, such as predicative constructions (118), titles (119) and idioms (120):

(117) *Trots* **daghemmens**$_{HH}$ *otillräckliga resurser* ...
despite kindergarten-DEF.GEN inadequate resources ...
'Despite the kindergarten's inadequate resources ...'

(118) ...*för att bli en bra* **soldat**_
...for to become a good soldier
'...in order to become a good soldier'

(119) ...*menar* **biskop**_ *Hellsten*
...thinks bishop Hellsten
'thinks bishop Hellsten'

(120) *ta* **studenten**_
take student-DEF
'graduate from highschool (lit. take the student)'

Annotation errors There is some variation in annotation which we suspect are annotation errors, e.g., (121)–(122) below. We also find instances that were assigned to the wrong lemma due to mistakes in lemmatization (e.g. *moder* 'mother' lemmatized to *mod* 'courage').

(121) *han får inte föra de direkta förhandlingarna på minst*
he may not lead the direct negotiation-DEF.PL in at-least
ett **halvår**$_{HH}$
one half-year
'he is not allowed to lead the direct negotiation for at least another half year'

(122) *Om djup disharmoni mellan* **föräldrarna_** *dessutom*
if deep disharmony between parents-DEF also
äventyrar barnens hälsa ...
adventure child-DEF.GEN health ...
'If a deep disharmony between the parents also jeopardizes the children's health ...'

It is interesting to note that the main variation in annotation stems precisely from difficulties in determining reference, either due to bleak denotational properties such as for the abstract nouns, or due to properties of the context, as in the reference shifting constructions.

7.2 Data preliminaries

This section presents the data sets employed in the scaled up classification experiments. We examine the set of nouns, as well as feature representation and feature extraction, which all constitute important elements in a supervised machine learning experiment.

7.2.1 Talbanken05 nouns

These data sets consist of the noun lemmas with corresponding class (person/non-person) extracted from the Talbanken05 material, detailed above.[122] Following the conclusions at the end of section 7.1.2, we here approximate the class of 'animate' to 'person' and the class of 'inanimate' to 'non-person'. Table 7.6 provides an overview of the data set resulting from extraction from

[122]The treebank was lemmatized prior to extraction (Kokkinakis 2001).

Class	Types	Tokens covered
Animate	644	6010
Inanimate	6910	34822
Total	7554	40832

Table 7.6: The animacy data set from Talbanken05; number of noun lemmas (Types) and tokens in each class.

Talbanken.[123] Intersective elements, see section 7.1.2, were assigned to their majority class.[124]

It is clear that the data is highly skewed towards the non-person class, which accounts for 91.5% of the data instances. We may also note that the type-token ratio differs somewhat for the two classes. Person nouns exhibit less lexical variation than non-person nouns; each person noun type occurs on average nine times, whereas the corresponding figure for non-person nouns is five.

7.2.2 Features

In chapter 6 we made use of a set of theoretically motivated, distributional features to represent various aspects of the syntactic properties of the nouns that were classified. In particular, we found that the features encoding subject, direct object and genitive were strong features for animacy classification. Whether or not these features are important also in the current setting remains to be tested empirically. There may also be other features which are important in the scaling to a new, larger set of nouns and a new, although closely related, language. We therefore construct a very general *feature space* for animacy classification, which makes use of distributional data regarding the general syntactic properties of a noun, as well as various morphological properties. It is clear that in order for a syntactic environment to be relevant for animacy classification it must be, at least potentially, nominal. We define the *nominal potential* of a dependency relation as the frequency with which it is realized by a nominal element (noun or pronoun) and determine empirically a threshold of 0.10. The syntactic and morphological features in the general feature space are presented below:

Syntactic features A feature for each dependency relation with nominal po-

[123]Note that the figures in table 7.6 differ from those presented in table 7.2 above, as the current data set only contains common nouns, not proper names.

[124]When there is no majority class, i.e. in the case of ties, the noun was removed from the data set. 12 lemmas were consequently removed from the data set.

tential: (transitive) subject (SUBJ)[125], object (OBJ), prepositional complement (PA), root (ROOT)[126], apposition (APP), conjunct (CC), determiner (DET), predicative (PRD), complement of comparative subjunction (UK). We also include a feature for the complement of a genitive modifier, the so-called 'possessee', (GENHD).

Morphological features A feature for each morphological distinction relevant for a noun: gender (NEU/UTR), number (SIN/PLU), definiteness (DEF/IND), case (NOM/GEN). Also, the part-of-speech tags distinguish dates (DAT) and quantifying nouns (SET), e.g. *del*, *rad* 'part, row', so these are also included as features.

7.2.3 Feature extraction

In chapter 6, the distributional data for the individual noun lemmas was extracted from a fairly large, automatically parsed corpus of Norwegian. For extraction of distributional data for the set of Swedish nouns we make use of the Swedish Parole corpus, see section 5.1.2. To facilitate feature extraction, we part-of-speech tag the corpus and parse it with the MaltParser, which assigns a dependency analysis.[127]

Table 7.7 shows an overview of the aggregated mean values, along with standard deviations, from the Parole corpus for each class of Talbanken05 noun (Animate or Inanimate) broken down by the various features. Despite the fact that these values are from a noisy, automatically annotated corpus, we observe many of the same tendencies as noted in the treebank material discussed earlier. We find clear distributional differences between the classes in a range of syntactic relations, most notably in argument positions (SUBJ, OBJ), as prepositional complement (PA) etc. For the extraction of the SUBJ and OBJ features in chapter 6, we took advantage of the containment of ambiguity which characterizes Constraint Grammar analysis, see section 5.1.3, and extracted only subjects and objects which were structurally unambiguous. The data extraction for Swedish is in this respect more noisy, since the dependency analysis

[125] An element is a transitive subject if it has a direct object sibling.

[126] Nominal elements may be assigned the root relation in sentence fragments which do not include a finite verb.

[127] For part-of-speech tagging, we employ the MaltTagger – a HMM part-of-speech tagger for Swedish (Hall 2003). The pretrained model for Swedish employs the SUC tagset (http://spraakbanken.gu.se/parole/tags.phtml). For parsing, we employ MaltParser, see section 5.3.1 with the pretrained model for Swedish, which has been trained on the SUC-tags output by the tagger. It makes use of a smaller set of dependency relations than those found in Talbanken05.

			Animate		Inanimate	
			Mean	SD	Mean	SD
Syntactic		SUBJ	0.21	0.12	0.08	0.07
		OBJ	0.13	0.07	0.19	0.13
		PA	0.21	0.10	0.40	0.18
		ROOT	0.03	0.03	0.03	0.05
		APP	0.03	0.03	0.01	0.03
		CC	0.12	0.08	0.09	0.08
		DET	0.12	0.16	0.04	0.09
		PRD	0.07	0.08	0.04	0.06
		UK	0.04	0.05	0.01	0.03
		GENHD	0.03	0.05	0.04	0.06
Morphological	*gender*	NEU	0.05	0.21	0.29	0.45
		UTR	0.95	0.21	0.71	0.45
	number	SIN	0.51	0.34	0.75	0.30
		PLU	0.48	0.34	0.24	0.29
	definiteness	DEF	0.34	0.24	0.33	0.25
		IND	0.66	0.24	0.66	0.25
	case	NOM	0.93	0.17	0.96	0.12
		GEN	0.07	0.17	0.03	0.09
	date	DAT	0.00	0.00	0.01	0.07
	set	SET	0.00	0.00	0.01	0.08

Table 7.7: Mean relative frequencies and standard deviation for each feature by class following feature extraction from Parole for nouns of absolute frequencies >10.

does not indicate structural ambiguity. It is interesting to note that the tendencies are still very similar despite the noise of the data.

With respect to morphological properties, we observe differences in distribution with respect to gender and number where animates have a stronger preference for non-neuter gender (UTR) than inanimate, and, conversely, inanimate nouns exhibit a stronger preference for neuter gender than animate nouns. With respect to number, we, somewhat surprisingly, note that there is a stronger preference for singular number for inanimate nouns than animate, and the converse with respect to plurality. However, these features exhibit a high degree of variation and we find that certain nouns which almost exclusively occur in singular or plural affect these aggregated results. For instance, abstract inanimate nouns like *död* 'death' or *ansvar* 'responsibility' occur exclusively in the singular. Recall, however, that the feature representations of the nouns consist

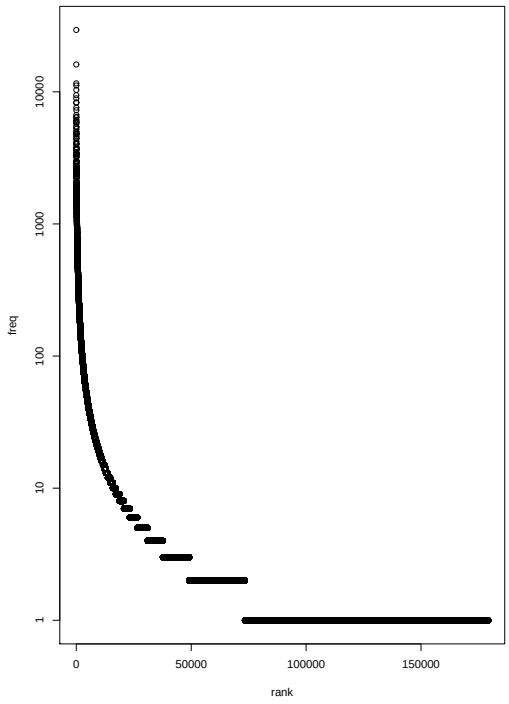

Figure 8: Rank frequency profile of all Parole nouns.

of normalized counts for that specific noun and not the aggregated means for each class as presented in table 7.7. Any lexical preferences with respect to morphology are thus properties of the individual nouns supplied to the classifier.

In chapter 6, we examined the effect of sparse data on classification. It is to be expected that the problem of sparse data becomes more severe as we attempt to scale up the animacy classification. The rank/frequency profile of common nouns in Parole is illustrated in figure 8.[128] It shows a Zipfian curve which is typical of word frequencies in natural language, where a few noun lemmas are highly frequent and an increasing number of lemmas have lower frequencies. The greatest number of lemmas, as illustrated by the "tail" in figure 8 occur only once, so-called *hapax legomena*.

In the experiments in chapter 6, we sorted the data into various frequency

[128] A rank/frequency profile illustrates the token frequencies of the ranked types. The frequency is plotted on a logarithmic scale, since there is such a large discrepancy between the token frequencies of the top ranked types, compared to the lower ranked ones (Baroni 2007).

		Animate		Inanimate		Total
Bin	Freq	#	%	#	%	#
∼1000	>1000	31	4.8	260	3.8	291
∼500	999-500	35	5.4	271	3.9	306
∼100	499-100	92	14.3	979	14.2	1071
∼50	99-50	57	8.9	553	8.0	610
∼10	49-10	132	20.5	1376	19.9	1508
∼1	9-1	132	20.5	1563	22.6	1695
0	0	165	25.6	1908	27.6	2073
Total		644	100.0	6910	100.0	7554

Table 7.8: Animate and inanimate Talbanken05 nouns in frequency bins by Parole frequency.

	Animate		Inanimate		Total	
Bin	#	%	#	%	#	%
>1000	31	10.7	260	89.3	291	100.0
>500	66	11.1	531	88.9	597	100.0
>100	158	9.5	1510	90.5	1668	100.0
>50	215	9.4	2063	90.6	2278	100.0
>10	347	9.2	3439	90.8	3786	100.0
>0	479	8.7	5002	91.3	5481	100.0

Table 7.9: Animate and inanimate Talbanken05 nouns in accumulated frequency bins by Parole frequency.

bins in order to examine the effect of sparse data on the classification performance. In table 7.8 we see the nouns from Talbanken05 organized into frequency bins by their absolute frequencies in the Parole corpus. For both animate and inanimate nouns, we find the same general tendency illustrated by the rank/frequency profile, indicated by an increasing number of types with lower frequencies. We observe that 30% of the Talbanken05 noun lemmas do not occur at all in the Parole corpus, hence will not be included in the data sets for classification.

Since the main focus of the current chapter is to scale up the animacy classification to realistic data sets, we mostly employ data sets consisting of *accumulated frequency bins*, which include all nouns with frequencies above a certain threshold. The data organized into accumulated frequency bins is presented in table 7.9.

7.3 Experiments

The experiments described in the following address some important issues in the scaling up of our method from chapter 6. In particular, we discuss how the skewed distribution of classes, as noted above, affects the results and examine the interaction with the additional, complicating factor of data sparseness. The overall focus will be on locating features which are stable class predictors across different machine learners and for data sets of varying properties.

7.3.1 Experimental methodology

In chapter 6, we compared eager and lazy machine learning algorithms for the task of animacy classification. We looked at the use of decision-trees acquired with the eager c4.5 algorithm (Quinlan 1993) and compared it with memory-based learning which employs lazy learning with the *k*-nearest neighbor algorithm. In section 6.5 we did not find any statistically significant differences between the two learning algorithms and conjectured that the size of the data set influenced the measure of significance conservatively. In the current chapter we have a considerably larger data set, hence we may once again compare performance between the two machine learning algorithms. In the following experiments we will continue to employ both types of learners and contrast the two wherever appropriate. For decision tree learning, we employ C5.0 with boosted classifiers, see section 5.2.1, and for memory-based learning we employ TiMBL, see section 5.2.2, with the *basic* settings, resulting from the parameter optimization described in section 6.5.2, unless otherwise stated.[129] For training and testing of the classifiers, we make use of leave-one-out cross-validation. The baseline represents assignment of the majority class (inanimate) to all nouns in the data set. Due to the skewed distribution of classes, as noted above, the baseline accuracy is very high, usually around 90%. Clearly, however, the class-based measures of precision and recall, as well as the combined F-score measure are more informative for these results. The baseline F-score for the animate class is thus 0.0%, and a main goal is to improve on the rate of true positives for animates, while limiting the trade-off in terms of performance for the majority class of inanimates, which start out with F-scores approaching 100. For calculation of the statistical significance of differences in the performance of classifiers tested on the same data set, McNemar's test (Dietterich 1998) is employed.

[129] Recall that the basic settings correspond to $k = 1$ with no feature weighting.

Bin	Baseline	MBL	DecTree
>1000	89.3	89.0	90.7
>500	88.9	90.3	93.3
>100	90.5	89.8	93.7
>50	90.6	89.4	93.3
>10	90.8	89.0	92.2
>0	91.3	90.0	92.1

Table 7.10: Accuracy for MBL and DecTree classifiers with the original feature set (SUBJ, OBJ, GEN) on Talbanken05 nouns in accumulated frequency bins.

7.3.2 Original features

The experiments on Norwegian in chapter 6 showed that the three features subject, object and genitive case were the most robust features. Table 7.10 shows the results from classification of the Talbanken05 nouns with the distributional features SUBJ, OBJ and GEN extracted from Parole, as described in sections 7.2.2–7.2.3 above. The experiments were run on accumulated frequency bins, where each data set contains all data instances of higher frequencies, e.g. the $>$ 50 data set contains all nouns of frequencies higher than 50.

We observe a clear difference between the results for the lazy (MBL) and eager (DecTree) machine learners. The performance of the MBL-classifier is never significantly better than the baseline and for the >100, >50, >10 and >0 data sets, the performance is in fact significantly worse than the baseline. The DecTree-classifier in contrast performs significantly better than baseline on all data sets of frequencies <1000.[130]

Tables 7.11 and 7.12 show the experimental results relative to class for the lazy and eager learner, respectively. For both classifiers, we find that the performance for the inanimate class is fairly stable, whereas the performance for the animate class deteriorates as more infrequent nouns are added to the data set. We find that the performance for the animate class is quite low (varying between top 66.1 and bottom 31.2), regardless of learner, and performance is clearly affected by the frequency of the data instances. If we compare the class results for the two learners, we find that the main difference is found in a better animate precision and inanimate recall for the DecTree-classifier. These are clearly advantageous properties in dealing with the skewed class distribution and counteracting overgeneralization from the less frequent class.

[130]The decision tree classifier does not differ significantly from the baseline for the >1000 data sets, but differs significantly from the baseline at the p<.001-level for the >500, >10 and >0 data sets, and at the p<.0001-level for the >100 and >50 data sets.

	MBL					
	Animate			Inanimate		
	Precision	Recall	Fscore	Precision	Recall	Fscore
>1000	48.5	51.6	50.0	94.2	93.5	93.8
>500	56.1	56.1	56.1	94.5	94.5	94.5
>100	46.3	47.5	46.9	94.5	94.2	94.4
>50	43.7	42.0	42.8	94.0	94.4	94.2
>10	38.8	34.3	36.4	93.4	94.5	94.0
>0	39.4	26.5	31.7	93.2	96.1	94.6

Table 7.11: Precision, recall and F-scores for the two classes in MBL-experiments with original features (SUBJ, OBJ, GEN).

	DecTree					
	Animate			Inanimate		
	Precision	Recall	Fscore	Precision	Recall	Fscore
>1000	57.1	51.6	54.2	94.3	95.4	94.8
>500	75.0	59.1	66.1	95.0	97.6	96.3
>100	79.8	44.9	57.5	94.5	98.8	96.6
>50	72.8	46.0	56.4	94.6	98.2	96.4
>10	67.6	28.2	39.8	93.2	98.6	95.8
>0	65.8	20.5	31.2	92.9	99.0	95.8

Table 7.12: Precision, recall and F-scores for the two classes in DecTree-experiments with original features (SUBJ, OBJ, GEN).

7.3.2.1 Uniform distribution

Based on the results from the experiments with original features, it is difficult to say anything about the general applicability of these features (SUBJ, OBJ, GEN) to the Swedish nouns. This is mainly due to the fact that the data exhibits a very skewed distribution of classes, hence training data for the animate class is limited. In order to test the generalizability of the original distributional features further and tease apart the influence of a skewed class distribution from that of data sparseness, an additional experiment is performed on data sets with a uniform distribution of classes.[131]

Table 7.13 shows the results for the uniformly distributed data sets and

[131]The uniform data sets are constructed from all animate instances in a data set and the corresponding proportion of randomly selected inanimate instances, balanced with respect to absolute frequencies. This technique for dealing with skewed data sets is known in the machine learning literature as 'down-sampling' and denotes the removal of instances of the majority class for training, see, e.g., Hoste 2005.

Bin	Baseline	MBL	DecTree
$>1000_{Uni}$	50.0	90.3	90.3
$>500_{Uni}$	50.0	88.6	83.3
$>100_{Uni}$	50.0	79.1	82.0
$>50_{Uni}$	50.0	78.6	83.3
$>10_{Uni}$	50.0	72.8	76.8
$>0_{Uni}$	50.0	68.6	72.2

Table 7.13: Accuracy for MBL and DecTree learners with the original feature set (SUBJ, OBJ, GEN) on Talbanken05 nouns with uniform class distribution in accumulated frequency bins.

	MBL		DecTree	
	Animate	Inanimate	Animate	Inanimate
$>1000_{Uni}$	90.3	90.3	90.0	90.6
$>500_{Uni}$	88.7	88.5	84.1	82.5
$>100_{Uni}$	78.8	79.4	81.7	82.2
$>50_{Uni}$	78.3	78.9	83.0	83.5
$>10_{Uni}$	72.8	72.7	73.3	79.5
$>0_{Uni}$	67.0	70.0	68.6	75.1

Table 7.14: F-scores for the two classes in experiments with original features (SUBJ, OBJ, GEN) and uniform class distribution.

contrasts eager and lazy learning, as before. We observe a clear reduction in error rate (80.6%-37.2%) for all classifiers compared to a random baseline. As the F-scores for each class in table 7.14 illustrate, the uniform distribution of classes gives balanced results for the individual classes as well. This shows that the set of motivated, robust features identified in the previous chapter are good class predictors also for Swedish and larger sets of naturally occurring nouns. It is also clear, as discussed in chapter 6, that data sparseness has a clear effect on the results, regardless of the class distribution. Results deteriorate gradually as more infrequent nouns are added, from accuracies of 90.3 for the $>1000_{Uni}$ data set to an average 70.4 for the $>0_{Uni}$ data set.

In the previous section we observed a difference in performance between the lazy and eager learner. In the present experiment, we find significant differences only for the data sets $>50_{Uni}$, $>10_{Uni}$ and $>0_{Uni}$.[132] This indicates that it is the ability to deal with data sparseness which is the main source of difference

[132]The performance of the decision tree classifier differs significantly from that of the MBL-learner at the $p<.05$ level for the $>50_{Uni}$ and $>10_{Uni}$ data sets, and at the $p<.01$-level for the $>0_{Uni}$ data set.

Bin	Baseline	MBL	DecTree
$\sim 1000_{Uni}$	50.0	90.3	90.3
$\sim 500_{Uni}$	50.0	88.6	88.6
$\sim 100_{Uni}$	50.0	79.1	81.6
$\sim 50_{Uni}$	50.0	78.6	79.8
$\sim 10_{Uni}$	50.0	72.8	71.2
$\sim 1_{Uni}$	50.0	68.6	60.0

Table 7.15: Accuracy for MBL and DecTree classifiers with the original feature set (SUBJ, OBJ, GEN) on Talbanken05 nouns with uniform class distribution in individual frequency bins.

between the two, rather than the skewed distribution of data. In chapter 6 we noted effects to the opposite, but, without enough noun instances to conclude on significant effects. We find that the data sparseness is better dealt with by the decision tree learner, given that there is sufficient data to generalize over. We must differentiate between the *size* of the data set and the *sparsity* of the data set. Table 7.15 illustrates this point further, showing results for experiments which are run on individual frequency bins, rather than accumulated ones. This provides an identical setting to the experiments on lower frequency nouns in chapter 6. These data sets are thus considerably smaller than their accumulated counterparts and once again, we find no significant differences between the two classifiers. The conclusion is therefore that decision trees perform better than MBL over sparse instances, given a larger data set than MBL. We must, however, note that the notion of similarity embodied in the MBL-settings is not updated to take into account a larger data set, which gives a somewhat unfair comparison, a point to which we return in the next section.

We may conclude from the above that both data sparseness and skewed class distribution are serious issues in the scaling up of our classification task. We find that the skewed distribution causes an unbalanced result for the lower frequency class of animate nouns. We also observe the general detrimental tendency of sparse data, regardless of class distribution and size of data set. We find that it may be partially counteracted by the size of the data set, however there is clearly room for improvement. In dealing with more infrequent nouns it is clear that the three features employed above do not provide sufficient class discrimination. In the following we will therefore investigate some strategies to obtain more informed learners. In particular, we examine an extended feature space, as well as optimizing the notion of similarity employed during classification.

7.3.3 General feature space

The general feature space described in 7.2 above gives more distributional data for each individual noun. This can be an advantage in the light of the skewed distribution and data sparseness discussed above, since it enables a more informed measure of similarity between the instances. However, whether the more general feature space capture generalizations which are relevant to the animacy dimension is a claim which has to be tested empirically.

Prior to the experiments, the TiMBL settings were optimized on a subset of the full data set, giving us a set of optimized MBL_{opt} classifiers.[133] The parameter optimization shows that a larger set of nearest neighbors, as well as feature weighting and weighted class voting provide for better generalizations over the data. All of these parameters contribute to a more discriminating notion of similarity which is an important factor in successfully exploiting the information contained in an enlarged feature space, as well as the earlier mentioned skewed distribution and data sparseness.

Table 7.16 shows the accuracy obtained with all features in the general feature space in terms of accuracy. We find significant improvements compared to the baseline for all data sets except the >0 data set, where performance for the unoptimized, lazy learner (MBL) is at baseline. The DecTree and MBL_{opt} classifiers are clearly superior to the unoptimized MBL classifier, hence will be focused on in the following.[134]

We observe a clear improvement with the general feature space compared to the baseline. The performance of the DecTree classifier on the >1000 data set is significant at the $p<.01$ level, whereas the MBL_{opt}-classifier differs at the $p<.001$ level on this same data set. Performance on all the other data sets show highly significant reduction of errors ($p<.0001$) for both classifiers. As we recall, the data sets are successively larger, hence it seems fair to conclude that the size of the data set partially counteracts the lower frequency of the test nouns. It is not surprising, however, that a method based on distributional features suffers when the absolute frequencies approach 1. Tables 7.17-7.18 present the experimental results relative to class. We find that, as noted earlier in chapter 6, it is largely the animate class which suffers from the addition of lower frequency nouns. Even so, the classification of animate instances is

[133] For parameter optimization we employ the paramsearch tool, supplied with TiMBL, see http://ilk.uvt.nl/software.html. paramsearch implements a hill climbing search for the optimal settings on iteratively larger parts of the supplied data. We performed parameter optimization on 20% of the total >0 data set, where we balanced the data with respect to frequency, concatenating equal proportions from each respective frequency bin. The resulting settings are $k = 11$, GainRatio feature weighting and Inverse Linear (IL) class voting weights.

[134] Differences between the two MBL-classifiers with general features are significant for all data sets.

Bin	Baseline	MBL	DecTree	MBL$_{opt}$
>1000	89.3	96.2	94.5	97.3
>500	88.9	95.1	96.1	97.3
>100	90.5	95.6	96.6	96.8
>50	90.6	94.8	95.7	96.1
>10	90.8	93.1	94.6	95.4
>0	91.3	91.9	93.9	93.9

Table 7.16: Accuracy for MBL and DecTree classifiers with a general feature space on Talbanken05 nouns in accumulated frequency bins.

	DecTree					
	Animate			Inanimate		
	Precision	Recall	Fscore	Precision	Recall	Fscore
>1000	82.6	61.3	70.4	95.5	98.5	97.0
>500	86.4	77.3	81.6	97.2	98.5	97.8
>100	89.1	72.8	80.1	97.2	99.1	98.1
>50	87.3	64.2	74.0	96.4	99.0	97.7
>10	76.8	59.1	66.8	96.0	98.2	97.1
>0	79.8	40.5	53.7	94.6	99.0	96.7

Table 7.17: Precision, recall and F-scores for the two classes in DecTree-experiments with a general feature space.

	MBL$_{opt}$					
	Animate			Inanimate		
	Precision	Recall	Fscore	Precision	Recall	Fscore
>1000	89.7	83.9	86.7	98.1	98.8	98.5
>500	89.1	86.4	87.7	98.3	98.7	98.5
>100	87.7	76.6	81.8	97.6	98.9	98.2
>50	85.8	70.2	77.2	97.0	98.9	97.9
>10	81.9	64.0	71.8	96.4	98.6	97.5
>0	75.7	44.9	56.4	94.9	98.6	96.7

Table 7.18: Precision, recall and F-scores for the two classes in MBL$_{opt}$-experiments with a general feature space.

notably improved compared to the experiment with original features. We also find that the performance for the inanimate class is quite stable throughout the experiments (ranging from 98.5-96.7), a fact which is important since these are in clear majority in the data set.

The MBL$_{opt}$-classifier consistently performs slightly better than the Dec-

Bin	Baseline	MBL-Morph$_{opt}$	MBL-Syntax$_{opt}$
>1000	89.3	89.7	94.5
>500	88.9	90.5	95.1
>100	90.5	91.5	95.5
>50	90.6	91.3	95.0
>10	90.8	91.1	94.0
>0	91.3	91.5	93.2

Table 7.19: Accuracy for MBL$_{opt}$-classifiers with feature subsets on Talbanken05 nouns in accumulated frequency bins.

Tree classifier, although only differences for the >1000 and >10 data sets are significant.

7.3.4 Feature analysis

Unlike the experiments reported in chapter 6, the features employed for representation of the nouns in the general feature space are not all linguistically motivated indicators of an animacy distinction. The above experiments, however, indicate that these features provide important clues for the animacy of nouns. In the following we analyze the influence of the various features from different perspectives.

7.3.4.1 Feature subsets – syntax vs. morphology

The general feature space consists of both syntactic and morphological features and the above experiments have indicated the importance of both of these feature types. The extent to which it is morphology or syntax which is most important in ascertaining animacy, however, is not clear. One way of contrasting the importance of syntactic and morphological distribution in determining animacy, is to run classification experiments with feature subsets of syntactic and morphological features.

In order to test the influence of syntactic versus morphological features we trained and optimized MBL classifiers for each of these feature subsets, as defined in section 7.2.2 and summarized in table 7.7 above. The overall results in terms of accuracy are presented in table 7.19.

The results clearly indicate that the syntactic features are the strongest indicators of animacy. The classifiers employing only morphological features perform around baseline or slightly above ($p<.05$) for the >100–>0 data sets. It

is clear that the increased size of these data sets enable the acquisition of generalizations regarding morphological clues for animacy, but these are clearly not sufficient. The classifiers employing syntactic features perform notably better on their own, with all differences from the baseline significant ($p<.0001$).[135] Even so, the necessity of both types of features is also corroborated by the results – the syntactic classifiers never outperform the classifiers combining morphological and syntactic evidence, all of which perform significantly better for all data sets.

7.3.4.2 Decision tree

An advantage of decision tree learning is that the result of learning provides a generalization over the data set which may be inspected. A decision tree consists of a set of weighted, disjunctive tests which at each node in the tree assigns an appropriate test to an input, and which proceeds along one of its branches, representing possible outcomes of the test. All features are usually not employed in the tree, since smaller trees are preferred and the tree is pruned prior to application, see section 5.2.1 for more details. As an indicator of feature importance we may therefore examine the decision trees in a bit more detail.

Figure 9 presents the decision tree constructed for the >100 data set.[136] The disjunctive tests applied at each step are of the form $\langle attr\ Test\ value \rangle$, where *attr* is a feature, *value* is a possible value of that feature and *Test* is the test operator. In this case all values are numerical and the operators are the binary numeric operators `<`, `=` and `>`. Each terminal node of the tree represents an assigned class, and information regarding the correct/incorrect ratio of instances covered by that particular node is provided in the example tree.

The decision tree in figure 9 embodies generalizations observed several places above. We find that the subject feature partitions the data set initially, with a cut-off of approximately 0.14. In fact, all the decision trees for the various accumulated data sets employ the subject feature for initial partitioning. The largest branch (lines 1-13) is characterized by instances with lower proportions of subject occurrences and is dominated by inanimate terminals. In the same branch, a higher proportion of objects is employed to ascertain the inanimate class (line 5) and vice versa (line 6). We noted earlier on the distribu-

[135] To be precise, the difference from the baseline for the >1000 data set is at the $p<.05$ level, most likely due to the small size of the data set.

[136] The decision tree in figure 9 was constructed over the entire data set and is in this respect an idealization. Minor variations of this tree were actually employed under the experiments, since we applied leave-one-out cross validation.

```
1   subj <= 0.1374663:
2   :...app > 0.04225352:
3   :   :...prd <= 0.06315789: inan (19)
4   :   :   prd > 0.06315789:
5   :   :   :...obj <= 0.2116788: anim (10/1)
6   :   :       obj > 0.2116788: inan (5)
7   :   app <= 0.04225352:
8   :   :...cc <= 0.1607717: inan (1176/7)
9   :       cc > 0.1607717:
10  :       :...subj <= 0.1105528: inan (108/4)
11  :           subj > 0.1105528:
12  :           :...ind <= 0.6509434: inan (3)
13  :               ind > 0.6509434: anim (5)
14  subj > 0.1374663:
15  :...uk <= 0.008849557: inan (135/8)
16      uk > 0.008849557:
17      :...prep > 0.3342618:
18          :...sin <= 0.1395349: anim (3)
19          :   sin > 0.1395349: inan (43/1)
20          prep <= 0.3342618:
21          :...nom > 0.9925373: inan (18/4)
22              nom <= 0.9925373:
23              :...app > 0.02214452: anim (34)
24                  app <= 0.02214452:
25                  :...subj > 0.221519:
26                      :...uk > 0.01548673: anim (47)
27                      :   uk <= 0.01548673:
28                      :   :...neu <= 0.0001496558: anim (12/3)
29                      :       neu > 0.0001496558: inan (4)
30                      subj <= 0.221519:
31                      :...root > 0.04166667: inan (4)
32                          root <= 0.04166667:
33                          :...nom <= 0.9231928:
34                              :...sin <= 0.1753731: anim (2)
35                              :   sin > 0.1753731: inan (8)
36                              nom > 0.9231928:
37                              :...app <= 0.02048417: anim (29/4)
38                                  app > 0.02048417: inan (3)
```

Figure 9: Decision tree acquired for the >100 data set in experiments with a general feature space.

tional asymmetry with respect to prepositional complementation and we find that this generalization is also represented in the decision tree, where we find the majority of animate instances in the subtree dominated by a test for lower proportions of this construction (line 20). The earlier mentioned preference for genitive case is present through a restriction on the proportion of nominative occurrences (lines 21-38), which is mutually exclusive from the genitive.

We also find some predictive environments which have not been studied in detail earlier. This is partially due to the fact that the parse model employed to parse Parole makes use of a slightly different tag set than the one found

in Talbanken05, the subject of our study of animacy in section 7.1.2 above. For instance, the UK-tag is employed for predicative modifiers, as in (123) and (124), and the APP-tag is employed for appositions, as in (125), all taken from the Parole corpus:

(123) *Vi jobbar som* **barnflickor** *och* ...
we work as nannies and ...
'We work as nannies ...'

(124) *Han ser ut som en dansk* **murare**
he looks out like a danish mason
'He looks like a Danish mason'

(125) *Hannes Sköld,* **stiftaren**, *var knäckt* ...
Hannes Sköld, founder-DEF was devastated ...
'Hannes Sköld, the founder, was devastated ...'

We observe that a lower value for the UK feature directly determines the inanimate class (line 15) for a set of nouns with a higher proportion of subject occurrences, but which are still predominantly inanimate. The classification of these instances has an accuracy of 94%. If we examine the classified instances, we find predominantly non-concrete, inanimate nouns like *förslag* 'suggestions', *råd* 'advice', *studie* 'study', *utredning* 'investigation', as well as a group of collective and organization nouns (31.1% of the instances), such as *förening* 'association', *grupp* 'group', *kommun* 'municipality', *ledning* 'board', *personal* 'personell' etc. As we noted in chapter 6, these nouns have high proportions of subject occurrences but are in Talbanken05 annotated as non-person referring. It is clear, however, that the nouns classified by this node occur more seldom as predicational modifiers, a construction which semantically requires more concrete and individuated arguments.

In the decision trees, we observe a general tendency for syntactic features to appear higher in the tree, with morphological features occurring closer to the leaf nodes. In particular, the aforementioned SUBJ feature, as well as the features DET, UK, CC, PREP, APP and PRD are recurring features with high coverage in all the decision trees. The morphological feature representing singular number SIN occurs in all decision trees, although with less general coverage. This indicates that the syntactic features provide more general indications of animacy status, but that the morphological features provide the more fine-grained information which ultimately determines the class. Thus both types of features are needed, a result which the experiments clearly showed.

Backward feature selection:

- Generate a pool of features $PF = \{F_i\}$
- Initialize the set of removed features RF with the empty set, $RF = \{\}$
- **loop: for each** feature $F_i \in PF$
  - run a cross-validation on the training set without the features in RF and F_i
  - **if** improvement of accuracy: add F_i to RF
- **goto** loop until no more improvement

Figure 10: Algorithm for automatic feature selection with backward search.

7.3.4.3 Automatic feature selection

The general feature space was constructed by including features for all annotation relevant to nouns. The above experiments showed that extending the feature space proved beneficial to classification for all data sets. In order to prune the feature space for unnecessary features, we performed backward feature selection from the general set of features. Backward feature selection starts out with the whole feature set and successively removes features, testing for improvement of results at each step. The algorithm for automatic feature selection employing backward search is presented in figure 10 and has been adapted from the forward algorithm presented in Mihalcea 2002. We remove only on statistical significance of improvement. We perform automatic feature selection on the >0 data set and find that the accuracy of the classifier improves slightly, from 93.9 to 94.0, but significantly ($p < .05$), following feature selection. The small difference is caused by an improvement in the classification of the animate class, in particular in terms of precision which improves from 75.7 to 77.1.

The advantage of backward selection is that it also gives us information regarding the importance of each individual feature along the lines of the "all minus one" testing in section 6.3.2. Important features will cause a deterioration of results when removed. We find that the only features which cause statistically significant deterioration of results on removal are the syntactic features SUBJ ($p<.01$), OBJ ($p<.05$) and DET ($p<.001$). As we saw in table 7.7 above, there was a clear distributional difference between the class of animate

>10 nouns		
(a)	(b)	← classified as
222	125	(a) class animate
49	3390	(b) class inanimate

Table 7.20: Confusion matrix for the MBL_{opt} classifier a general feature space on the >10 data set on Talbanken05 nouns.

and inanimate with respect to the syntactic relation of determiner. It turns out that animate determiners are predominantly genitives, so these three features in fact embody little more than the subset of robust features established following chapter 6.

The removal of the GENHD feature is the only case in which we find a significant improvement of results, on which this feature is permanently removed from the feature pool. Moreover, we find that all of the morphological features cause small, but insignificant deteriorations of results, as do the syntactic features PREP, ROOT, APP and PRD.

7.3.5 Error analysis

In chapter 6, we examined a small set of nouns in more detail and the current chapter has largely dealt with quantitative analysis of performance results on the scaled up data sets. We found that the morphosyntactic features supported a more fine-grained notion of animacy and explored a three-way classification task. It might be interesting to examine the output from the scaled up classifier in a bit more detail, and, in particular, we may examine the errors. The error analysis examines the performance of the MBL_{opt}-classifier employing all features on the > 10 data set in order to abstract away from the most serious effects of data sparseness.

Table 7.20 shows a confusion matrix for the classification of the nouns. Recall from section 7.1.2 above that the person reference annotation of the Talbanken05 nouns distinguishes only the classes corresponding to 'human' and 'inanimate' along the animacy dimension. There is no intermediate notion of animacy or expression of gradience. An interesting question is whether this choice affects the results. If so, we would expect erroneously classified inanimate nouns to contain nouns of intermediate animacy, such as animals and organizations.

If we examine the errors for the inanimate class we indeed find evidence of gradience within this category. The errors contain a group of nouns referring

to animals and other living beings (bacteria, algae), as listed in (126), as well as one noun referring to an "intelligent machine", included in the intermediate animacy category (Zaenen et al. 2004). Collective nouns with human reference and organizations are also found among the errors, listed in (128). Both of these are more frequent among the errors (ANIM:18.4%; ORG:12.2%) than in the corpus sample studied in section 7.1 above (ANIM:1.8%; ORG:3.4%). We also find some nouns among the errors with human denotation, listed in (129). These are nouns which typically occur in dereferencing contexts, such as titles, e.g. *herr* 'mister', *biskop* 'bishop' and which were annotated as non-person referring by the human annotators.[137] Finally, a group of abstract, human-denoting nouns are also found among the errors, as listed in (130). In summary, we find that nouns with gradient animacy properties account for 53.1% of the errors for the inanimate class.

(126) Animals/living beings:
alg 'algae', *apa* 'monkey', *bakterie* 'bacteria', *björn* 'bear', *djur* 'animal', *fågel* 'bird', *fladdermöss* 'bat', *myra* 'ant', *mås* 'seagull', *parasit* 'parasite'

(127) Intelligent machines:
robot 'robot'

(128) Collective nouns, organizations:
myndighet 'authority', *nation* 'nation', *företagsledning* 'corporate-board', *personal* 'personell', *stiftelse* 'foundation', *idrottsklubb* 'sport-club'

(129) Human-denoting nouns:
biskop 'bishop', *herr* 'mister', *nationalist* 'nationalist', *tolk* 'interpreter'

(130) Abstract, human nouns:
förlorare 'loser', *huvudpart* 'main-party', *konkurrent* 'competitor', *majoritet* 'majority', *värd* 'host'

For the animate nouns which are misclassified we have, as noted above, the additional influence of distributional factors and data sparseness. It is therefore more difficult to find any clear patterns in the misclassified nouns. It is interesting to note, however, that there are several nouns which recur as errors in the experiments for both Norwegian and Swedish. Among the animate nouns, we find among the highly frequent, misclassified nouns the instances *barn* 'child' and *vän* 'friend' which recurred in the error analyses for the experiments in chapter 6.

[137] In fact, both of these showed variable annotation in the treebank and were assigned their majority class – inanimate – in the extraction of training data.

7.3.5.1 Human versus automatic annotation

In chapter 6, we investigated gradience of the animacy dimension as evidenced by distributional data for Norwegian. We examined the case of organization nouns in more detail and experimented with supervised and unsupervised learning of a more fine-grained animacy distinction. The annotation study performed initially in this chapter, applied an even more fine-grained annotation scheme for animacy to a Swedish corpus sample. We concluded that intermediate categories are infrequent and settled for a binary distinction in the ensuing classification experiments. Even so, the often noted gradience is evident both in human and automatic annotation.

The manual annotation for person reference in the Talbanken05 treebank showed inconsistencies for certain instances. We found that these were a result of difficulties in ascertaining denotation and/or reference for the noun in question. For instance, nouns with bleak denotationalac properties, such as *element* 'element', *part* 'party', were assigned varying annotation by the human annotators. We also found examples like *soldat* 'soldier' and *student* 'student', where denotational properties with respect to animacy are clear, however, where dereferencing properties of the context caused annotation inconsistency. The main distinction between the classes of human and inanimate was fairly straightforward to apply, however, and the annotation was consistent with respect to intermediate categories such as animals and organizations. The human annotators clearly did not have difficulties in assigning animals to the non-person category, as instructed. This is not surprising since these constitute a clearly defined category, separate from persons.

The experimental results show clear evidence of gradience. In the feature analysis, for instance, we noted that a group of organization nouns were classified under a separate node in the decision tree, which tested for properties compatible with both classes. The error analysis of this section has shown that the inanimate class does not easily incorporate animals on terms of linguistic distribution. It is interesting to note that both the human and automatic annotation showed difficulties in ascertaining class for a group of abstract, human-denoting nouns, like *individ* 'individual', *motståndare* 'opponent', *kandidat* 'candidate', *representant* 'representative'. These were all assigned to the animate majority class during extraction, but were misclassified as inanimate during classification.

Comparing human and automatic annotation we find that these elucidate different properties of the animacy dimension. If we contrast the type of gradience found in the human and the automatic annotation, we may note some differences. The automatic classification deals purely with animacy as a *linguistic* category; i.e. animacy as evidenced in linguistic use. It also per defin-

intion treats animacy as a denotational category since the data representation abstracts over individual contexts of usage. Human annotators clearly have available world knowledge, and in particular also the animacy categories as *ontological* categories (Dahl 2008). As we saw above, animals constitute a fairly clearly delimited ontological category and is not confused with the category of humans, regardless of their linguistic behaviour. Furthermore, the task of human annotation differs from the automatic classification task in that the annotation is token-level and the influence of specific linguistic contexts clearly influences the annotation.

Human and automatic annotation also show a great deal of overlap in the treatment of animacy. The fact that we may, through machine learning based on distributional data from language use, replicate the annotation fairly successfully shows that animacy is largely a denotational property of nouns and we find that the animacy of a noun influences its linguistic distribution consistently and over large amounts of data.

7.4 Summary of main results

This chapter has discussed the scaling up of animacy classification. In the following, we address the questions posed initially in this chapter.

A prerequisite for supervised learning of animacy information is an annotated set of instances. We investigated referential and denotational approaches to *animacy annotation* through an annotation study performed by the author, as well as a corpus study of the annotation for person reference found in the Swedish treebank Talbanken05. These highlight problematic constructions for both types of approaches. In particular, we find that dereferencing constructions are problematic for referential approaches, whereas elements with vague or abstract denotational properties are problematic under a denotational approach. We conclude that a denotational approach is to be preferred for lexical acquisition of animacy information based on distributional evidence and that the material in Talbanken05 largely follows a denotational practice, hence, is well suited as training data. We also conclude that the dimension of person reference largely overlaps with animacy, and may be employed to approximate animacy.

In chapter 6 we developed a set of motivated features for animacy classification and we showed that a subset of these proved to be reasonably robust to data sparseness. A question posed initially concerns whether a *transfer of the method* to Swedish and a different data set is viable. The experimental results in section 7.3.2 indicate that this is indeed the case. By abstracting away from the skewed distribution of the data, as well as data sparseness, we showed that

the robust features SUBJ, OBJ and GEN provide comparable results to those obtained for Norwegian. The features proved to differentiate between the two classes well, resulting in balanced class results around 90% accuracy and F-scores.

Two main obstacles have been identified in the scaling of the animacy classification task: *data sparseness* and a *skewed class distribution*. As noted earlier in chapter 6, data sparseness is bound to be a problem for any method relying on distributional data, so also in the current chapter. We may conclude that these two factors are independent factors, but clearly also interact. We found that a skewed class distribution causes unbalanced class results for non-sparse data (in the $>$1000 experiments), and we found that data sparseness had detrimental effects on performance for non-skewed data sets, in the experiments with uniform class distribution. An advantage under the present setting is that we have available a notably larger data set. A key question therefore concerns how properties of the data representation, as well as learner properties should be defined in order to fully capture the information contained in the data and thereby alleviate some of the problems caused by the density and distributional properties of the data set.

A general feature space was constructed which took into account both morphological and syntactic evidence. *Feature importance* in classification was analyzed both experimentally, through classification with feature subsets and automatic feature selection, as well as manually, through the manual inspection of decision trees. Whereas the syntactic features were clearly most important, the morphological features provided useful clues, resulting in a combined effect in terms of performance. We obtain results for animacy classification, ranging from 97.3% accuracy to 94.0% depending on the sparsity of the data. With an absolute frequency threshold of 10, we obtain an accuracy of 95.4%, which constitutes a 50% reduction of error rate. With respect to class, we find that classification of the inanimate class is quite stable throughout the experiments, whereas the classification of the minority class of animate nouns suffers from sparse data. It is an important point, however, that it is largely recall for the animate class which goes down with increased sparseness, whereas precision remains quite stable. All of these properties are clearly advantageous in the application to realistic data sets, where a more conservative classifier is to be preferred.

An initial comparison between eager and lazy *machine learning algorithms* highlighted the need for a more discriminating notion of similarity in vector space for the memory-based learner. A parameter optimization stage was therefore introduced, which gave significant improvements in combination with a general feature space. With optimized lazy learners, we found no striking differences between the two learning algorithms. In general, it seems that the

property of data generalization prior to classification is an advantage of the eager learner, given enough data to generalize over. Another advantage is the possibility for manual inspection of the decision trees, a feature which may be exploited in feature and error analysis. Even so, the optimized, memory-based learner in general performed slightly better than the decision trees, however, with few differences being significant. We may therefore conclude that both types of learning algorithms are well suited for animacy classification.

Both in the preceding and current chapters we have expressed the underlying goal of elucidating properties of the animacy dimension and, in particular, the way in which it influences the linguistic distribution of nouns. This has been accomplished through corpus studies and experimental studies where a focus has been on feature and error analysis. One question with respect to animacy has been its *gradience*, i.e. whether the animacy dimension is a strictly binary one – animate and inanimate – or whether there are elements which have properties of both polarities. This question was addressed above with data both from human annotation and experimental results. Under the assumption of a binary animacy opposition, we showed how annotation inconsistency as well as classification errors provide different perspectives on the gradience of animacy. The fact that the human annotation classified tokens highlighted the influence of the linguistic context on classification, and the problems identified there were largely caused by elements which were denotationally or referentially variable. In the case of the automatic classification, a different picture emerges since the task in this case is to abstract over the totality of contexts for a particular noun. With no world-knowledge available, the automatic annotation deals strictly with animacy as a linguistic category. In the error analysis, we found that this approach causes the proposed, intermediate categories along the animacy dimension to emerge.

Part III

Parsing

8 ARGUMENT DISAMBIGUATION IN DATA-DRIVEN DEPENDENCY PARSING

Arguments tend to differ along a range of linguistic dimensions. We have examined one of these dimensions, namely animacy, in detail in part II and we found that statistical tendencies in syntactic realization proved to be reliable indicators of animacy. We have discussed the notion of soft, probabilistic constraints and presented evidence from a range of languages, and across various linguistic subdisciplines, suggesting that argument differentiation is influenced by these types of constraints. In syntactic parsing, argument status is assigned automatically employing various types of information, such as part-of-speech, lexical form etc. Parsing is thus a practical task where argument differentiation is put actively to use and it provides us with a setting where we may study the influence of various types of information in a set of controlled experiments.

This chapter introduces data-driven dependency parsing and motivates its choice in the current context. We present a methodology for error analysis of parse results and apply this methodology to the results for Swedish from MaltParser, a data-driven dependency parser. The error analysis sets the scene for the experiments presented in chapter 9, where we investigate the influence of a range of different linguistic properties on argument disambiguation.

8.1 Syntactic parsing

Whereas parsing in the general sense provides an interesting task and a controlled testing ground for argument differentiation, the types of generalizations which may be arrived at are clearly influenced by properties of the parser. For instance, in a grammar-driven parser, the grammar strictly defines the set of possible output strings and the grammar formalism chosen will also influence the analysis and possibly also the general expressivity.

The parser to be employed in these experiments should be compatible with the general assumptions established in chapters 2 and 3. We want our parser to deal with natural language and to be robust in assigning an analysis to any input string. Implicit in such a choice is a view of grammaticality as not being strictly defined by a grammar. We have discussed the central role of frequency, both from the point of view of linguistic constraints in argument differentiation and in terms of the modeling of these constraints as soft in the probabilistic sense. In a *data-driven* parser, parsing is per definition guided by frequencies in language use. This allows us to make as few assumptions as possible with respect to formulations of constraints on arguments, as well as their interaction, in terms of a grammar. Even so, it is clear that the nature of the data on the basis of which parsing is approximated directly determines the analyses constructed. With respect to arguments and argument differentiation, we have tried to make as few theoretical assumptions as possible. In particular, we do not want to commit to a structural definition of argument status. Rather, a view of grammatical functions as primitive notions, hence separated from surface linguistic properties such as linear precedence and morphological realization, enables investigations into mismatches between levels of linguistic analysis. In *dependency analysis* functional argument structure is separated from structural positioning and formulated as dependency relations. Structural assumptions are furthermore stripped down to the minimal relation between a head and its dependent, highlighting the link to semantic interpretation.[138]

8.1.1 Data-driven parsing

A distinction is often made between grammar-driven and data-driven parsing, where the former is characterized by a generative grammar which defines the language under analysis and the latter is not (Carroll 2000). This distinction has, however, become less clear-cut due to the extensive use of empirical methods in the field in recent years. Most current parsers are data-driven in the sense that they employ frequencies from language data to induce information to improve parsing. Data-driven parsing may thus be characterized, first and

[138] It is interesting to note that a recent line of investigation in syntactic parsing with phrase-structural representations has focused on the task of *function-labeling* (Blaheta and Charniak 2000; Merlo and Musillo 2005), where syntactic function labels are assigned to enrich the phrase-structure trees either in a separate post-processing stage or as an integral part of parsing. The more direct link to semantic analysis is cited as a main motivation for this task (Merlo and Musillo 2005). Although English has been the main language under study thus far, work on function labeling for Spanish highlights the particular importance of this type of information in dealing with languages that are less configurational than English (Chrupała and van Genabith 2006).

foremost, by the use of inductive inference, rather than by the use or dispension of a grammar in the traditional sense (Nivre 2006). The development is also a temporal one, where the early parsers consisting solely of hand-crafted grammars expressed as rules, observably ran into serious difficulties resulting from ambiguity in natural language, see section 2.3.1.

As a step towards disambiguation, statistical models have been widely employed in parsing. These models assign probabilities to syntactic structure by decomposing syntactic representations and defining probability distributions over these. Statistical models may thus be employed for parse selection following purely grammar-driven approaches, since these assign probability scores to the analyses returned by the grammar. Grammars may also be extended stochastically to produce probabilistic versions. The probabilistic extensions of context-free grammars (PCFGs) (Charniak 1996), for instance, define probability distributions over non-terminal nodes, where the probability of a syntactic analysis is simply the product of all its subtrees. Most statistical parsing models can be viewed as *history-based*: they decompose the parse tree into a set of parse *decisions* associated with a certain probability. The particular decomposition chosen is an important component in defining statistical parsing models (Collins 1999). PCFGs display some well-known weaknesses resulting from precisely the independence assumptions made in the statistical model, where the application of a phrase-structural rule depends only on the local subtree to which it applies, disregarding the larger structural context, as well as any lexical dependencies which may hold between elements lower in the tree. Ensuing work has focused on lexicalization of PCFGs (Collins 1996; Charniak 1997, 2000), as well as alternative decompositions to increase context-sensitivity (Collins 1999; Johnson 1998; Klein and Manning 2003).

The availability of treebanks has been crucial to the development of data-driven parsing, supplying data for inductive inference in terms of estimation of parameters for statistical parse models or even for the induction of whole grammars, so-called *treebank grammars* (Charniak 1996). A system for data-driven parsing of a language L may be defined by three components (Nivre 2006: 27):

1. A formal model M defining permissible analyses for sentences in L.
2. A sample of text $T_t = (x_1, \ldots, x_n)$ from L, with or without the correct analyses $A_t = (y_1, \ldots, y_n)$.
3. An inductive inference scheme I defining actual analyses for the sentences of any text $T = (x_1, \ldots, x_n)$ in L, relative to M and T_t (and possibly A_t).

As we have seen above, the formal model *M* may consist of a hand-crafted or induced grammar. The inductive inference scheme may consist simply in maximum likelihood estimation from a corpus, as in the case of PCFGs discussed above.

In strictly data-driven approaches, a grammar, whether hand-crafted or induced, does not figure at all. Hence, the formal model *M* is not a grammar and the sample of text T_t is a treebank containing the correct analyses with respect to *M*, which constitutes the training data for the inductive inference scheme *I*. Parsing in this respect does not rely on a definition of the language under analysis independently of the input data. Without a formal grammar, data-driven models condition on a rich context in the search for the most probable analysis, hence are clearly history-based. Magerman (1995) describes an early, purely data-driven parser for English which decomposes phrase-structural trees into a set of *features* (lexical form, part-of-speech tag, structural position etc.) and employs decision trees to score individual decisions during parsing. Collins (1999) shows that decomposition in terms of head-modifier dependencies produces a significantly more accurate parser.

The fact that parsing is unconstrained by a grammar gives a very large search space and there are various strategies for making search tractable. Typically, some sort of pruning of the search space is necessary to prevent computation of the probability of all possible parses. Deterministic processing constitutes another, very efficient strategy where the probability of each decision is maximized at each deterministic choice-point during the derivation.

8.1.2 Dependency parsing

The use of dependency representations, see section 5.1.1, in syntactic parsing has recently received extensive attention in the NLP community (Buchholz and Marsi 2006; Nivre et al. 2007). One of the arguments in favour of parsing with dependency representations is that dependency relations are much closer to the semantic relations which figure between words in a sentence. As automatic parsing often is viewed as a means to a semantic interpretation of a sentence, dependency analysis represents a step in the right direction.

We may define the task of dependency parsing informally as the mapping from natural language sentences to well-formed dependency structures. As before, this mapping may be defined explicitly and exhaustively by a grammar, and it may be data-driven to various extents, as discussed above.

8.1.3 Data-driven dependency parsing

In data-driven dependency parsing, the formal model *M* defining permissible analyses is given by a definition of a dependency graph – a labeled acyclic graph with certain properties, see for instance section 5.1.1 above with references therein. Decomposition of the dependency graph for induction is one point where approaches to data-driven dependency parsing differ and we may distinguish between *transition-based* approaches, where dependency graphs are decomposed into parse transitions (see, e.g., Yamada and Matsumoto 2003; Nivre, Hall and Nilsson 2004) and *graph-based* approaches, where dependency graphs are decomposed into subgraphs or individual dependency arcs, (see, e.g., McDonald et al. 2005).

MaltParser is a language-independent system for data-driven dependency parsing, which is based on a deterministic parsing strategy (Nivre 2003; Nivre, Hall and Nilsson 2004), in combination with treebank-induced classifiers for predicting parse transitions. It allows for explicit formulation of features employed during parsing by means of a feature model and is optimal with respect to incrementality.

8.1.3.1 Parsing strategy

The parsing strategy consists in a non-deterministic parsing algorithm which is made deterministic by a parse guide. The parsing algorithm is an adaptation of the shift-reduce algorithm for context-free phrase structure grammars for application to dependency graphs (Nivre 2003).

The parsing algorithm in MaltParser constructs parsing as a set of *transitions* between *parse configurations*. A parse configuration is a triple $\langle S, I, G \rangle$, where *S* represents the parse stack – a list of tokens which are candidates for dependency arcs, *I* is the queue of remaining input tokens, and *G* represents the dependency graph defined thus far (Nivre 2006). There are four possible transitions between parse configurations (where *top* is the token on top of the stack, and *next* is the next token in the input) (Nivre et al. 2006: 1):

SHIFT:	Push *next* onto the stack.
REDUCE:	Pop the stack.
RIGHT-ARC(r):	Add an arc labeled *r* from *top* to *next*; push *next* onto the stack.
LEFT-ARC(r):	Add an arc labeled *r* from *next* to *top*; pop the stack.

The parsing algorithm described above is clearly non-deterministic in allowing for several possible transitions out of most parse configurations. The choice to

make the parsing strategy deterministic is taken primarily on grounds of efficiency (Nivre 2006). However, we will see below that it also has the effect that the parser is incremental.

The parse *guide* predicts the next parse action (transition), based on the current parse configuration. The guide is trained employing discriminative machine learning, which recasts the learning problem as a classification problem: given a parse configuration, predict the next transition. Prediction thus relies on a decomposition of the gold standard data set into parse configurations and a feature model which defines the relevant attributes of the configuration for use by the classifier.[139]

8.1.3.2 Feature model

As mentioned above, the parse guide predicts the next parse action based on the current parse configuration.[140] The feature model defines the relevant attributes of tokens in a parse configuration. There are generally two main types of attributes – *static* and *dynamic*. Static attributes are constant and defined by the input to the parser, whereas dynamic attributes are updated during parsing. Examples of static attributes are *lexical form* and *part-of-speech*. The feature model in MaltParser also enables the use of dynamic attributes of the dependency graph under construction, in particular the *dependency relation*.

Parse configurations are represented by a set of features, which focus on attributes of *top*, *next* and neighboring tokens in the stack, input queue and dependency graph under construction. Figure 11 shows an example of a feature model which employs the word form (FORM), part of speech (POS), and dependency relation (DEP) of a given token. The feature model is depicted as a matrix where rows denote tokens in the parser configuration, defined relative to the stack (S), input queue (I) and dependency graph (G), and columns denote attributes. Each cell containing a + corresponds to a feature of the model. Examples of the features include part-of-speech for the top of the stack, lexical form for the next and previous (*next-1*) input tokens and the dependency relation of the rightmost sibling of the leftmost dependent of *top*.

[139] See Nivre 2006 for details about the derivation of training data.

[140] To be precise, classification is performed on the basis of equivalence classes of configurations, where equivalence classes are constructed in terms of parse configurations and the features employed to represent them. The function Φ defines an equivalence relation over properties of configurations and is composed of a set of feature functions which each pick out a certain property of the current configuration (Nivre 2006).

	FORM	POS	DEP
S:*top*	+	+	+
S:*top*+1		+	
I:*next*	+	+	
I:*next*−1	+		
I:*next*+1	+	+	
I:*next*+2		+	
G: head of *top*	+		
G: leftmost dependent of *top*			+
G: rightmost dependent of *top*			+
G: leftmost dependent of *next*	+		+
G: leftmost dependent of head of *top*			+
G: leftmost sibling of rightmost dependent of *top*			+
G: rightmost sibling of leftmost dependent of *top*	+		
G: rightmost sibling of leftmost dependent of *next*		+	+

Figure 11: Feature model for Swedish; S: stack, I: input, G: graph; $\pm n$ = *n* positions to the left(−) or right (+).

8.1.3.3 Training and parsing

Parsing with the MaltParser system involves two phases – a *training* phase and a *parsing* phase (Nivre 2006). Training involves the extraction of feature vectors from the gold standard data set, and the induction of a parse guide. Parsing proceeds by extraction of feature vectors for every non-deterministic configuration and querying of the parse guide. Classifiers can be trained using any machine learning approach, but the best results have so far been obtained with support vector machines, using LIBSVM (Chang and Lin 2001) with a quadratic kernel $K(x_i, x_j) = (\gamma x_i^T x_j + r)^2$, see (Nivre et al. 2006) for more detail.

8.1.3.4 Incrementality

Strict incrementality in parsing involves connectedness at each point during the analysis of the input string. Unlike many other data-driven parsers, MaltParser approaches incrementality. Nivre (2004) shows that while incrementality in the strict sense is not attainable in dependency parsing, the arc-eager parsing algorithm employed in MaltParser is optimal in that it provides a close approx-

imation of incrementality.[141]

8.2 Error analysis

In grammar-driven systems, errors may be directly related to properties of the grammar and error analysis can exploit this more transparent relation in diagnosing and eliminating errors.[142] In many data-driven systems, however, there is no explicit grammar responsible for errors, hence error analysis consists solely in analysis of the relation between the input and output data. A deeper analysis of specific error sources in data-driven parsing is clearly an important step towards a further advancement of the state of the art.

The rest of this chapter will be devoted to an in-depth error analysis of argument assignment in a data-driven dependency parser – MaltParser, as trained on the Swedish treebank Talbanken05, see section 5.1.1. We start out by formulating a methodology for error analysis which allows us to quantify parser performance with respect to specific dependency relations and specific error types over these relations. The following error analysis will focus on errors in the assignment of argument relations and relate these errors to morphosyntactic properties of the arguments, the set of which is bounded by the features employed during parsing and specified explicitly in the feature model.

We will attempt to provide answers to the following questions:

Error analysis How may we characterize the errors performed by the parser?

Argument errors What characterizes the errors in argument assignment? May the errors be related in any consistent way with variation in the linguistic expression of arguments?

Generalizations Which types of generalizations may be acquired regarding syntactic arguments in a strictly data-driven setting?

Argument disambiguation To what extent may syntactic arguments be distinguished based solely on surface properties like lexical form, morphology and word order?

[141] The arc-eager algorithm differs from the standard shift-reduce algorithm for dependency structures in treating right and left dependencies differently, hence coping with chains of right dependencies which requires stacking of waiting head/dependents. Incrementality is then redefined as a restriction on connectedness between the stack and the graph under construction.

[142] We may distinguish between *error analysis* and *error mining*, where error analysis has a focus on characterizing (and possibly correcting) errors and makes use of a gold standard to locate the errors, usually in the form of a treebank or a test suite. Error mining, on the other hand, focuses primarily on locating errors, see for instance van Noord 2004.

8.2.1 A methodology for error analysis

An error analysis of parse results may typically be characterized by the following two steps:

1. locate the errors
2. characterize the errors

Step 1 involves comparison of the parser output with a gold standard corpus and the application of an evaluation metric. Step 2 is less straightforward and more dependent on the aim of the analysis. We will treat errors as sets of nodes in a dependency structure and characterize these by lexical and structural properties, such as part-of-speech and dependency label.

8.2.1.1 Evaluation metric

In the evaluation of dependency parsing, overall parsing accuracy is commonly reported using the standard metrics of *unlabeled attachment score* (UAS) and *labeled attachment score* (LAS), i.e., the percentage of nodes that are assigned the correct head *without* (unlabeled) or *with* (labeled) the correct dependency label:

$$\text{UAS} = \frac{\text{\# correctly attached tokens}}{\text{\# tokens}}$$

$$\text{LAS} = \frac{\text{\# correctly attached and labeled tokens}}{\text{\# tokens}}$$

For analysis of performance in the assignment of specific dependency labels, we employ the standard measures of precision and recall, as well as the combined, balanced F-score. Note that only correctly labeled *and* attached tokens are considered as true positives.

8.2.1.2 Error sets and types

The dependency relations in the treebank data adhere to the single-head constraint, see 5.1.1, hence we may equate errors with token nodes in a dependency graph with corresponding properties, such as head, dependency relation etc.[143] In general, we will treat error analysis as dealing with sets of errors, i.e.

[143] As Nivre (2006) notes, the single-head constraint allows for the assignment of dependency relations to nodes, rather than arcs, which simplifies the formulation of the labeling performed during parsing. In the same way it allows for error analysis dealing with dependent nodes.

token nodes, and comparisons between parsers as standard set-theoretic operations over these sets. The sets of errors will vary depending on the evaluation metric, but will always constitute the complement set of the set of correct instances according to the chosen metric. For the UAS and LAS evaluation metrics defined above, we define the sets of correct instances, $correct_U$ and $correct_L$, and the sets of errors, $error_U$ and $error_L$, as follows:

correct$_U$ $= \{x|$head of x is correct$\}$

error$_U$ $= \{x|$head of x is wrong$\}$

correct$_L$ $= \{x|$head of x is correct and dependency label of x is correct$\}$

error$_L$ $= \{x|$head of x is wrong and dependency label of x is wrong$\} \cup \{x|$head of x is wrong and dependency label of x is correct$\} \cup \{x|$head of x is correct and dependency label of x is wrong$\}$

In comparing sets of errors P_A and P_B for two parsers, we may examine properties of their *intersection*, $P_A \cap P_B$, and *difference* $P_A - P_B$, where:

$P_A \cap P_B = \{x \mid x \in P_A \text{ and } x \in P_B\}$

$P_A - P_B = \{x \mid x \in P_A \text{ and } x \notin P_B\}$, and $P_A - P_B = P_A - (P_A \cap P_B)$

When parser A is the gold standard data set, we obviously have that $P_A = \{\}$. In a comparison between two parsers, we wish to locate the differences responsible for a general improvement or deterioration in overall results. We may thus define improvement and deterioration of results for the error set of a new parser P_N, compared to that of a baseline parser P_{Bl}, as follows:

improvement if $\mid P_{Bl} \mid > \mid P_N \mid$. We may then examine properties of the *corrected errors* in the difference set $P_{Bl} - P_N$ further.

deterioration if $\mid P_{Bl} \mid < \mid P_N \mid$. We may then examine properties of the *new errors* in the difference set $P_N - P_{Bl}$ further.

In characterizing the results we may sort the sets of errors defined above into *error types*, based on various, relevant properties. With our main objective being an analysis of argument disambiguation, the main focus will be on analysis of labeled results, with a focus on the argument dependency relations. For the analysis of the labeled results we create error types based on dependency relations:

- the gold dependency relation in the error (Dep_{gold})

- the gold and system assigned dependency relation in the error ($Dep_{gold}\_Dep_{sys}$).[144]

For analysis of unlabeled results, the error types will be given by the part-of-speech of the dependent and/or head:

- part-of-speech of dependent in the error (POS_{dep})
- part-of-speech of dependent and gold head in the error ($POS_{dep}\_POS_{head}$)
- part-of-speech of dependent and gold head and erroneous system head in the error ($POS_{dep}\_POS_{head}\_POS_{err}$)

8.2.2 Data

The data for the error analysis of argument assignment in Swedish was obtained by parsing the written part of Talbanken05 with MaltParser.[145] We employed the settings optimized for Swedish in the CoNLL-X shared task (Nivre et al. 2006), with the feature model presented in figure 11. As we can see, the features employed during parsing are part-of-speech (POS), lexical form (FORM) and structural properties of the dependency graph under construction (DEP). We employed 10-fold cross validation for training and testing, and the overall result for unlabeled and labeled dependency accuracy is 89.87 and 84.92, respectively.[146]

8.2.3 General overview of errors

Table 8.1 presents a list of the overall most frequent error types ($Dep_{gold}\_Dep_{sys}$) in the data, sorted by absolute frequency. The most frequent error type, ET_OA exemplified by (131), consists largely of prepositional, post-nominal modifiers which have been analyzed as object adverbials.[147] In other words, these errors, as well as the converse OA_ET errors in (132), are prepositional attachment errors.

[144] Note that it might be the case that $Dep_{gold} = Dep_{sys}$ when head attachment alone is the source of error.

[145] All examples in the current chapter and chapter 9 are taken from the written sections of Talbanken05, unless otherwise stated.

[146] Note that these results are slightly better than the official CoNLL-X shared task scores (89.50/84.58), which were obtained using a single training-test split, not cross-validation. Note also that, in both cases, the parser input contained gold standard part-of-speech tags.

[147] Object adverbials (OA) are adverbials which are closely related to the verb, much like objects, without necessarily being subcategorized for by the verb. They are predominantly (90.3%) headed by a preposition and the choice of preposition is governed by the verb (Teleman 1974).

(131) *Genom* *deGaulle* *bröts* *länken* **med** *NATO*
through deGaulle broke-PASS link-DEF with NATO
'Through deGaulle the ties with NATO were severed'

(132) *På* *fredagen* *disputerar* *Åke* *Nilsson* **på** *avhandlingen*
on friday-DEF defends Åke Nilsson on thesis-DEF
'This Friday, Åke Nilsson will defend the thesis'

We also find a range of other adverbial relations among the errors presented in table 8.1. Recall from section 5.1.1 that the annotation in Talbanken05 makes numerous, fine-grained distinctions in adverbial functions (spatial, temporal, modal, comparative etc.). These clearly prove difficult for the parser to replicate.

Among the most frequent errors, we also find a large group involving the core argument relations of subjects – regular and formal subjects – and direct object. In particular, confusion of the two argument functions of subject and direct object (SS_OO, OO_SS) are among the top ten most frequent error types with respect to dependency assignment.

8.3 Errors in argument assignment

In section 5.1.1 we provided an overview of the dependency relations in Talbanken05 and divided these into argument and non-argument relations. The argument relations were either subcategorized for by the verb or thematically entailed by the verb. We may examine the parse performance for each of the argument relations in terms of the class-based performance measures of precision, recall and F-score, see table 8.2.

It is quite clear that there is a direct relation between the frequency of the dependency relation in the treebank and the parser performance. The most frequent relations are also the relations for which the parser performs best – SS (90.25), SP (84.82), OO (84.53).

Table 8.1 shows the most frequent error types involving argument relations.[148] We find frequent error types involving different kinds of subjects (SS, FS , ES), objects (OO, IO) and predicatives (SP). In the following sections we examine these errors in more detail. In order to relate the errors to properties of Scandinavian type languages we briefly examine the realization of these argument relations in Scandinavian in section 8.3.1, supplied with quantitative

[148]Table 8.1 includes both errors where the gold standard relation is an argument relation and/or where the proposed system relation is an argument relation, since both of these will affect the results for the argument relation in terms of precision and recall, respectively

Gold	System	#
ET	OA	450
SS	OO	446
OA	ET	410
AA	RA	404
AA	OA	398
TA	AA	372
RA	AA	311
OO	SS	309
RA	OA	308
AA	TA	290
FS	SS	281
OA	RA	270
OA	AA	269
SS	ROOT	265
AA	ET	251
SS	FS	133
RA	ET	244
SP	SS	240
SS	DT	238
ET	AA	232
PA	DT	231

Gold	System	#
SS	OO	446
OO	SS	309
FS	SS	281
SS	ROOT	265
SP	SS	240
SS	DT	238
OO	ROOT	221
SS	SP	206
DT	SS	146
SS	CC	137
SP	AA	136
SS	FS	133
OO	PA	126
OO	AA	103
OO	DT	99
IO	OO	97
ES	OO	95
ET	OO	91
DT	OO	90
ROOT	SS	86

Table 8.1: 20 overall most frequent error types (left) and 20 most frequent argument error types (right), where SS=subject, OO=object, AA=other adverbial, OA=object adverbial, ET=nominal post-modifier, RA=spatial adverbial, TA=time adverbial, FS=formal subject, SP=subject predicative, DT=determiner, CC=second conjunct, AA=adverbial, PA=prepositional complement, IO=indirect object, ES=logical subject, ET=nominal post-modifier.

data from Talbanken05. We then examine the most frequent error types for argument relations in a bit more detail. Table 8.1 shows that regular subjects and direct objects are commonly confused for each other, hence these will be treated together in section 8.3.2. Section 8.3.3 will examine errors involving formal or expletive subjects, as well as the relation of logical subjects and in sections 8.3.4–8.3.5 we will examine indirect objects and predicative constructions, respectively.

Deprel		Gold	Correct	System	Recall	Precision	F-score
SS	subject	19383	17444	19274	90.00	90.51	90.25
SP	subject predicative	5217	4416	5196	84.65	84.99	84.82
OO	direct object	11089	9639	11718	86.92	82.26	84.53
IO	indirect object	424	276	301	65.09	91.69	76.14
AG	passive agent	334	249	343	74.55	72.59	73.56
VO	object inf.	121	84	112	69.42	75.00	72.10
ES	logical subject	878	562	687	64.01	81.80	71.82
FS	formal subject	884	578	737	65.38	78.43	71.31
VS	subject inf.	102	47	58	46.08	81.03	58.75
FO	formal object	156	70	91	44.87	76.92	56.68
OP	object predicative	189	42	112	22.22	37.50	27.91
EO	logical object	22	2	3	9.09	66.67	16.00

Table 8.2: Dependency relation performance: total number of gold instances (Gold), system correct (Correct), system proposed (System), recall, precision and F-score

8.3.1 Arguments in Scandinavian

In chapter 4, we noted that the expression of grammatical function in Scandinavian is governed largely by linear order in the clause, as expressed through the fields schemas. The linearizations of grammatical functions in main and subordinate clauses presented in section 4.2.3 are repeated in (133) and (134) below:

(133) Linearization of grammatical functions in declarative, main clauses:
XP | V_{fin} SUBJ S-ADV | $V_{non-fin}$ OBJ_{ind} OBJ_{dir} ADV

(134) Linearization of grammatical functions in subordinate clauses
subj | SUBJ S-ADV V_{fin} | $V_{non-fin}$ OBJ_{ind} OBJ_{dir} ADV

We recall that the initial field is characterized by variation and may be filled by pretty much any constituent (XP), whereas subordinate clauses are non-V2. The interpretation of (133)–(134) is that if core grammatical functions do not occur preverbally, they are predicted to be linearized in this order. As we shall see, however, the argument relations differ with respect to how likely it is that they appear in initial position. As we also recall, morphological marking is not a very reliable indicator of grammatical function, and only a subset of the personal pronouns are marked for case.

Part-of-speech		#	%
PO	pronoun	9549	49.3
N	noun	9178	47.4
V	verb	405	2.1
AJ	adjective	136	0.7
PR	preposition	50	0.3
R	numeral	43	0.2
	other	22	0.1
Total		19383	100.0

Table 8.3: Part-of-speech for subjects (SS) in Talbanken05.

8.3.1.1 Subjects

Subjects in Scandinavian are realized by nominal expressions: various types of noun phrases, as in (135)–(136) where the subject is a noun and pronoun, respectively, or subordinate clauses, as in (137) where the subject is a subordinate clause.

(135) **Speciallläraren** *kan också komma till klassrummet*
special-teacher-DEF can also come to classroom-DEF
'The special education teacher may also come to the classroom'

(136) **De** *har alltså ansvar och omsorg om barnen*
they have so responsibility and care for children-DEF
'So, they have the responsibility to care for the children'

(137) **Att värderingarna förändrats** *är helt säkert riktigt*
that value-PL.DEF changed-PASS is totally certain correct
'That the values have changed is almost certainly correct'

Table 8.3 shows the distribution of the various parts-of-speech over the subject relation in the written sections of Talbanken05.[149] We find that pronouns and nouns account for 96.6% of the subjects, and 2.1% of the subjects are subordinate clauses, listed as 'verb' in the overview since verbs are clausal heads in the dependency representation. We also observe that almost half of all subjects are expressed by a pronoun, supporting cross-linguistic tendencies in the referentiality of subjects, noted in section 3.4.

[149] Since verbs are the heads of clauses in dependency grammar, clausal arguments are represented as 'verb' in the overviews over parts-of-speech for the different argument relations, e.g., table 8.3.

	Before		After		Total	
Deprel	#	%	#	%	#	%
SS	14958	77.2	4425	22.8	19383	100.0
FS	609	68.9	275	31.1	884	100.0
ES	8	0.9	870	99.1	878	100.0
OO	611	5.5	10478	94.5	11089	100.0
IO	5	1.2	419	98.8	424	100.0
SP	491	9.4	4726	90.6	5217	100.0

Table 8.4: Ordering relative to verb for argument relations in Talbanken05.

The linearization of grammatical functions in main clauses places the subject in the Midfield, directly following the finite verb. In table 8.4, we find an overview of the linear position of various arguments with respect to the head verb. For subjects, the head verb is always the finite verb, and we find that only 22.8% of the subjects follow the finite verb (after). These counts include both subjects of main and subordinate clauses and if we restrict our counts to subjects in main clauses, we find slightly more variation; 35.7% of the main clause subjects occur following the finite verb.

In chapter 4, we noted that the fields schema does not directly express the tendency for subjects to occur preverbally. Regardless of the status of the clause, this tendency is certainly supported by the data: 77.2% of the total subjects occur preverbally, and 64.3% of the main clause subjects.

Formal subjects

Formal or expletive subjects are characterized by a lack of semantic content. They occupy a subject position, but do not share thematic properties with regular subjects. The Scandinavian languages, like English and the other Germanic languages, enforce a subject requirement, also known as the Extended Projection Principle (Chomsky 1981) and the Subject Condition (Baker 1983), which requires that all declarative main clauses must contain a subject. As a consequence, a formal subject is employed when a thematic subject for various reasons may not occupy a subject position.[150]

We may discern six general types of constructions where an expletive subject figures in Scandinavian (Teleman, Hellberg and Andersson 1999):

[150] Note that the reasons for using a formal subjects vary; the thematic subject may be demoted, as in impersonal passives, or prefer a postposed position due to discourse-oriented constraints, as in presentational constructions.

1. Existential/presentational constructions:

   (138) **Det** *finns olika slags barnhem*
   it exists different sorts orphanages
   'There are different kinds of orphanages'

2. *Det* with oblique subject:

   (139) *För somliga räcker* **det** *med bröstet eller flaskan*
   for some is-sufficient it with breast-DEF or bottle-DEF
   'For some it is sufficient with breastfeeding or the bottle'

3. Extraposed finite or non-finite subordinate clause, as in (140):

   (140) **Det** *är ytterst lätt att gå ur kyrkan*
   it is extremely easy to go out-of church-DEF
   'It is very easy to leave the church'

4. Impersonal passive:

   (141) **Det** *syndas ofta utan tvekan* ...
   it sin-PASS often without doubt ...
   'There is often sinning going on, without doubt ...'

5. Weather-verbs or verbs of perception denote an event or state with no clear agent:[151]

   (142) **Det** *regnar/åskar/snöar*
   it rains/thunders/snows
   'It rains/thunders/snows'

   (143) *Nu luktar* **det** *torkad frukt i källaren*
   now smells it dried fruit in cellar-DEF
   'It smells of dried fruit in the cellar'

6. Cleft constructions:

   (144) **Det** *är hon som svarar för de inre relationerna*
   it is she who answers for the internal relation-PL.DEF
   'It is she who is responsible for the internal relations'

[151] The examples in (142)–(143) are from Teleman, Hellberg and Andersson 1999.

Part-of-speech		#	%
V	verb	492	56.0
N	noun	329	37.5
PO	pronoun	47	5.4
AJ	adjective	7	0.8
PR	preposition	2	0.2
P	participle	1	0.1
Total		878	100.0

Table 8.5: Part-of-speech for logical subjects (ES) in Talbanken05.

Formal subjects are distinguished from regular subjects in Talbanken05 and assigned a separate dependency relation (FS). The category of formal subject is employed only in cases where there is also an expressed logical subject (ES) and is exclusively realized as the impersonal 3rd person pronoun *det* 'it'.[152] The logical subject may be realized by a nominal element or a subordinate clause. When it is nominal it is normally realized as an object and is in complementary distribution with regular objects in existentials and impersonal passives of transitive verbs. As table 8.5 shows, subordinate clauses are the most common logical subjects (56%), followed by nouns (37.5%) and pronouns (5.4%). It is interesting to note that the distribution of part-of-speech over the logical subject function is very similar to that of regular objects, see table 8.6, in contrast to regular subjects, cf. table 8.3. The main criteria for the assignment of the FS and ES dependency relations may be summarized as follows:

Expressed logical subject the categories of formal (FS) and logical subject (ES) entail each other (bidirectionally); annotation for FS only when there is an expressed ES.

Replacement replacement of the formal subject with the logical subject should result in a grammatical sentence

As a consequence of the above criteria, not all of the construction types listed above of are annotated as involving a formal subject. Talbanken05 identify types 1, 3 and some cases of 4 as formal subjects (FS). The first criterion, demanding an expressed logical subject, excludes impersonal passives of in-

[152] We find a total of 884 formal subjects in the written sections of Talbanken and 99.6% of these are realized as *det* 'it'. The remaining three are verbal and must be attributed to annotation mistakes.

transitive verbs, as well as the weather-type verbs of type 5.[153] The second criterion excludes expletive subjects expressed as obliques (type 2) and cleft constructions. These subjects are annotated as regular subjects (SS), rather than formal subjects.

The constructions analyzed as containing a formal subject (FS), and consequently also a logical subject (ES), are thus the existentials (type 1), the extrapositions (type 3) and impersonal passives of transitive/ditransitive verbs (type 4). [154]

8.3.1.2 Objects

In section 3.1, we introduced the notion of subjects and objects as the *core* arguments of a clause, denoting its main participants. Objects are also nominal constituents of the verb and may be headed by pronouns and nouns, as in (145) as well as subordinate clauses, as in (146):

(145) *Men* *människorna* *vill* *ha* *mera* **bostäder**
but people-DEF want have more housing
'But the people demand more housing'

(146) *Vi* *måste* *kolla* **om** **dubbelröstning** **skett**
we must check whether double-voting happened
'We have to check whether or not double-voting has taken place'

Table 8.6 shows the distribution of the various parts-of-speech over the direct object relation (OO) in Talbanken05. We find a clear difference from the subject relation in table 8.3, where the proportion of subordinate clauses (verb) constitutes the most striking difference. These account for 18.3% of the direct

[153]The annotation manual mentions a few exceptions to the criterion demanding an expressed logical subject. In sentences where the subject is the adverbial pronoun *här* 'here', we may get a logical subject analysis without a formal subject (Teleman 1974: 46). In sentences where a clause containing a formal subject is analyzed as modifying the logical subject clause, the subordinate formal subject does not have a corresponding logical subject (Teleman 1974: p. 46). As a consequence, we find that the treebank contains slightly different numbers of elements annotated as formal (FS) and logical subject (ES) – 884 vs. 878 instances, respectively.

[154]We may note that the replacement described in the replacement criterion above, can be related to the argument status of the subject and is not randomly chosen. Neither is the group of remaining formal subject constructions without certain commonalities. It has been argued several places in the literature that the subject of weather verbs are quasi-arguments which therefore cannot readily be replaced with another argument (Chomsky 1981; Falk 1993). The exclusion of subjects of impersonal passives over intransitive verbs from the group of formal subjects, however, is unfortunate. Clearly, these should also be expletives on a par with their transitively formed counterparts.

Part-of-speech		#	%
N	noun	6377	57.5
PO	pronoun	2247	20.3
V	verb	2034	18.3
AJ	adjective	282	2.5
PR	preposition	60	0.5
R	numeral	34	0.3
P	participle	30	0.3
AB	adverb	15	0.1
	other	10	0.1
Total		11089	100.0

Table 8.6: Part-of-speech for direct objects (OO) in Talbanken05.

objects, but only 2.1% of the subjects. Note however, that clausal objects include a wide category of syntactic constituents in the Talbanken05 annotation scheme, including infinitival complements of control and raising verbs.[155] We furthermore observe that direct objects are clearly less commonly expressed pronominally (20.3%), than subjects (49.3%). As indicated by the linearization of the fields analysis in (133)–(134), objects are positioned in the End field, following any finite and non-finite verbs. Objects are often claimed to be in a closer structural relation with the verb than the subject and depending on the valency of the lexical verb, there may be one or two objects – a direct (OO) and an indirect object (IO). The frequencies of ordering with respect to the lexical verb in table 8.4 above, clearly show the strong preference for postverbal position in the case of both types of objects. We find that 97.3% of all direct objects and 98.8% of the indirect objects are postverbal (after).

Indirect objects are in general much less frequent than direct objects and in Talbanken05 there are only a total of 424 instances. These show a strong preference for pronominal realization, as we see from table 8.7. In section 7.1.2 we also noted that indirect objects show a general preference for animate denotation. As we also saw in chapter 7, these properties are not unrelated, since animate reference is typically expressed pronominally. There is, however, a complicating factor in the annotation of indirect objects; the reflexive argument of transitive reflexive verbs is annotated as an indirect object, as in (147) below.

[155]The annotation manual proposes a replacement test for objecthood of subordinate clauses: if the clause can be replaced by a pronoun, e.g. *något* 'something' or *detta* 'this', it is annotated as object.

Part-of-speech		#	%
PO	pronoun	320	75.5
N	noun	98	23.1
AJ	adjective	5	1.2
ID	idiom	1	0.2
Total		424	100.0

Table 8.7: Part-of-speech for indirect objects (IO) in Talbanken05.

(147) *Barnens* *hus* *tänker* *jag* **mig** *hellre* *av* *modell* *äldre*
children-GEN house think I myself rather of model older
villa
villa
'The children's house I rather imagine as an older villa'

In these types of constructions, the pronominal argument may not be anything but a reflexive pronoun, coreferent with the subject. The reflexive pronoun accounts for 44.4% of all the indirect object instances in Talbanken, which shows that indirect objects are in fact even more infrequent than first assumed.

8.3.1.3 Predicatives

Predicatives establish a core argument – a subject or object – as having a certain property, being a member of a certain class of referents, or establishes referential identity.

In Scandinavian, predicatives are largely realized by adjectives, participles and nominals (Teleman 1974), as illustrated by the overview of part-of-speech for subject predicatives in table 8.8. The adjectival predicatives agree with the predicated argument in gender, definiteness and number. Subject predicatives are complements of a small set of verbs - *vara* 'be', *bliva* 'become', *varda* 'become', *heta* 'named', *kallas* 'be-called', *förefalla* 'seem', *verka* 'seem', *se . . . ut* 'look' and typically occupy the object position, with which it is in complementary distribution, following the finite and possibly non-finite verb(s).

(148) *Deras* *val* *av* *äktenskapspartner* *blev* *kanske*
their choice-SG.NEUT of marriage-partner became possibly
slumpmässigt
random-SG.NEUT
'Their choice of partner was possibly random'

(149) *Per-Ola Larsson som är* **sekreterare** *i organisationen …*
Per-Ola Larsson who is secretary in organization-DEF …
'Per-Ola Larsson who is the secretary in the organization …'

(150) *Detta är* **EEC-organisationen** *i dag*
this is EEC-organization-DEF to day
'This is the EEC organization at present'

We may distinguish semantically and referentially between *descriptive* and *identifying* predicatives (Teleman, Hellberg and Andersson 1999), where the former classify or characterize the predicated argument – either subject or object – further, as in (149), whereas the latter establish a relation of strict coreference, as in (150), i.e. the extension of the two arguments are identical. Identifying predicatives occur only with the copula verbs *vara, bli, förbli* 'be, become, remain', and the nominal predicative is typically definite, as in (150). Descriptive predicatives in contrast, are usually indefinite, and may occur without indefinite article, as in (149) above. The use of an evaluative adjective is typical, where the noun is commonly a hypernym of the predicational argument and the main semantic contribution of the predicative is in the information expressed by the adjective. In the treebank data, we find that the nominal subject predicatives exhibit a preference for indefinite expression; 89.6% of the subject predicatives expressed as nouns are indefinite.[156] This indicates that descriptive predicatives are most common in Swedish.

Talbanken05 distinguishes subject predicatives (SP), exemplified by (148)–(150) above, and object predicatives (OP), exemplified by (151):

(151) *Lämna aldrig spädbarn* **ensamma** *hemma*
leave never infant-PL alone-PL home
'Never leave an infant home alone'

There are 5223 subject predicatives in Talbanken05 and these are distributed across the various parts-of-speech as shown in table 8.8 The object predicatives are highly infrequent and there are only 190 instances in the treebank. These show similar distributional properties with respect to part-of-speech as the subject predicatives.[157]

With respect to word order placement, subject predicatives are in the postverbal position in a clear majority of cases (91.6%), see table 8.4 above. We also

[156] Out of all nominal (noun or pronoun) subject predicatives, 77.6% are indefinite. We then count all pronouns as definite.

[157] We find that 41.6% of the object predicatives are adjectives, 28.9% are nouns and 16.8% participles. One difference is in the fact that there are no pronominal object predicatives in the data set.

Part-of-speech		#	%
AJ	adjective	2262	43.4
N	noun	1801	34.5
P	participle	572	11.0
PO	pronoun	280	5.4
PR	preposition	149	2.9
V	verb	85	1.6
AB	adverb	46	0.9
R	numeral	20	0.4
	other	2	0.0
Total		5217	100.0

Table 8.8: Part-of-speech for subject predicatives (SP) in Talbanken05.

observe variation in position for this relation and 9.4% of the subject predicatives are located preverbally.[158] Object predicatives, on the other hand, are almost exclusively postverbal (99.0%).

8.3.2 Subject and direct object errors

We noted above that the two most frequent error types involving argument relations were errors analyzing subjects as objects (SS_OO) and vice versa (OO_SS). Table 8.9 shows an overview of the main error types in the assignment of the subject and direct object dependency relations.

In addition to the confusion of subjects and objects, which constitutes the most common error type for both relations, we find that both subjects and objects are quite commonly assigned status as the root of the dependency graph (ROOT).[159] For both argument relations we also observe error types indicating confusion with other argument relations. For subjects we observe confusion with the other main argument functions, such as subject predicatives (SP) and expletive subjects (FS), as well as confusion with determiners (DT)

[158] This figure may be compared to direct objects which are only found preverbally in 2.7% of the cases in Talbanken.

[159] The root relation is the default head for all nodes in the dependency graph since the graph is initialized with all nodes attached to the root to ensure connectedness. The dependency graph is thus not guaranteed to be a tree in the technical sense, but is always a set of subtrees attached to the artificial root. These are thus errors where the parser has not located a more appropriate attachment and label. Nominal elements may very well be attached to the root, for instance in sentence fragments which lack a finite verb, but for the error types involving erroneous assignment to the root, this is clearly not the case.

Error types for subjects (SS)		
Gold	Sys	#
SS	OO	446
SS	ROOT	265
SS	DT	238
SS	SS	216
SS	SP	206
SS	CC	137
SS	FS	133
SS	PA	53
...	...	...

Error types for objects (OO)		
Gold	Sys	#
OO	SS	309
OO	ROOT	221
OO	OO	149
OO	PA	126
OO	AA	103
OO	DT	99
OO	ET	58
OO	OA	57
...	...	...

Table 8.9: Error types for subjects (left) and objects (right).

and prepositional complements (PA). For objects we observe primarily confusion with various adverbial relations (AA, ET, OA), as well as confusion with prepositional complements (PA) and determiners (DT). We may note that confusion with DT and PA indicate that a phrasal reading rather than a clausal one has been chosen. Since there is no explicit notion of phrases in a dependency analysis, these errors in dependency labeling are primarily errors also in head attachment, as opposed to the errors confusing argument relations. For the SS_OO and OO_SS errors only 19.5% and and 17.8% involve incorrect head assignments, whereas the corresponding proportions for the SS_DT and OO_PA errors are 92.9% and 100%, respectively.

There are various sources of errors in subject/object assignment. Common to all of them is that the parts of speech that realize subjects and objects are compatible with a range of dependency relations. Pronouns, for instance, may function as subjects, objects, determiners, predicatives, conjuncts, prepositional objects, etc. In addition, we find "traditional" attachment ambiguity errors, for instance in connection with coordination, subordination, particle verbs, etc. These represent notorious phenomena in parsing, and are by no means particular to Swedish. This language, however, in addition exhibits ambiguities in morphology and word order which complicate the picture further. The confusion of subjects and objects follows from lack of sufficient formal disambiguation, i.e., simple clues such as word order, part-of-speech and word form do not clearly indicate syntactic function. The reason for this can be found in ambiguities on several levels.

With respect to word order, we have seen that subjects and objects may both precede or follow their verbal head, but these realizations are not equally likely. Subjects are more likely to occur preverbally, whereas objects typically

		Before		After		Total	
Gold	System	#	%	#	%	#	%
SS	OO	103	23.1	343	76.9	446	100.0
OO	SS	103	33.3	206	66.7	309	100.0

Table 8.10: Ordering relative to verb for the SS_OO and OO_SS error types.

occupy a postverbal position. Based only on the word order preferences discussed above, we would expect postverbal subjects and preverbal objects to be more dominant among the errors than in the treebank as a whole (23% and 6% respectively), since they display word order variants that depart from the canonical, hence most frequent, ordering of arguments. This is precisely what we find. Table 8.10 shows a breakdown of the errors for confused subjects and objects and their position with respect to the verbal head.

We find that postverbal subjects (after) are in clear majority among the subjects erroneously assigned the object relation. Due to the V2 property of Swedish, the subject must reside in a position following the finite verb whenever another constituent occupies the preverbal position, as in (152) where a direct object resides sentence-initially or (153) where we find a sentence-initial adverbial:

(152) *Samma erfarenhet gjorde* **engelsmännen**
same experience made englishmen-DEF
'The same experience, the Englishmen had'

(153) *År 1920, och först då, fick den gifta* **kvinnan**
Year 1920, and first then, got the married woman-DEF
fullständig myndighet
complete rights
'It was not until 1920 that the married woman recieved full civil rights'

Whereas the postverbal subjects are in a non-canonical position, the preverbal subjects should be easier to locate since their structural position is a strong indicator for subjecthood. As table 8.10 shows, preverbal subjects are also in minority among the errors. However, when preverbal subjects are mistakenly assigned the object function, it is typically in cases where the parser has not been able to determine their clause-initial position. In subordinate clauses without complementizers, as in (154), this error is indicated also through erroneous head assignment to the matrix verb instead of the following head verb:

(154) *På denna grund tycker jag ett* **äktenskap** *ska byggas*
on this ground think I a marriage should build-PASS
'On these foundations I think that a marriage should be built'

For the confused objects we find a larger proportion of preverbal elements than for subjects, which is the mirror image of the normal distribution of syntactic functions among preverbal elements. As table 8.10 shows, the proportion of preverbal elements among the subject-assigned objects (33.3%) is notably higher than in the corpus as a whole, where preverbal objects account for a miniscule 6% of all objects.

The preverbal objects are topicalized elements which precede their head verb, which may be either the matrix verb, as in (155)–(157), or the verb of a following subordinate clause, as the relative clause in (158) below:[160]

(155) **Detta** *anser tydligen inte Stig Hellsten*
this means apparently not Stig Hellsten
'This, Stig Hellsten apparently does not believe'

(156) *Vilken* **uppfattning** *har mannen om kvinnans 'rätta*
which opinion has man-DEF about woman-DEF.GEN 'right
plats' i hemmet?
place' in home-DEF?
'Which opinion does the man have about the woman's place in the home?'

(157) *Kärlekens innersta* **väsen** *lär inte något politiskt parti*
love-DEF.GEN inner nature seems not any political party
kunna påverka
can-INF influence
'The inner nature of love, it seems that no political party can influence'

(158) **Vad** *Hellsten uppfattar som något tryggt och fast,*
what Hellsten interprets as something safe and firm,
blir …
becomes …
'What Hellsten interprets as something safe and firm, becomes …'

Contrary to our initial hypothesis, however, we find a majority of postverbal objects among the objects confused for subjects. These objects are interpreted as subjects because the local preverbal context strongly indicates a subject analysis. This includes verb-initial clauses as in (159) where we find a clause-initial imperative or cases of VP coordination, as in (160), as well as constructions where the immediate preverbal context consists of an adverbial and the subject is non-local, as in (161) and (162) below.

[160] Note that raising verbs, like *lär* in example (157), are analyzed as normal auxiliary verbs, hence the topicalized objects are annotated as dependents of these.

(159) *Glöm aldrig det* **löfte** *om trohet för livet*
forget never that promise of faithfulness for life-DEF
'Never forget that promise of faithfulness for life'

(160) *. . . om man tidigare varit gul i ögonen, haft* **gulsot**
. . . if one earlier been yellow in eyes-DEF, had jaundice
eller . . .
or . . .
'. . . if one earlier has had yellow eyes, had jaundice or . . . '

(161) *Ungdomarna blir med barn och det sociala trycket*
teenagers become with child and the social pressure-DEF
nästan tvingar **dem** *att gifta sig*
almost forces them to marry themselves
'The teenagers become pregnant and social pressure almost forces them to get married'

(162) *Eftersom man har full frihet att enkelt och snabbt ingå*
because one has full freedom to easily and quickly enter
äktenskap
marriage
'Because one has the freedom to easily and quickly get married'

The example in (161) is particularly interesting as it violates the V2-property, assumed to be a categorical constraint of Swedish. We may note that the examples in (159)–(162) above indicate acquisition of argument ordering resulting from the V2 requirement; when there is no preverbal argument or when the preverbal argument is not a good subject candidate, the argument following the verb is analyzed as subject. Recall, however, that the parser does not have information on tense or finiteness, hence overgeneralizes to examples like (162), where the verb is non-finite.

In addition to the word order variation discussed above, Swedish also has limited morphological marking of syntactic function, see section 4.1. Recall that nouns are only marked for genitive case and only pronouns are marked for accusative case. There is also syncretism in the pronominal paradigm. There are pronouns which are invariant for case, e.g. *det, den* 'it', *ingen/inga* 'no', and furthermore may function as determiners. This means that with respect to word form, only the set of unambiguous pronouns clearly indicate syntactic function. We may predict that subject/object confusion errors frequently exhibit elements whose syntactic category and/or lexical form does not disambiguate, i.e., nouns or ambiguous pronouns. Table 8.11 shows the distribution of nouns, functionally ambiguous and unambiguous pronouns and other parts of speech

Gold	System	Noun		Pro_{amb}		Pro_{unamb}		Other		Total	
SS	OO	324	72.6%	53	11.9%	29	6.5%	40	9.0%	446	100%
OO	SS	215	69.6%	74	23.9%	9	2.9%	11	3.6%	309	100%

Table 8.11: Part of speech for the SS_OO and OO_SS error types – nouns, ambiguous pronouns, unambiguous pronouns and other parts of speech.

for confused subjects/objects.[161] Indeed, we find that nouns and functionally ambiguous pronouns dominate the errors where subjects and objects are confused. Since case information is not explicitly represented in the input, this indicates that case is acquired quite reliably through lexical form.[162] The fact that we find a higher proportion of ambiguous pronouns among the objects erroneously assigned subject status indicates that the parser has acquired a preference for subject assignment of pronouns compatible with the difference in frequency for pronominal realization (SS_{pro} 49.2%, OO_{pro} 10.1%[163]).

As discussed earlier, not all of the objects erroneously assigned subject status are in preverbal position. In fact, a slight majority are still in postverbal position. This is in part due to another type of ambiguity in terms of syntactic category; both nouns and a subset of the pronouns may function as determiners (DT). In the postverbal position, word order demands (V2) create clusters of arguments where both a phrasal and a clausal interpretation is possible, i.e. the nouns are erroneously analyzed as modifiers instead of phrasal heads. In (163)–(164) we see examples which illustrate the phenomenon. The subject, expressed either as a noun or a (ambiguous) pronoun, is parsed as a determiner of the following noun and the following argument as the subject of the verb.

(163) *Naturligtvis knyter barnen* **kontakter** *utanför hemmet*
naturally attaches children contacts outside home-DEF
'Naturally, the children bond outside the home'

(164) *I ett äktenskap har ingen* **äganderätt** *över den andre*
in a marriage has no ownership over the other
'In a marriage, nobody has ownership over the other'

[161]The 'other' category consists mainly of verbs (heads of subordinate clauses), adjectives, participles and numerals functioning as nominal heads.

[162]Since pronouns are a closed class and it is mainly the set of personal pronouns that are marked for case, we would assume that this property can be acquired reliably without large amounts of training data.

[163]The proportion of pronouns is higher for both subjects and objects if we normalize over only nominal instances, i.e. excluding subordinate clauses: SS_{pro} 50.6%, OO_{pro} 23.7%. Even so, there is a clear difference between the two relations in terms of pronominal expression.

Error types for formal subjects (FS)		
Gold	System	#
FS	SS	281
FS	DT	10
FS	OO	6
FS	ROOT	3
FS	FO	1
FS	KO	1

Table 8.12: Error types for formal subjects.

The initial error analysis shows that the confusion of subjects and objects constitutes a frequent and consistent error during parsing. It is caused by ambiguities in word order and morphological marking and we find cases that deviate from the most frequent word order patterns and are not formally disambiguated by part-of-speech information. In order to resolve these ambiguities, we have to examine features beyond part-of-speech category and linear word order.

8.3.3 Formal subject errors

Table 8.12 presents the errors in dependency relation assignment performed by the parser for the relation of formal subject. We see that the confusion is almost exclusively with that of the regular subject function (SS). The errors for the function of logical subject, shown in table 8.13, vary over the functions compatible with a postverbal realization (objects (OO), object adverbial (OA), subject predicatives (SP) etc.). This is to be expected since the logical subject may be realized by both nominal phrases and clausal elements.

The confusion between regular and formal subjects is clearly caused by the fact that they may be realized by the same word form and may occupy the same structural positions (both pre- and postverbally). After all, formal subjects are subjects in all structural respects.[164] The impersonal, third person pronoun *det* 'it', is maximally ambiguous and may occupy a wide range of dependency relations, as table 8.14 illustrates. It occurs in all major argument relations

[164]One might argue that distinguishing formal subjects from other subjects is unduly complicating. It has been shown, however, that whereas more fine-grained dependency labels may affect parsing accuracy negatively, it improves semantic analysis, such as semantic role labeling (Johansson and Nugues 2007). To the extent that syntactic parsing is not simply a goal in itself, it seems that a more fine-grained analysis is worthwhile, and, in particular, with respect to phenomena like expletive categories which clearly have profound effects on the semantic interpretation of the arguments.

Error types for logical subjects (ES)		
Gold	System	#
ES	OO	95
ES	OA	55
ES	SP	26
ES	ET	23
ES	KA	20
ES	SS	19
ES	ROOT	19
ES	PA	15
ES	AA	15
ES	CC	5

Table 8.13: Error types for logical subjects.

Dependency relations of *det*		
Deprel	Abs	%
SS (regular subject)	1305	37.5
FS (formal subject)	881	25.3
DT (determiner)	791	22.7
OO (direct object)	226	6.5
HD (head of idiom)	86	2.5
PA (prep. complement)	79	2.3
SP (subject predicative)	33	0.9
...	...	...
Total	3480	100.0

Table 8.14: Dependency relations for the 3rd person pronoun *det* 'it' in Talbanken05.

and also occurs frequently as a definite determiner. The pronoun is not case-marked and hence is formally invariant in all the relations exemplified above. Since formal subjects are structural subjects it is not possible to differentiate the two categories based on structural properties of word order.

A relevant question then relates to how the parser manages to recognize any formal subjects at all. We may examine closer the errors performed by the parser and also the instances which were parsed correctly. Since the form does not vary it is clear that the analysis of the pronoun as formal or regular subject will be largely dependent on the relation with and properties of the verbal head. We saw earlier that position relative to the verbal head proved to

		Before		After		Total	
Gold	System	#	%	#	%	#	%
FS	FS	401	68.9	181	31.1	582	100.0
FS	¬FS	208	68.8	94	31.1	302	100.0
FS	SS	201	71.5	80	28.5	281	100.0
¬FS	FS	100	63.7	57	36.3	157	100.0

Table 8.15: Ordering relative to verb for formal subjects: correctly located (FS_FS) and errors – all (FS_¬FS) and confusion with subject relation (FS_SS).

be a contributing factor in the error analysis of regular subjects. As table 8.15 shows, however, there are no clear differences between correctly located formal subjects and ones that were not located by the parser or elements which were erroneously assigned the FS relation. These error sets exhibit similar distributions with regard to ordering with respect to the verbal head (before/after) as well as distance (difference between immediately preceding/following and the total before/after bins).

The occurrence of a formal subject is very much dependent on properties of the predicate and often reflects the argument structure of the verb. With no other formal or structural clues, we must assume that the interpretation of *det* 'it' as either a formal or a regular subject relies heavily on the lexical form of the verb. We may in addition predict that the correctly located formal subjects will be arguments of a smaller group of verbs which are frequently found with a formal subject, whereas the set of errors will exhibit a more heterogeneous group of verbal heads.

Table 8.16 compares the ten most frequent head verbs for the correctly located formal subjects with those of the errors.[165] We find that 37.3% of the correctly analyzed formal subjects are actually arguments of the existential predicate *finns* 'exists', as in (138) above. Overall, the set of head verbs for the correct formal subjects is smaller and we find on average 5 instances per head verb type, whereas the corresponding figure for the errors is 2.3. This indicates that the parser acquires lexical generalizations regarding these verbs and their argument structure. We may also note that the percentage of hapax legomena in the set of head verbs for the errors is 30%, but only 11% for the correctly recognized formal subjects – another observation which adds to the difficulty of correct analysis based on frequency for these arguments. The list of head verbs for the correct subjects consists largely of verbs which typically take a formal subject – existential predicates *finns* 'exists', *står* 'stands', complex

[165]When the verb is the copula *är* 'is', the subject predicate has been included to form a complex predicate of the type *vara_svårt* 'be_difficult'.

Head verbs - CORRECT	
finns 'exists'	217
står 'stands'	29
vara_svårt 'be_difficult'	16
vara_viktigt 'be_important'	16
har 'have'	13
går 'goes'	11
fordras '?'	9
vara_klart 'be_clear'	9
kan 'can'	8
måste 'must'	8
...	...

Head verbs - ERRORS	
kan 'can'	19
har 'have'	15
blir 'becomes'	12
kommer 'comes'	10
måste 'must'	9
skulle 'should'	9
går 'goes'	7
vara_plikt 'be_duty'	6
ska 'shall'	5
skall 'shall'	5
...	...

Table 8.16: 10 most frequent finite head verbs for the formal subjects in Talbanken05 – correctly located by the parser (left) and errors (right).

copular predicates *vara_svårt* 'be_difficult', *vara_viktigt* 'be_important' etc., whereas the list of head verbs for the errors consists to a large part of functional verbs – auxiliaries (*har* 'have', *blir* 'becomes') and modals (*måste* 'must'). In fact, only 9.6% of the head verbs for the correctly assigned subjects are functional verbs, whereas 30% of the head verbs for the errors are modal or temporal auxiliaries. This indicates that our earlier comments on distance to the verbal head are somewhat diffused. Although the distance to the finite verb serving as head for the subject may not be large, the distance to the lexical head verb indicating its argument status is on average longer for the errors than the correctly identified formal subjects.

The following picture of the difficulties in assigning the formal subject relation emerges: with no formal or structural clues available, the correct identification relies on lexical information regarding the head verb. In order for this information to be employed during parsing, it must represent a reliable source of information – being frequent, fairly unambiguous and available at attachment time.[166] It is also clear that the analysis of the logical subject relation (ES) relies largely on the correct analysis of the formal subject. Further confusion with the object and object adverbial functions is therefore to be expected as a result of error propagation.

[166]Recall that depending on the presence or absence of a logical subject (ES), the subject of one and the same predicate may be annotated as regular (SS) or formal (FS), respectively.

Error types for indirect objects IO		
Gold	System	#
IO	OO	97
IO	DT	31
IO	PA	6
IO	ROOT	3
IO	SS	2
IO	AT	2
IO	HD	1
IO	AA	1
IO	SP	1

Table 8.17: Error types for indirect objects.

8.3.4 Indirect object errors

The general trend in the parse results for indirect objects is that precision is high (93%), whereas recall is considerably lower (66%). This means that the parser has difficulties locating candidate indirect objects in general, and chooses an indirect object analysis only if there is clear evidence for it. This evidence, we may assume, is given in part by morphology. It is not surprising that indirect objects should be difficult to locate, as they occupy the postverbal position examined earlier, hence may be confused with both subjects and objects. However, person denoting indirect objects are marked by accusative case when expressed pronominally, a property which clearly sets them apart from subjects. Differentiation from direct objects, however, is more difficult, as table 8.17 clearly illustrates. We find that confusion with the direct object relation (OO) is the most frequent error type for indirect objects.

If we examine the instances which are correctly recognized as indirect objects by the parser, we find that the majority of these consist of case marked pronouns (83%) which to a large part are reflexive pronouns, see 8.3.1 above. With respect to word order, these sentences display the canonical ordering of arguments shown in (133)–(134) in section 8.3.1, to a large extent (87.5%). In these, we find either a preverbal subject or no realized subject at all in the case of subordinate clauses. The direct objects in the set of correct instances are most often realized by noun phrases, hence differ from the indirect object in this respect. Although both the factors of word order and part-of-speech can contribute towards differentiation from the direct object, we must assume that lexical properties of the verb are also important. After all, it is a property of the verb that it takes two objects. We find that a set of verbs are recurrent in the

correctly analyzed sentences, most notably the verbs *ge* 'give', *lära* 'teach', *skaffa* 'obtain', *fråga* 'ask', *tänka* 'think' which alone account for around 70% of the sentences. There is on average 5 indirect objects per verb type and only 8.9% of the verbs are hapax legomena.

The error types depicted in table 8.17 indicate that the main confusion for indirect objects is with the direct object function. The other prominent error type is confusion with the determiner function. As mentioned earlier, confusion with the direct object is not surprising due to the fact that these may occur in the same position. Confusion with the determiner function results from formal ambiguity – both pronouns and nouns may function as determiners. Since the following direct object is often a noun, the indirect object is analyzed as a determiner.

If we examine the errors performed by the parser along the same parameters as above - part-of-speech, word order and verbal properties, we find that the errors contain a lower proportion of pronominal indirect objects (61%), hence more nouns. Since indirect objects typically are pronominal and direct objects nouns, this is one property which contributes to the confusion by the parser. Also, there is a somewhat lower proportion of the unmarked word order pattern (70.8%), compared to the set of correctly analyzed indirect objects (87.5%). With a different ordering of arguments, e.g. postverbal subject, the confusion possibilities obviously multiply. The biggest difference from the set of correctly analyzed indirect objects can be found in the lexical heterogeneity of the verbal head. The same set of five ditransitive verbs which accounted for 70% of the correct sentences here only account for 33% of the sentences and the number of indirect object instances per verb type is 1.6. Half of all the verbs are hapax legomena, indicating that they have never been observed prior to parsing.

The above analysis shows that in the analysis of indirect objects, the main difficulty is in in distinguishing it from the direct object. This relies on the interplay of several factors. An unmarked word order, differences in nominal realization as well as an acquired generalization over the verb and its ditransitive argument structure are all factors which contribute to argument disambiguation.

8.3.5 Subject predicative errors

Table 8.18 shows the most frequent errors for the dependency relation of subject predicative (SP). We find that the most common error consists in the confusion of subject predicatives for subjects. This is not surprising, as these usually accompany each other and may be realized by the same types of con-

Error types for subject predicatives (SP)		
Gold	System	#
SP	SS	240
SP	AA	136
SP	ROOT	84
SP	OO	65
SP	OA	28
SP	CC	27
SP	DT	25
SP	AT	23
SP	KA	23
SP	PA	16

Table 8.18: Error types for subject predicatives.

stituents.[167] In parallel, we saw earlier that subjects are often confused for subject predicatives.

A clear majority (80.8%) of the SP_SS errors are nominal, i.e. either pronouns or nouns. More than half of these are in preverbal position, a distribution which clearly deviates from that of the corpus as a whole. We observe that the percentage of indefinite nominals is lower for these errors (53.6% of the nominal SP_SS errors are realized by an indefinite noun), a property which may be ascribed to their preverbal position. Clause-initial position is usually correlated with given information, hence also exhibits a greater tendency for definite expression. This property, however, makes these arguments difficult to parse correctly and, in particular, to distinguish from subjects. Subjects and nominal subject predicatives are notoriously difficult to differentiate, even for humans, as the annotation manual also makes clear (Teleman 1974: p. 59).

The second most common error type in the baseline analysis of subject predicatives concerns the adjectival predicatives in all majority. The confusion of subject predicatives with the regular adverbial function (AA) occurs first and foremost for adjectives and adverbs. These are almost exclusively postverbal (92.6% of the errors).

[167] Subject predicatives are, however, found without a corresponding subject in infinitival clauses.

Gold	System	#
OO	AA	103
ET	OO	91
AA	SP	82
AA	OO	76
TA	OO	76
OA	OO	71
OO	ET	58
OO	OA	57
ET	SS	56
TA	SS	53
SS	AA	45
OA	AG	38
AA	SS	37
SS	ET	36
KA	OO	35

Table 8.19: 15 most frequent argument/non-argument error types, where OO=object, AA=other adverbial, ET=nominal post-modifier, SP=subject predicative, TA=temporal adverbial, OA=object adverbial, SS=subject, AG=passive agent, KA=comparative adverbial.

8.3.6 Argument and non-argument errors

We have in the above sections focused largely on an error analysis of the argument relations and have found that confusion of the various argument relations is a common error. We mentioned initially in section 8.2.3 that confusion of non-argument relations, and in particular adverbials, is also a common error. In section 2.4, we noted that the distinction between arguments and non-arguments has been proposed to be gradient and probabilistic in nature (Manning 2003). The error analysis also shows that this distinction is not always straightforward, and we find error types where arguments are confused for non-arguments and vice versa.

In table 8.19 we find an overview of the 15 most frequent error types involving arguments and non-arguments. We find both error types where arguments are confused for non-arguments, e.g., OO_AA, OO_ET, SS_AA and, a somewhat larger group, of error types where non-arguments are confused for arguments, e.g., ET_OO, AA_SP, AA_OO, TA_OO etc. First of all, we may note that these error types are not nearly as common as those involving confusion within the groups of arguments and non-arguments. This indicates that the distinction

is acquired to a certain extent. Moreover, we find that the errors further support defining properties of the groups of arguments and non-arguments given the features employed by the parser. The errors confusing arguments for non-arguments largely involve categorially non-canonical arguments – adjectival and verbal objects and subjects, as in the OO_AA error in (165) below. In a parallel fashion, the non-arguments analyzed as arguments are predominantly nominal, as in the ET_OO and TA_SS errors in (166) and (167) below.

(165) *Samma sak gäller vuxna*
same case concerns adults
'The same goes for adults'

(166) *Den rätten har vi* **kvinnor** *haft sedan* ...
that right-DEF have we women had since ...
'We, the women, have had that right since ...'

(167) *Varje* **morgon** *åker tre miljoner människor* ...
every morning travel three million people ...
'Every morning, three million people travel ...'

With respect to word order, we find that preverbal, nominal adverbials are erroneously analyzed as subjects and postverbal adverbials as objects, indicating acquisition of word order preferences in line with our earlier findings.

8.3.7 Head distance

The overview of the distribution of the various argument relations in Scandinavian showed that they differ in their ordering preferences with respect to the verb. In the error analysis we have seen clear evidence for the acquisition of these preferences in the fact that the errors are largely instances which depart from the most frequent ordering. For subjects, for instance, we have seen that the set of errors contains notably less preverbal elements. Separate from the issue of ordering, however, is the issue of general distance to the head. Given the incremental, deterministic nature of our parser, we may assume that longer dependency arcs will be less accurate and more error-prone (McDonald and Nivre 2007).

In order to evaluate the distance factor with respect to the argument relations, we may compare head distance in the sets of correctly parsed arguments with the corresponding sets of errors. Figure 12 shows the proportions of *adjacent* (± 1) and *close* ($\pm 1, 2, 3$) dependents with respect to the head in the sets of correct as opposed to errors for the subject (SS), formal subject (FS),

direct (OO) and indirect (IO) objects, as well as subject predicatives (SP) dependency relations. First of all, it is clear that the argument relations differ in

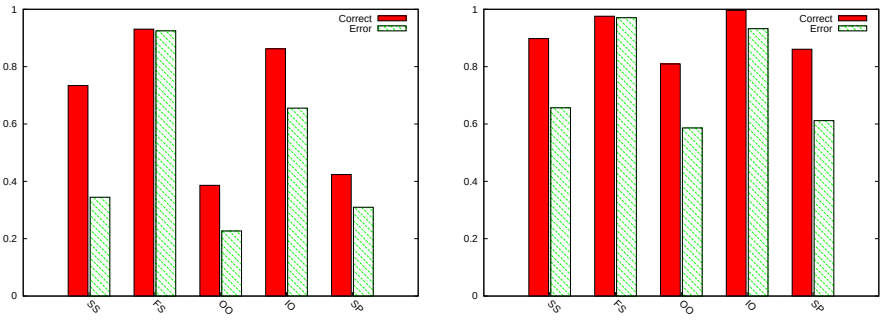

Figure 12: Proportion of adjacent (± 1) dependents (left) and close ($\pm 1, 2, 3$) dependents (right) in correct versus error sets for argument relations.

their preferences with respect to distance to the head, as seen by the general height of the bars. Note however, that since we measure distance in terms of linear position, distance preferences will also be interfused with preference for *short* expression. For instance, subjects (SS, FS) and indirect objects (IO) have a preference for pronominal realization and are also shown to be highly local. It is not surprising then, that subjects (SS, FS) and indirect objects (IO) show higher proportions of adjacent dependents than direct objects. More importantly however, we observe that the sets of correctly and erroneously analyzed arguments clearly differ in proportions of adjacent and close dependents. We find that their proportion is notably higher in the sets of correct instances for all relations except the formal subjects.[168] The difference is most clear in the case of adjacent dependents in figure 12, but the same tendency is also present in figure 12, where we define the set of close dependents to be within a linear distance of $\pm 1, 2, 3$ from the head.

8.4 Setting the scene

This chapter has dealt with argument disambiguation in Swedish. We have introduced data-driven dependency parsing and argued that it provides a framework for studying the effect of frequency-derived constraints in argument differentiation. Data-driven parsing has the advantage that syntactic analysis is directly conditioned on properties of the data, and, in particular, on frequency of language use. Dependency analysis provides a framework which maintains

[168] Recall that formal subjects are realized as pronouns and positioned either immediately preceding or following the finite verb.

a separate level of grammatical functions, which enables the acquisition of linguistic constraints on grammatical functions, rather than structural position. The system employed for data-driven dependency parsing, MaltParser, makes it possible to constrain the information employed during analysis, hence providing an experimental setting where we may study the effect of different factors.

With the general aim of studying the generalizations which may be acquired during data-driven dependency parsing of Swedish, we have argued that an in-depth *error analysis* can provide some of the answers. The errors may be characterized by properties like part-of-speech and dependency relation into error types. By comparing tendencies in sets of correctly and erroneously analyzed arguments, we have approached a characterization of *argument errors*. In particular, we have seen that confusion of argument relations are a frequent type of error, for instance the error types SS_OO, FS_SS, IO_OO discussed above. Based purely on frequency in the training data, a range of generalizations regarding the structural and formal preferences of the argument relations have been shown to be reliably acquired. Some of these acquired *generalizations* include:

- canonical ordering of arguments, e.g. subjects typically precede the verb and objects follow.
- verb second (V2)
- lexical/formal preferences:
  - case preferences for pronouns, e.g. indirect objects are accusative
  - verbal subcategorization, e.g. ditransitive verbs, existential predicates
- categorial preferences:
  - the core argument relations are nominal
  - tendencies in referentiality, e.g. subjects and indirect objects are also more likely to be pronominal than direct objects
- main distinction between arguments and non-arguments

The error analysis has also made clear that morphology and word order do not provide sufficient evidence for *argument disambiguation* in all cases. The errors in argument assignment may often be attributed to global or local ambiguities caused by word order variation and lack of morphological marking. We have also seen that degree of head locality is a factor in the remaining

errors. Clearly, local properties of the arguments are even more important in an incremental, deterministic setting. The above error analysis thus sets the scene for further investigations into argument differentiation through a set of experiments where the influence of linguistic features discussed earlier in this thesis may be explicitly evaluated in terms of argument disambiguation in a data-driven dependency parser.

9 PARSING WITH LINGUISTIC FEATURES

Despite the dramatic improvement in accuracy for data-driven parsers in recent years, we still have relatively little knowledge about the exact influence of data-derived features on the parsing accuracy for specific linguistic constructions. There are a number of studies that investigate the influence of different features or representational choices on overall parsing accuracy, within a variety of different frameworks, (Bod 1998; Megyesi 2002; Klein and Manning 2003; Bikel 2004; Charniak and Johnson 2005). There are also attempts at a more fine-grained analysis of accuracy, targeting specific linguistic constructions or grammatical functions (Buchholz 2002; Carroll and Briscoe 2002; Kübler and Prokić 2006). But there are few studies that combine the two perspectives and try to tease apart the influence of different features on the analysis of specific constructions, let alone motivated by a thorough linguistic analysis.

In this chapter, we present an in-depth study of the influence of certain linguistic features, such as animacy, definiteness, and finiteness, on the parsing accuracy for argument relations. In chapter 8, we saw that characteristics of argument realization in Scandinavian type languages pose special problems for the identification of argument relations due to limited case marking and ambiguous word order patterns. In the following we will experiment with the addition of morphosyntactic and lexical semantic features that approximate the distinguishing properties of the argument functions discussed in chapter 3. We will isolate features of the arguments and the verbal head, as well as combinations of these, and evaluate their effect on overall parsing results as well as on argument disambiguation specifically.[169] We will address the following questions:

Linguistic features How will a set of linguistic features expressing inherent properties of arguments affect argument disambiguation? How will linguistic features expressing morphological and semantic properties of the verb affect argument disambiguation?

[169] A shorter version of the first experiments is found in Øvrelid and Nivre 2007.

Parser To what extent is argument disambiguation dependent on specific properties of the parser?

Scalability Are the results scalable? May the linguistic features be acquired automatically?

9.1 Linguistic features

Argument differentiation depends on a range of linguistic dimensions, some of which are recurrent in a range of languages and others the result of more language-specific properties of syntax and morphology. In order to test the influence of these features in argument differentiation, we must provide empirical approximations of these dimensions which may be derived from a corpus.

In chapter 3, we examined linguistic dimensions which have been claimed to influence argument differentiation across a range of languages and in different types of linguistic studies. In particular, we examined animacy, definiteness and referentiality in detail. The close correlation between *animacy* and different distinctions in argumenthood has been the subject of large parts of the previous chapters, hence need not be repeated at length here. In Part II we found that syntactic distribution provided a reliable indicator of animacy. In the present context we may also make use of the systematic influence of animacy on arguments and examine the effect of animacy in argument disambiguation. The dimension of *definiteness* expresses the extent to which a referent is identifiable and unique. It is thus concerned with the status of the referent, either in the linguistic discourse or in terms of cognitive status. With respect to argument differentiation, definiteness is in particular important in distinguishing the external argument, the subject, from other arguments such as objects and subject predicatives. The dimension of *referentiality* expresses the way in which reference is determined for a linguistic expression, or rather, the extent to which the determination of reference relies on the linguistic context. A highly referential element may be referred to by solely relying on context and may therefore be referred to with a pronoun, whereas a common noun relies to a greater extent on lexical or denotational semantic knowledge. Referentiality is a factor in argument differentiation and in section 8.3.1 we saw that subjects are more likely to be expressed pronominally than objects. Differentiation within the group of objects, i.e. between direct and indirect objects, is also influenced by referentiality.

Chapter 4 and section 8.3.1 outlined some important properties of Scandinavian morphosyntax. We have seen that the structural expression of arguments in Scandinavian is characterized by initial variation along with rigid verb

Linguistic feature	Treebank feature
animacy	person reference
definiteness	morph. definiteness
referentiality	pronoun type, part-of-speech
finiteness	tense
case	morph. case

Table 9.1: Linguistic features and their empirical counterparts.

placement. Recall that the V2 constraint requires that the finite verb be the second constituent of declarative main clauses and *finiteness* has been claimed to be a defining property of Scandinavian syntax (Holmberg and Platzack 1995; Eide 2008). Even if the morphological marking of arguments in Scandinavian is not extensive or unambiguous, *case* may distinguish arguments when expressed pronominally.

9.1.1 Empirical approximations

In table 9.1 we find an overview of the linguistic dimensions discussed above with their corresponding treebank feature. It distinguishes between the features discussed in chapter 3, representing soft, cross-linguistic tendencies in argument differentiation, and the more language-specific features of Scandinavian discussed in chapter 4. We map the linguistic features to a set of empirical features representing information which is found in the annotation of the Talbanken05 treebank.

Recall that the Talbanken05 treebank explicitly distinguishes between person- and non-person referring nominal elements, a distinction which overlaps fairly well with the traditional notion of animacy. See section 7.1.2 for a detailed overview of the information on person reference in Talbanken05. Morphological definiteness is marked for all common nouns in Talbanken05; definite nouns are marked as DD and indefinite nouns are unmarked (Ø). Talbanken05 contains morphological case annotation for pronouns which distinguishes between nominative (Ø) or accusative case (AA). Common nouns distinguish nominative (Ø) and genitive (GG) case. The morphosyntactic features which are expressed for the part-of-speech of verb in Talbanken are tense (present, past, imperative, past/present subjunctive, infinitive and supine) and voice (Ø/passive; PA).

Pronouns are furthermore annotated with a set of pronominal classes which distinguish between e.g. 1st/2nd person and 3rd person pronouns, reflexive,

reciprocal, interrogative, impersonal pronouns etc. For the third person neuter pronoun *det* 'it' and demonstrative *detta* 'this', the annotation in Talbanken05 distinguishes between an impersonal and a personal or "definite" (DP) usage. The impersonal class includes expletives, as well as pronouns which refer to a preceding clause, as in (168) below.

(168) **Det** *skall Bayless hålla reda på*
that shall Bayless hold order on
'Bayless will keep track of that'

The impersonal pronominal class is employed for *non-referential* pronouns.[170] The two classes of pronouns have quite distinct syntactic behaviours. The impersonal pronouns never function as determiners (DT), whereas the definite pronouns often do (71.4%). Also, the impersonal pronouns are more likely to function as formal subjects FS (32.4%) than the definite pronoun (1.1%).[171]

9.2 Experiments with linguistic features

The experiments are aimed at investigating argument differentiation in a disambiguation task where frequency drives analysis. There is no explicit formulation of constraints, rather syntactic analysis is constrained directly by frequency of language use.

In chapter 8 we found that a set of features expressing only word form, part-of-speech and preceding analysis served to guide the acquisition of general patterns of argument realization discussed in chapter 4. However, the observed errors in argument assignment were caused in part by ambiguities precisely in lexical form, morphological marking and word order patterns. We will here examine the effect of additional differentiating properties of arguments.[172]

9.2.1 Experimental methodology

The main goal of the experiments is to evaluate the effect of the linguistic features discussed in section 9.1 on argument disambiguation. The experimental setup should therefore enable us to isolate the effects of different fea-

[170] Note that we here employ 'referential' in a narrow sense, which only includes reference to entities. The category of 'non-referential pronouns' consequently includes pronouns which do not refer, i.e., expletives, as well as pronouns which refer to propositions.

[171] The fact that a few definite/referential pronouns are annotated as formal subjects (15 instances, 1.1%) and that a few impersonal pronouns function as determiners (11 instances, 0.4%) must be assumed to stem from annotation error.

[172] All examples in the current chapter are taken from the written sections of Talbanken05.

	FORM	POS	DEP	FEATS
S:*top*	+	+	+	+
S:*top*+1		+		
I:*next*	+	+		+
I:*next*−1	+			+
I:*next*+1	+	+		+
I:*next*+2		+		
G: head of *top*	+			+
G: leftmost dependent of *top*			+	
G: rightmost dependent of *top*			+	
G: leftmost dependent of *next*	+		+	+
G: leftmost dependent of head of *top*			+	
G: leftmost sibling of rightmost dependent of *top*			+	
G: rightmost sibling of leftmost dependent of *top*	+			+
G: rightmost sibling of leftmost dependent of *next*		+	+	

Figure 13: Extended (FEATS) feature model for Swedish; S: stack, I: input, G: graph; $\pm n = n$ positions to the left($-$) or right ($+$).

tures, as well as evaluate and compare them. We take the parser evaluated in chapter 8 as our baseline system against which we compare and quantify improvement/deterioration in results. All experiments are performed using 10-fold cross-validation for training and testing on the entire written part of Talbanken05.

Recall from section 8.1.3 that the feature model of MaltParser defines the attributes employed to describe the parse configurations at each point during parsing. In order to incorporate information on our linguistic features we therefore extend the feature model with an additional, static attribute, FEATS. The extended version of the feature model is depicted in figure 13, including all four columns.[173] What is varied in the experiments is thus only the information contained in the FEATS features (animacy, definiteness, etc.), while the tokens for which these features are defined remain constant. This provides a controlled setting for the testing of our linguistic features. Note that the FEATS features added in this way, like the POS features inherited from the baseline parser, are initially taken from the gold standard annotation in the treebank, which means that the results may give an over-optimistic view of the accuracy that can be expected when parsing new text. We will return to this point later in this chapter.

[173]Preliminary experiments showed that it was better to tie FEATS features to the same tokens as FORM features (rather than POS or DEP features). Backward selection from this model was tried for several different instantiations of FEATS but with no significant improvement.

In terms of evaluation we wish to be able to quantify the effect of the added features, as well as perform more in-depth error analyses. Evaluation will be performed at the different levels of analysis established in section 8.2.1 – overall accuracy employing labeled/unlabeled attachment scores, performance per dependency relation, as well as overview in terms of error types. Statistical significance is checked using Dan Bikel's randomized parsing evaluation comparator.[174] [175] Since the main focus is on argument analysis, significance testing and error analysis will focus on labeled accuracy, unless otherwise stated. We report accuracy for specific dependency relations, measured as a balanced F-score. In order to summarize improvement with respect to dependency relation assignment when comparing two parsers, we rank the relations by their frequency-weighted difference of F-scores.[176]

In the error analysis in chapter 8, we examined various error types in terms of confusion classes of dependency relations. We will employ two different comparative measures to compare parsers with respect to specific error types: (i) the difference in total number of errors of a certain type for the compared parsers, and (ii) the number of corrected or newly added errors. In set-theoretic terms, with respect to the set of errors for a baseline parser P_{Bl} and a new parser P_N, the two measures are defined as follows:

(i) $\mid P_{Bl} \mid - \mid P_N \mid$

(ii) $\mid P_{Bl} - P_N \mid$ or $\mid P_N - P_{Bl} \mid$

Whereas the former measure compares the overall tendency of a parser to make a certain type of error, the latter allows us to compare the parsers' performance for the same set of errors that were examined during the initial error analysis and specifically targeted in the experiments. Examining sets of corrected or newly added errors provides us with more detailed information regarding the effect of the added information. Total number of errors are presented in the form of confusion matrices, see, e.g., table 9.4 on page 215, where we give the total number of occurrences of each error type for the baseline parser, together

[174] http://www.cis.upenn.edu/∼dbikel/software.html

[175] The main idea in randomized parsing evaluation is that given a null hypothesis of no difference between two sets of results, shuffling the results from one system with those of the other should produce a difference in overall results equal to or greater than the original difference, since the individual scores then should be equally likely. If the performance between two sets differ significantly, on the other hand, the shuffling of the predictions will very infrequently lead to a larger performance difference. The shuffling is iterated 10,000 times and the total number of differences in results equal to or larger than the original is recorded. The relative frequency of the number of differences is then interpreted as significance of the difference.

[176] For each dependency relation, the difference in F-scores is weighted by the relative frequency of the dependency relation, $\frac{Deprel}{\sum_i Deprel_i}$, in the treebank.

	Unlabeled	Labeled
NoFeats	89.87	84.92
Anim	89.93	85.10
Def	89.87	85.02
Pro	89.91	85.04
Case	89.99	85.13
Verb	90.15	85.28
ADPC	90.17	85.45
ADPCV	90.42	85.73
All	90.73	86.32

Table 9.2: Overall results expressed as unlabeled and labeled attachment scores.

with the percentage of each error type out of all errors for the dependency relation. For the extended parsers, we give total numbers (#) along with the relative improvement compared to the baseline (%).

9.2.2 Animacy

As table 9.2 shows, the addition of information on animacy for nominal elements causes an improvement in overall results ($p<.0002$). We find that the added information has the greatest effect on the *labeling* of dependency results, rather than attachment.[177] The improvement may be summarized per dependency relation as in table 9.3 where the dependency relations are ranked by their frequency-weighted difference of balanced F-scores in order to indicate their relative impact in the improvement of the parse results.

The subject and object functions are the dependency relations whose assignment improves the most when animacy information is added. We also find a small improvement for indirect objects, F-scores improve from 77.2 to 78.8.[178] Furthermore, there is an effect in accuracy for a range of other functions where animacy is not directly relevant, but where the improved analysis of arguments contributes towards correct identification, e.g., adverbials and determiners.

If we take a closer look at the individual error types involving subjects and objects in table 9.4, we find that the addition causes a reduction of errors confusing subjects with objects (SS_OO), determiners (SS_DT) and subject pred-

[177] The difference in unlabeled results is not statistically significant.

[178] Since indirect objects are quite infrequent in the treebank, this dependency relation is not included in the ranked list in 9.3.

	Freq	NoFeats	Anim
SS	0.1105	90.25	90.81
OO	0.0632	84.53	85.04
DT	0.1081	94.14	94.48
TA	0.0249	70.29	71.07
PA	0.1043	94.69	94.81
CC	0.0343	78.02	78.34
++	0.0422	90.33	90.52
OA	0.0305	70.63	70.84
FO	0.0009	56.68	63.81
AT	0.0441	95.76	95.90

	Freq	NoFeats	Def
SP	0.0297	84.82	85.59
OO	0.0632	84.53	84.84
DT	0.1081	94.14	94.30
SS	0.1105	90.25	90.38
AA	0.0537	68.70	68.91
PA	0.1043	94.69	94.77
TA	0.0249	70.29	70.57
AN	0.0057	39.42	40.64
+F	0.0099	52.07	52.64
UK	0.0305	93.17	93.30

Table 9.3: 10 most improved dependency relations with added information on animacy (left) and definiteness (right), ranked by their weighted difference of balanced F-scores.

icatives (SS_SP) – all functions which do not exhibit the same preference for human reference as subjects. For the SS_SS error type, we do not find an improvement.[179]

The set of corrected errors shows an effect of acquired animacy preferences. For instance, all corrected indirect objects (20 instances, 13.5% of the baseline errors) are human. The added information corrects 15.6% of the baseline errors for the SS relation and 60% of these are refer to humans. The influence of the animacy information is clear also if we examine the individual error types. For subjects, we find an improved disambiguation from non-argument relations such as adverbials and determiners, as mentioned above. For instance, the corrected errors of the SS_AA error type, which account for 28.9% of the baseline errors, are all human, and the corrected errors for the SS_DT error type (21.4%) show a a clear majority of 84.3% human elements. With respect to the confusion of argument relations, we also observe the influence of the added information in the error sets. The percentage of corrected errors for the SS_OO and OO_SS error types, compared to the baseline parser, are 21.1% and 28.5%, respectively, when adding information on animacy. Whereas 57.4% of the corrected subjects of this error type are human, only 19.3% of the corrected objects are.

[179]Labeled attachment scores require both attachment and labeling to be correct, hence we find error types like SS_SS, where only the head attachment is incorrect.

Confusion matrix for subjects (SS)

	NoFeats		Anim		Def		Pro		Case		Verb		ADPC		ADPCV		All	
sys	#	% tot.	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
OO	446	23.0	388	13.0	425	4.7	401	10.1	419	6.1	365	18.2	361	19.1	293	34.3	296	33.6
ROOT	265	13.7	270	-1.9	284	-7.2	275	-3.8	277	-4.5	260	1.9	269	-1.5	266	-0.4	241	9.1
DT	238	12.3	196	17.6	230	3.4	218	8.4	205	13.9	239	-0.4	164	31.1	160	32.8	160	32.8
SS	216	11.1	222	-2.8	214	0.9	202	6.5	198	8.3	161	25.5	217	-0.5	166	23.1	153	29.2
SP	206	10.6	203	1.5	187	9.2	198	3.9	201	2.4	216	-4.9	188	8.7	187	9.2	195	5.3
CC	137	7.1	135	1.5	123	10.2	139	-1.5	139	-1.5	122	10.9	120	12.4	114	16.8	98	28.5
FS	133	6.9	141	-6.0	148	-11.3	148	-11.3	154	-15.8	151	-13.5	147	-10.5	153	-15.0	155	-16.5
PA	53	2.7	53	0.0	43	18.9	43	18.9	37	30.2	49	7.5	25	52.8	22	58.5	26	50.9
…	…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Confusion matrix for objects (OO)

	NoFeats		Anim		Def		Pro		Case		Verb		ADPC		ADPCV		All	
sys	#	% tot.	#	%)	#	%)	#	%)	#	%)	#	%)	#	%)	#	%)	#	%)
SS	309	21.3	263	14.9	288	6.8	280	9.4	273	11.7	259	16.2	251	18.8	215	30.4	212	31.4
ROOT	221	15.2	239	-8.1	224	-1.4	237	-7.2	229	-3.6	218	1.4	251	-13.6	245	-10.9	241	-9.0
OO	149	10.3	153	-2.7	151	-1.3	148	0.7	146	2.0	143	4.0	143	4.0	141	5.4	141	5.4
PA	126	8.7	122	3.2	129	-2.4	123	2.4	112	11.1	123	2.4	111	11.9	109	13.5	105	16.7
AA	103	7.1	94	8.7	97	5.8	92	10.7	106	-2.9	102	1.0	96	6.8	95	7.8	74	28.2
DT	99	6.8	95	4.0	94	5.1	99	0.0	85	14.1	99	0.0	81	18.2	70	29.3	72	27.3
ET	58	4.0	54	6.9	61	-5.2	57	1.7	59	-1.7	64	-10.3	49	15.5	49	15.5	49	15.5
OA	57	3.9	59	-3.5	58	-1.8	58	-1.8	57	0.0	65	-14.0	63	-10.5	66	-15.8	64	-12.3
…	…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 9.4: Confusion matrices for the assignment of the subject and object dependency relations for the baseline parser (columns 2–3) and for the extended feature models (columns 4–11).

9.2.3 Definiteness

The addition of information on definiteness during parsing causes a significant improvement of overall results ($p<.02$). The dependency relation for which we observe the largest improvement is the subject predicative relation (SP), as shown in table 9.3.

As we recall from section 8.3, subject predicatives are often confused with subjects, see table 9.4, and vice versa, see table 9.5. Predicatives in Swedish usually stand in a classifying relation to the subject, where the subject is established as being an instance of a class of some kind. As a consequence, the predicative is often denoted by a nominal expressing generic reference and typically realized by an indefinite noun phrase. If we examine the set of corrected errors compared to the baseline, we find that the added information causes a 14.2% reduction of the SP_SS errors, all of which are indefinite nouns.

We furthermore observe an improved performance in the analysis of indirect objects (IO) with an F-score improvement from 77.2 to 78.7, see the

Confusion matrix for subject predicatives (SP)																		
	NoFeats		Anim		Def		Pro		Case		Verb		ADPC		ADPCV		All	
sys	#	% tot.	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
SS	240	30.0	231	3.8	229	4.6	237	1.2	231	3.8	240	0.0	213	11.2	213	11.2	208	13.3
AA	136	17.0	140	-2.9	123	9.6	126	7.4	127	6.6	131	3.7	127	6.6	129	5.1	131	3.7
ROOT	84	10.5	86	-2.4	83	1.2	85	-1.2	80	4.8	86	-2.4	82	2.4	77	8.3	76	9.5
OO	65	8.1	71	-9.2	63	3.1	73	-12.3	66	-1.5	72	-10.8	70	-7.7	64	1.5	67	-3.1
SP	34	4.2	35	-2.9	32	5.9	34	0.0	35	-2.9	30	11.8	35	-2.9	32	5.9	30	11.8
OA	28	3.5	25	10.7	25	10.7	25	10.7	27	3.6	27	3.6	28	0.0	26	7.1	31	-10.7
CC	27	3.4	25	7.4	29	-7.4	27	0.0	27	0.0	25	7.4	27	0.0	25	7.4	11	59.3
DT	25	3.1	29	-16.0	24	4.0	22	12.0	25	0.0	28	-12.0	20	20.0	22	12.0	21	16.0
AT	23	2.9	22	4.3	20	13.0	24	-4.3	22	4.3	24	-4.3	20	13.0	22	4.3	22	4.3
KA	23	2.9	17	26.1	22	4.3	21	8.7	26	-13.0	21	8.7	21	8.7	21	8.7	19	17.4
PA	16	2.0	16	0.0	14	12.5	16	0.0	9	43.8	17	-6.2	8	50.0	7	56.2	9	43.8
…	…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 9.5: Confusion matrix for the assignment of the subject predicative dependency relation for the baseline parser (columns 2–3) and for the extended feature models (columns 4–11).

Confusion matrix for indirect objects (IO)																		
	NoFeats		Anim		Def		Pro		Case		Verb		ADPC		ADPCV		All	
sys	#	% tot.	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
OO	97	65.5	95	2.1	85	12.4	94	3.1	94	3.1	97	0.0	89	8.2	96	1.0	97	0.0
DT	31	20.9	31	0.0	33	-6.5	33	-6.5	23	25.8	33	-6.5	21	32.3	25	19.4	19	38.7
PA	6	4.1	5	16.7	5	16.7	5	16.7	6	0.0	5	16.7	5	16.7	5	16.7	5	16.7
IO	4	2.7	4	0.0	4	0.0	3	25.0	4	0.0	4	0.0	4	0.0	4	0.0	4	0.0
ROOT	3	2.0	0	0.0	2	33.3	1	66.7	3	0.0	3	0.0	1	66.7	1	66.7	0	0.0
AT	2	1.4	1	50.0	2	0.0	2	0.0	2	0.0	2	0.0	1	50.0	1	50.0	1	50.0
SS	2	1.4	1	50.0	3	-50.0	3	-50.0	2	0.0	3	-50.0	2	0.0	3	-50.0	3	-50.0
AA	1	0.7	0	0.0	1	0.0	1	0.0	1	0.0	1	0.0	0	0.0	1	0.0	0	0.0
HD	1	0.7	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0
SP	1	0.7	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0

Table 9.6: Confusion matrix for the assignment of the indirect object dependency relation for the baseline parser (columns 2–3) and for the extended feature models (columns 4–11); shows all errors.

confusion matrix for this dependency relation in table 9.6. As mentioned in chapter 3, the two objects in a double object construction are typically differentiated by several factors, among which definiteness has been shown to be one (Bresnan et al. 2005).

9.2.4 Pronoun type

The addition of pronoun type information causes a general improvement in overall parsing results ($p<.01$), as we can see from table 9.2. The dependency

	Freq	NoFeats	Pro
SS	0.1105	90.25	90.66
OO	0.0632	84.53	84.99
FS	0.0050	71.31	73.99
PA	0.1043	94.69	94.78
FO	0.0009	56.68	66.18
SP	0.0297	84.82	85.08
AA	0.0537	68.70	68.84
TA	0.0249	70.29	70.59
+F	0.0099	52.07	52.80
UK	0.0305	93.17	93.40

	Freq	NoFeats	NonRef
SS	0.1105	90.25	90.61
OO	0.0632	84.53	84.96
TA	0.0249	70.29	71.02
FS	0.0050	71.31	74.22
UK	0.0305	93.17	93.46
AN	0.0057	39.42	40.53
ES	0.0050	71.82	72.80
MA	0.0091	76.14	76.57
ROOT	0.0649	86.71	86.77
FO	0.0009	56.68	60.32

Table 9.7: 10 most improved dependency relations with added information on pronominal class (left) and non-referentiality (right), ranked by their weighted difference of balanced F-scores.

relations whose assignment improves the most are, once again, the core argument functions (SS, OO), see table 9.7. We also find a general improvement in terms of recall for the assignment of the formal subject (FS) and object (FO) functions, which are both realized by the third person neuter pronoun *det* 'it', annotated as non-referential in the treebank.

9.2.4.1 Non-referential pronouns

In a separate experiment (NonRef), we isolated the property of non-referentiality from the other pronominal classes. In this experiment, only information regarding non-referentiality was included as an additional feature during parsing. The results were slightly lower than the experiment with all information on pronominal class, but not significantly so and these results also differ significantly from those of the baseline parser ($p<.01$).

We would expect information on non-referentiality to be beneficial in the disambiguation of regular, referential subjects (SS) and formal subjects (FS). The error analysis in 8.3 showed that these are difficult to distinguish by form or word order alone and we found that verbal form was the main indicator for the baseline parser. Moreover, and as a consequence of an improved analysis of formal and regular subjects, we may expect improvement in the disambiguation of logical subjects (ES) and direct objects (OO).

If we examine table 9.7, we find that isolation of non-referentiality has a clear effect on the analysis of the SS, OO and FS relations. In fact, performance for the FS relation is slightly better in the NonRef experiment than in the ex-

Confusion matrix for formal subjects (FS)																		
	NoFeats		Anim		Def		NonRef		Case		Verb		ADPC		ADPCV		All	
sys	#	% tot.	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
SS	281	91.8	272	3.2	277	1.4	247	12.1	277	1.4	275	2.1	241	14.2	244	13.2	252	10.3
DT	10	3.3	7	30.0	12	-20.0	5	50.0	9	10.0	8	20.0	7	30.0	7	30.0	3	70.0
OO	6	2.0	5	16.7	7	-16.7	5	16.7	5	16.7	3	50.0	5	16.7	3	50.0	5	16.7
FS	4	1.3	7	-75.0	4	0.0	3	25.0	5	-25.0	3	25.0	3	25.0	4	0.0	3	25.0
ROOT	3	1.0	5	-66.7	4	-33.3	2	33.3	4	-33.3	3	0.0	3	0.0	2	33.3	2	33.3
KA	1	0.3	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0	1	0.0
FO	1	0.3	1	0.0	1	0.0	1	0.0	1	0.0	0	0.0	1	0.0	1	0.0	0	0.0

Table 9.8: Confusion matrix for the assignment of the formal subject dependency relation for the baseline parser (columns 2–3) and for the extended feature models (columns 4–11); shows all errors.

periment where all pronominal features were included (Pro). The identification of ES also improves. Note, however, that there is no direct mapping between non-referentiality and status as a formal subject. As mentioned earlier in section 8.3.3, non-referential pronouns in Talbanken05 are annotated as subjects when they are not linked to another argument, i.e., the logical subject. In fact, the most common dependency relation for the non-referential pronouns is SS (48.7%) and not FS (31.4%). Even so, it is clear that the added information contributes towards an improved recognition of the formal subject relation. Table 9.8 shows a confusion matrix for the FS relation, where the results for the NonRef feature are displayed in column 6. We find a reduction of total number of errors of 12.1% for the FS_SS error type, compared to 10.7% with all pronominal features and the set of corrected FS_SS errors are all non-referential. So, even though non-referentiality does not constitute unequivocal evidence for a formal subject analysis, it contributes important information along with the other available features, such as verb form. For the corrected SS_FS errors, a clear majority of these (70%) are referential, as the example in (169) below, where *det* 'it' refers to a narcotic substance:

(169) **Det** *visade sig vara vanebildande*
it showed itself be addictive
'It turned out to be addictive'

9.2.5 Case

When we employ case information during parsing we find a clear improvement in results ($p<.0001$). However, the improvement is not first and foremost caused by improvement in assignment of subjects and objects, but rather,

	Freq	NoFeats	Case
DT	0.1081	94.14	94.71
PA	0.1043	94.69	95.13
SS	0.1105	90.25	90.61
OO	0.0632	84.53	85.11
TA	0.0249	70.29	71.16
SP	0.0297	84.82	85.20
+F	0.0099	52.07	52.70
AN	0.0057	39.42	40.35
VG	0.0302	94.65	94.81
IO	0.0024	76.14	77.88

	Freq	NoFeats	Verb
SS	0.1105	90.25	90.88
VG	0.0302	94.65	96.61
ROOT	0.0649	86.71	87.61
OO	0.0632	84.53	85.45
+F	0.0099	52.07	55.08
MS	0.0096	63.35	66.43
UK	0.0305	93.17	93.70
++	0.0422	90.33	90.67
AG	0.0019	73.56	80.64
AN	0.0057	39.42	41.69

Table 9.9: 10 most improved dependency relations with added information on case (left) and verb (right), ranked by their weighted difference of balanced F-scores.

the assignment of determiners and prepositional complements, see table 9.9. The error analysis in section 8.3 showed evidence that pronominal case preferences were acquired through lexical form. However, the error analysis also noted ambiguities between status as phrasal modifier as opposed to phrasal head, e.g. SS_DT, OO_PA. Nouns in Swedish are inflected for genitive case and may then serve as determiners for other nouns. Knowledge that a noun is in genitive case in theory excludes it from having an argument relation, be it a clausal argument relation like subject and object, or a phrasal argument, such as prepositional complements (PA).

We find a clear effect of genitive case marking in the improved results. For the determiner relation we find improvements in total number of errors for error types indicating confusion with a range of clausal and phrasal argument relations, e.g. DT_PA (21.1%), DT_SS (17.1%). Clear majorities of the corrected errors compared are in genitive case; for instance, 79.3% of the DT_PA errors headed by a noun are in genitive case. We also observe an improvement in the total error counts of 25.5% for the converse PA_DT error type and find that all the corrected PA_DT errors are non-genitive, as in (170):

(170) *I dagens äktenskap accepterar de flesta kvinnor inte*
in todays marriage accept the most women not
utan **protest** *rollen som en undergiven maka*
without protest role-DEF as a subordinate spouse
'In modern marriages most women do not accept a role as a subordinate spouse without protest'

We may summarize, then, that case information has an effect mostly on ambiguities between phrasal head relations, e.g., PA, SS, OO, and modifier relations, e.g., DT, with positive effects for the analysis of both types of dependency relations.

9.2.6 Verbal features

In this experiment, all information available for the verbal category (Verb), i.e. voice and tense, was included during parsing. The addition of morphosyntactic information for verbs causes a clear improvement in overall results ($p<.0001$), shown in table 9.2.

Table 9.9 shows the top ten improved dependency relations with added morphosyntactic information for verbs. The added information has a positive effect on the verbal dependency relations – ROOT, MS, VG, as well as an overall effect on the assignment of the SS and OO argument relations. Information on voice also benefits the relation expressing the demoted agent (AG) in passive constructions, headed by the preposition *av* 'by', as in English.

The overview of the most common error types for the SS and OO relations, see confusion matrices in table 9.4, indicates that the addition of information on verbal features improves on the confusion of the main argument types – SS_OO, OO_SS, as well as SS_FS. We also find that head attachment of subjects (SS_SS) in particular improves. We know that the subject is always attached to the finite verb in the Talbanken05 analysis, so this is not surprising.

If we examine the set of baseline errors for the SS_OO and OO_SS error types, we find that 33.2% (SS_OO) and 37.2% (OO_SS) of these have been corrected with added verbal features. For the SS_OO errors, the corrected cases are almost exclusively postverbal subjects which all follow a finite head verb, as in (152) above and repeated here as (171). The OO_SS errors are also almost exclusively postverbal and a fair number of these (37%) have a non-finite head verb, as in (160), repeated here (172). As we remember, only objects may follow a non-finite verb. Also, among the corrected objects with a finite head, we find quite a few imperative forms, as in (159), repeated as (173).

(171) *Samma erfarenhet gjorde* **engelsmännen**
same experience made englishmen-DEF
'The same experience, the Englishmen had'

(172) *...om man tidigare varit gul i ögonen, haft* **gulsot**
...if one earlier been yellow in eyes-DEF, had jaundice
eller ...
or ...
'...if one earlier has had yellow eyes, had jaundice or ...'

	Unlabeled	Labeled
NoFeats	89.87	84.92
Verb	90.15	85.28
Voice	89.81	84.97
Tense	90.15	85.27
Finite	90.24	85.33

Table 9.10: Overall results for experiments with verbal features, expressed as unlabeled and labeled attachment scores.

(173) *Glöm aldrig det* **löfte** *om trohet för livet*
forget never that promise of faithfulness for life-DEF
'Never forget that promise of faithfulness for life'

The verbal properties then, are beneficial for the disambiguation of the core argument functions, in addition to aiding the correct identification of several verbal dependency relations.

9.2.6.1 Individual verbal features

In order to tease apart the influence of the various verbal features we also performed a set of experiments testing individual sets of verbal features. Three experiments were run with differing feature sets: only voice information (Voice), only tense information (Tense) and a final experiment where the categories in the tense feature were mapped to a binary distinction between finite and non-finite verb forms (Finite). The last experiment was performed in order to test explicitly for the effect of the finiteness of the verb.

Voice

The addition of information on voice in isolation has little effect on the results and the overall difference from the baseline is not statistically significant. This is somewhat surprising as voice alternations have such confounding effects on the argument structure and argument realization of a verb.

In Swedish, the passive may be expressed by a passive suffix -s on the verb, as in (174) below, or periphrastically (*be/become* + passive participle), as in (175):

(174) *I krig* **utförs** *all verksamhet i 'skarpladdad miljö'*
in war perform-PASS all business in 'sharploaded environment'
'During war all business is performed in a heavily loaded environment'

	Freq	NoFeats	Voice
OO	0.0632	84.53	85.29
SS	0.1105	90.25	90.40
SP	0.0297	84.82	85.32
TA	0.0249	70.29	70.85
AG	0.0019	73.56	80.00
ES	0.0050	71.82	72.96
AN	0.0057	39.42	40.26
UK	0.0305	93.17	93.32
CA	0.0073	67.66	68.08
FS	0.0050	71.31	71.69

	Freq	NoFeats	Finite
ROOT	0.0649	86.71	88.03
SS	0.1105	90.25	90.91
VG	0.0302	94.65	96.42
OO	0.0632	84.53	85.31
+F	0.0099	52.07	55.45
MS	0.0096	63.35	66.63
TA	0.0249	70.29	71.20
AA	0.0537	68.70	69.04
++	0.0422	90.33	90.67
NA	0.0422	92.46	93.56

Table 9.11: 10 most improved dependency relations with added information on voice (left) and finiteness (right), ranked by their weighted difference of balanced F-scores.

(175) …*man kan* **bli** **vald** *i en annan valkrets*
…one can become elected in a other constituency
'One may be elected in another constituency'

If we examine only the verbal, passive predicates, we find that passive predicates account for only 9.8% of all verbal predicates in Talbanken.[180] The passive suffix is by far the most common mode of expression for passive voice in Swedish and these account for 78.4% of the passive predicates in the treebank.

With the addition of information on voice we would expect an improvement for the SS and OO relations in particular, as well as the passive agent relation (AG). Table 9.11 shows the ranked list of improved dependency relations for the Voice experiment. We do find an improved assignment for subjects and objects, as well as the passive agent. The improvement for the AG is in fact notable, with F-scores improving from 73.6% to 80.0%. However, since this dependency relation is infrequent in the treebank, improvement has less effect on overall results and this relation is ranked lower in the list in table 9.11 than the more frequent argument functions.

The improvement in analysis of the OO relation is clearly linked to verbal

[180]We count as verbal predicates all elements annotated as verb (30767 tokens) or verbal participle (715 tokens), i.e., participles which are dependents of a verb. Moreover, we count as verbal passive predicates all verbs annotated as passives (*s*-suffixed verbs: 2413 instances), and all verbal participles annotated as passives (666 instances out of which 446 have *vara* 'be' as a head verb and 114 have *bliva* 'become'). The group of verbal passives does not include passive participles in attributive function, e.g., *tecknade symbolbilder* 'drawn pictures', *skalade potatisar* 'peeled potatoes'.

argument structure; a passive transitive verb does not take an object whereas its active version does. We find a 14.6% decrease in total number of errors for the OO_SS error type and we find that the corrected errors for this type (26.2% of the baseline errors) consist exclusively of objects of active verbs. We find a parallel improvement for the SS_OO error type, with a 10.9% decrease in total numbers of errors. The added information improves on 18.8% of the baseline errors, and we find that the corrected errors consist almost exclusively of postverbal subjects (93.1%) of passive verbs.

Table 9.11 also shows improvement for the subject predicative relation (SP). This is due to the fact that the participle in periphrastic passive constructions is encoded as subject predicative.

The fact that the addition of information on voice does not have a great overall effect can be attributed to several factors. First of all and as mentioned above, only around 9% of all predicates are passive in Talbanken, so we are not adding a wealth of new information. In addition, passives are not involved in that many errors performed by the baseline parser. In fact, only 4% of the dependency relation assignment errors involve either a passive head or dependent. Hence it seems clear that passives do not pose severe problems for the baseline parser. Also, and touching on a more general problem with verbal features, complex predicates containing a finite auxiliary encode their external argument as a dependent on the auxiliary verb but other arguments as dependents of the lexical head verb. In other words, in the analysis of periphrastic constructions with a preverbal subject, the information on voice expressed on the non-finite verb may be interceded by several constituents when the subject relation is assigned and hence not available as a feature of the history. Even so, the above error analysis showed evidence for the effect of voice information for errors relating to argument disambiguation and acquisition of differential argument structures for active and passive verbs.

Tense

Judging from the previous section, information on tense is responsible for a majority of the improvement in dependency relation assignment observed with the addition of verbal features. An experiment (Tense) was therefore performed with information only on tense. The results in table 9.10 show a significant improvement from the baseline ($p<.0001$). We find that the property of tense is clearly responsible for the main improvement when we add information on verbal features (Verb).

Finiteness

In order to ascertain the influence of finiteness, an additional experiment was performed where the various tense features were mapped to their corresponding class of 'finite' or 'non-finite'.[181] We see the results in table 9.10 and find a significant improvement from the baseline ($p<.0001$). It is clear that the simple property of finiteness captures the relevant distinctions shown by the tense features. In fact, the mapping to a binary dimension of finiteness causes a further improvement of overall results ($p<.03$), compared to the use of tense features. This clearly supports the central role of finiteness in Scandinavian syntax, and V2-languages in general. Recall that the finite verb provides a fixed position in the positioning and ordering of clausal elements. As table 9.11 shows, the addition of finiteness information causes improved analysis for verbal relations, the core argument relations (SS, OO), as well as non-argument, adverbial relations (TA, AA, NA).

9.2.7 Feature combinations

The following experiments combine the different nominal argument features, the nominal argument features with the verbal features, and finally all available grammatical features in Talbanken05. A question for the following is therefore whether or not we will observe a combined effect that improves the best result obtained for individual features.

The combination of the argument features of animacy, definiteness, pronoun type and case (ADPC), as well as the addition of verbal features to this feature combination (ADPCV) causes a clear improvement compared to the baseline *and* each of the individual feature experiments ($p<.0001$), see table 9.2). Since the results are better than the individual runs, we may conclude that there is a cumulative effect of the combined information.

Table 9.12 shows a ranked list of the dependency relations with the greatest effect on the improved results for the ADPCV experiment. We find an improvement for the main argument relations of subjects and objects (SS, OO), the verbal relations (MS, VG), as well as for the other functions which improved the most with the individual argument features – determiners (DT), subject predicatives (SP) and formal subjects (FS).

Table 9.13 shows a ranked list of improved argument relations. We find that the combined features results in improved performance for practically all ar-

[181] Note that we are not equating tense and finiteness, since there are untensed forms which are still finite, e.g. the imperative (Holmberg and Platzack 1995). Rather we map the present and past tenses, as well as the imperative to the class 'finite' and the rest to the 'non-finite' class.

	Freq	NoFeats	ADPCV
SS	0.1105	90.25	91.87
OO	0.0632	84.53	86.38
DT	0.1081	94.14	95.22
PA	0.1043	94.69	95.43
VG	0.0302	94.65	96.53
+F	0.0099	52.07	55.97
SP	0.0297	84.82	86.10
ET	0.0523	76.46	77.03
MS	0.0096	63.35	65.98
FS	0.0050	71.31	74.09

	Freq	NoFeats	All
SS	0.1105	90.25	92.10
CC	0.0343	78.02	82.21
OO	0.0632	84.53	86.77
DT	0.1081	94.14	95.22
PA	0.1043	94.69	95.53
AA	0.0537	68.70	70.18
MS	0.0096	63.35	70.26
VG	0.0302	94.65	96.67
TA	0.0249	70.29	72.71
+F	0.0099	52.07	57.08

Table 9.12: 10 most improved dependency relations with combined features (ADPCV; left) and all features (right), ranked by their weighted difference of balanced F-scores.

	Freq	NoFeats	ADPCV
SS	0.1105	90.25	91.87
OO	0.0632	84.53	86.38
SP	0.0297	84.82	86.10
FS	0.0050	71.31	74.09
AG	0.0019	73.56	79.75
FO	0.0009	56.68	67.65
ES	0.0050	71.82	73.67
VO	0.0007	72.10	84.72
VS	0.0006	58.75	65.56
OP	0.0011	27.91	30.28
IO	0.0024	76.14	77.09

Table 9.13: Improved argument relations with combined features (ADPCV), ranked by their weighted difference of balanced F-scores.

gument relations.[182] If we examine the confusion matrices for subjects and objects in table 9.4, we find a reduction of total errors for the SS_OO and OO_SS error types with 34.3% and 30.4% respectively. With respect to the specific errors performed by the baseline parser, we observe a substantial reduction of 44.6% for SS_OO and 46.0% for OO_SS. In the error analysis for the baseline parser in section 8.3, we concluded that word order and morphology does

[182]The only exception is the relation of logical object (EO) for which there is no change in accuracy compared to the baseline results. This is a very infrequent relation with only 22 instances in the treebank.

		Before		After		Total	
Gold	System	#	%	#	%	#	%
SS	OO	21	10.6	178	89.4	199	100.0
OO	SS	15	10.6	127	89.4	142	100.0

Table 9.14: Order relative to verb for corrected SS_OO and OO_SS errors in the ADPCV experiment.

		Noun		Pro_{amb}		Pro_{unamb}		Other		Total	
Gold	System	#	%	#	%	#	%	#	%	#	%
SS	OO	144	72.4	23	11.6	18	9.0	14	7.0	199	100.0
OO	SS	111	78.2	21	14.8	6	4.2	4	2.8	142	100.0

Table 9.15: Part of speech for corrected SS_OO and OO_SS errors in the ADPCV experiment.

not provide sufficient information for argument disambiguation in all cases. In particular, we noted that arguments which depart from the most common ordering and/or are not morphologically marked are overrepresented among the errors. We concluded that additional linguistic information is needed in order to resolve these ambiguities. In tables 9.14 and 9.15 we examine word order and part-of-speech for the corrected SS_OO and OO_SS errors in the ADPCV experiment. We see that the added information contributes to the reduction of precisely the types of errors which were identified in the error analysis. In particular, improvement is centered in postverbal positions, largely occupied by nouns and case ambiguous pronouns.

As we saw in section 5.1.1, Talbanken05 also contains a set of semantic features for other parts-of-speech, like adverbs, conjunctions and subjunctions. When we add the remaining set of linguistic features (All), the results improve further and differ significantly from the ADPCV experiment ($p<.0001$).

As table 9.12 shows, we observe a notable improvement for the conjunct relation (CC) as well as further improvement for argument relations (SS, OO), determiners, verbal relations and adverbials. It is largely the improved analysis for different types of coordinated relations (CC, MS, +F) and adverbials (AA, TA) which is the main difference from the ADPCV experiment. Improved analysis of coordinations and adverbials also influences the analysis of arguments. However, the confusion matrices for the various argument relations show that the added features have a different effect than the linguistic features studied above. For the error types expressing confusion of argument relations, such as SS_OO, OO_SS, SS_SP, FS_SS, we hardly observe any improvement at all. Rather, we find improvements in the error types involving arguments and co-

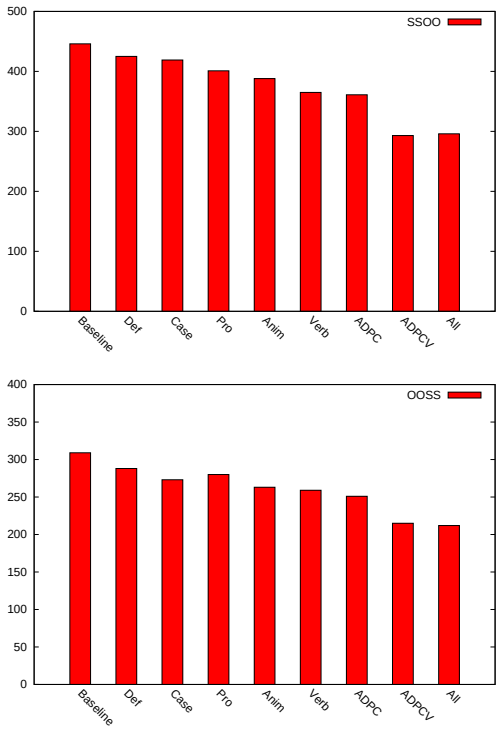

Figure 14: Total number of SS_OO errors (top) and OO_SS errors (bottom) in the various experiments.

ordinated elements, such as SS_CC, SP_CC, as well as argument and adverbial relations, such as OO_AA, ES_KA. Figure 14 shows the total number of SS_OO and OO_SS errors in the various experiments and clearly illustrate the observed reduction for this error type with the chosen set of linguistic features, as well as the lack of effect for the remaining features added in the All experiment. This indicates that the initial error analysis and the hypotheses formulated there provided a useful understanding of the problem. The linguistic features presented in 9.1, which were chosen to approximate dimensions of argument differentiation and defining properties of Scandinavian morphosyntax, contribute to the task of argument disambiguation.

9.2.8 Selectional restrictions

We have seen how additional information on semantic properties of arguments contributes to improved dependency relation assignment. However, argument

Class	Restriction
AnimSS	Selects animate subject
AnimOO	Selects animate object
InanSS	Selects inanimate subject
InanOO	Selects inanimate object

Table 9.16: Selectional restriction classes.

selection is clearly not only determined by isolated properties of the argument. On the contrary many would claim that the semantic class of the verb and the selectional restrictions posed on the argument by the verb are key properties to understanding the relation between the arguments and the head predicate. As we noted in section 3.5, animacy has figured among the categories used to define selectional restrictions from the very beginning. Later, computational approaches have made reference to more fine-grained semantic classes, usually taken from the English WordNet.

In the following we will present some experiments investigating the addition of information on selectional restrictions for verbs which focus on the category of animacy. We assign to verbs a selectional restriction class, see table 9.16, based on their occurrences in the treebank. The goal is thus to *enrich* the existing treebank annotation and furthermore to employ the extended features during parsing. We will in the following investigate the nature of selectional restrictions as absolute or gradient, as well as their effect on parse results.

9.2.8.1 Data extraction

We extract predicate-argument pairs from the treebank and generalize over these to determine the selectional restrictions for the predicate. In order to enable generalizations about verbs, the treebank is lemmatized prior to data extraction, using a lemmatizer for Swedish (Kokkinakis 2001).

The selectional restrictions of verbs constrain the semantic properties of arguments, operating at the syntax-semantics interface, and we must take into account mismatches in the mapping between syntactic structure and semantic predicate-argument relations. In extracting the relevant data, we have to determine for each verb-argument pair i) its semantic predicate, and ii) the grammatical relation between the argument and the predicate.

With regard to the first point, the treebank annotation of so-called “verb groups” must be taken into account. These consist of a finite auxiliary or modal verb along with one or more non-finite verbs. The subject argument of a verb

group is annotated as a dependent of the finite verb, whereas complements are annotated as dependents of the non-finite lexical main verb. For instance, as the dependency representation in figure 15 for example (176) illustrates, the subject, *man* 'one', is the structural argument of the finite verb, the modal auxiliary *kan* 'can', whereas the direct object, *struntsaker* 'trivialities', is the structural argument of the lexical verb *diskutera* 'discuss':

(176) *Där* **kan** *man* **diskutera** *struntsaker*
there can one discuss trivialities
'There, one can discuss trivialities'

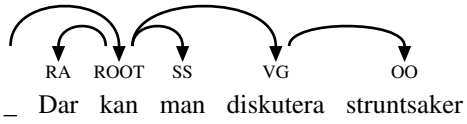

Figure 15: Dependency representation of example (176)

The dependency representation of (176) does not indicate that *man* 'one' semantically also acts as an argument of *diskutera* 'discuss'. In order to locate the lexical main verb in the dependency graph, we pursue each finite verb's, possibly null, chain of non-finite dependents. For the example sentence in (176) then, we extract the following pairs:

```
diskutera-SS:anim
diskutera-OO:inan
```

In addition to determining the semantic predicate, we must also determine the grammatical relation which holds between the predicate and its arguments. In most cases, this corresponds directly to the dependency label of the argument, but in passive constructions the structural assignment of grammatical functions does not directly reflect the semantic relations of the predicate. For instance, in example (177) below, the verb lemma *respektera* 'to respect' should not be recorded as selecting an inanimate subject in this particular case, but rather an inanimate object.

(177) *Parternas* *integritet* *måste* **respekteras** ...
part-DEF.GEN integrity must respect-PASS ...
'The integrity of the parties must be respected'

So, in the case of passive verbs, the argument relations are inverted and we record the subject as an object of the verb:

```
respektera-OO:inan
```

9.2.8.2 Selectional association

We wish to assign to each verb in the treebank a selectional restriction class expressing the semantic restrictions it places on its argument. Following Resnik (1996), we will base our notion of selectional restriction on the *selectional association* between a predicate and a semantic class, which allows us to quantify the extent to which a predicate selects for an animate, as opposed to inanimate, subject or object. The selectional association of a verb with respect to a class will determine its assigned selectional restriction class, see table 9.16. As we mentioned in section 3.5, theoretical and computational work on selectional restrictions have differed in their view on selectional restrictions as categorical or gradient. We will base our selectional restrictions on a probabilistic measure, approximated by corpus data, where verbs with categorical constraints on their arguments are simply the verbs with a selectional association of 1. We may thus experiment with different degrees of gradience in the selectional restriction classes by manipulating a threshold of selectional association.

Resnik (1996) presents a method for acquisition of selectional restrictions which is based on an information-theoretic approach to verbal argument selection. The approach quantifies the overall extent to which a predicate constrains the semantic class of its arguments as its *selectional preference strength*.[183] The contribution of the semantic class of the argument to the selectional preference strength of the predicate is expressed as the *selectional association* between a predicate and the particular class. Resnik (1996) looks at selectional restrictions for verbs and their objects and employs the WordNet resource (Fellbaum 1998) for obtaining semantic classes. Since nouns may belong to several classes in WordNet and no sense-tagged corpus is available, the estimation of frequencies distributes the ambiguity evenly over the nouns of a class.

Rather than defining selectional association through the additional notion of selectional preference strength, we define the association of a verb *v* with an argument class *c* directly as the conditional probability of the class given the

[183] Selectional preference strength expresses the amount of information a predicate carries regarding its argument by looking at the difference in the prior distribution of a semantic class ($P(c)$) in an argument position and the resulting distribution when taking the specific predicate into account, expressed by the conditional probability $P(c|p_i)$. More precisely, it is defined as the relative entropy (Kullback-Leibler divergence) between the prior distribution of semantic classes of arguments and the distribution for a particular predicate (verb) (Resnik 1996):

$$S(p_i) = D(P(c|p_i)|P(c)) = \sum_c P(c|p_i) log \frac{P(c|p_i)}{P(c)}$$

predicate (Manning and Schütze 1999: 293):

$$A(v,c) = P(c|v)$$

Resnik (1996) proposes an estimation of the joint probability $P(v,c)$, which takes into account class ambiguity for individual nouns. However, since we have annotated data with respect to the person/non-person distinction in Talbanken05, the estimation reduces to the following:

$$P(v,c) = \frac{freq(v,c)}{N}$$

where N is the total number of occurrences with respect to an argument type – subject or object, and $freq(v,c)$ is the count of occurrences of a verb with an argument of a certain class. $P(v)$ is estimated as the maximum likelihood estimate $freq(v)/\sum_i freq(v_i)$ expressing the relative frequency of the verb with respect to all verbs.

For each verb lemma in Talbanken, selectional association with the semantic classes of animate and inanimate is calculated based on the extracted data from the treebank, as detailed above. Verbs are then assigned a selectional restriction class based on their association with the two classes. In doing so, we set a threshold expressing the level of gradience embodied by the selectional restriction. A threshold of 1.0 will assign a class only to verbs which impose categorical constraints on their arguments, whereas a lower threshold, e.g. 0.9, will allow for some variation. In the following we will experiment with both categorical and gradient selectional restriction classes.

Selectional association is calculated separately for subjects and objects. This means that the sets of verbs in the classes for subjects and objects are not disjoint. A transitive verb lemma may be assigned both a subject class (`AnimSS/InanSS`) and an object class (`AnimOO/InanOO`). There are thus complex classes for transitive verbs which cover subsets of the simple classes in table 9.16. However, since the parser allows for bundles of individual features there is a clear advantage in tagging the data for simple classes.[184] This choice of annotation will facilitate generalization over the selectional restrictions of verbs of differing valencies. This will also assure that we do not mix in the notion of subcategorization with that of selectional restriction.

[184] As we recall from section 8.1.3, the parse guide classification is performed employing support vector machines. Depending on the kernel function chosen, feature combinations are constructed of size *n* internally. Here we employ a quadratic kernel function, hence n=2 and *pairs* of all features are constructed for classification. So for instance, pairs such as `AnimSS&HH`, will express a selectional restriction class of animate subjects occurring with a nominal with person reference (`HH`).

Class	Types	Tokens	Examples
AnimSS	388	1549	*kasta*, 'throw', *katalogisera* 'catalogue', *kissa* 'pee', *klaga* 'complain', *klamra* 'cling', *klandra* 'blame', *klappa* 'pat'
InanSS	333	1824	*spegla* 'mirror', *spetsa* 'sharpen', *spoliera* 'spoil', *spricka* 'shatter', *sprida* 'spread', *spruta* 'squirt', *sticka* 'sting'
AnimOO	143	618	*adoptera* 'adopt', *akta* 'beware', *aktivera* 'activate', *avskräcka* 'scare', *be* 'ask', *befatta* 'involve', *befria* 'free'
InanOO	739	5178	*rengöra* 'clean', *reparera* 'repair', *representera* 'represent', *restaurera* 'restore', *revidera* 'revise', *rikta* 'direct', *riva* 'tear-down'

Table 9.17: Verb lemmas by selectional restriction class with $A(v,c) = 1.0$.

9.2.8.3 Gradience of selectional restrictions

Table 9.17 presents an overview of the verb lemmas and tokens assigned to each of the four classes under a categorical definition of selectional restrictions, i.e., where $A(v,c) = 1.0$. As we see from the examples in table 9.17, the selectional restriction classes for subjects cut across valency classes and include both intransitive and transitive verbs. All together, 1181 unique verb lemmas verb were assigned at least one selectional restriction class. This gives us a coverage of 75.0% of the total number of unique verb lemmas in Talbanken (1573 lemmas in total). As expected, among the verb lemmas not covered by the classification we find verbs which may function as auxiliary verbs (copula, modals etc.) and hence may take any type of argument depending on the non-finite lexical verb. It is therefore not surprising that the classification coverage in terms of verb tokens is low; the auxiliary verbs are, after all, the overall most frequent verbs. As a consequence, only 24.5% (7523 tokens out of total 30767) of the verb tokens in Talbanken receive a selectional restriction class.

Another factor which influences the coverage is clearly the fact that the restrictions are categorical and do not allow for any variation. Many verbs are simply not that restrictive with respect to their arguments. For instance, the verb *visa* 'show' may occur with both animate and inanimate subjects, as in (178) and (179) below, as well as animate and inanimate object, as in *visa*

någon ‘show someone’ and (178). As a consequence, this verb is not assigned a selectional restriction class with a threshold of 1.0.

(178) *Han skall* **visa** *intyg*
he shall show proof
‘He will show proof’

(179) *Konsumentprisindex skall* **visa** *prisförändringar*
consumer-price-index shall show price-changes
‘The consumer price index should indicate price changes’

In addition to variation with respect to argument selection, there are also elements of the annotation in Talbanken05 which contribute to the lack of coverage. First of all, the animacy annotation with respect to collective nouns and organizations contribute to the low coverage, since these are annotated as inanimate. For instance, in example (180) below, the subject of the verb *skriva* ‘write’ is the noun *länderna* ‘countries’ which is annotated as inanimate but clearly employed metonymically to refer to the animate representatives from the countries.

(180) *. . . som de sex länderna* **skrev** *under den 25 mars*
. . . which the six countries-DEF wrote under the 25 march
1957 i Rom
1957 in Rome
‘. . . which the six countries signed on the 25th of March 1957 in Rome’

In fact, all other active instances of this verb occur with a human subject.[185] Due to the example in (180) then, the verb lemma is not assigned the `AnimSS` class. Since it is a transitive verb it should in principle be assigned an object class as well, preferably the `InanOO` class. After all, one usually writes something, not someone. However, it is not assigned an object class since it occurs with a direct object (*‘mamma’* ‘mom’) annotated as person referring:

(181) *När Bowlby* **skriver** *‘mamma’ . . .*
when Bowlby write ‘mom’ . . .
‘When Bowlby writes ‘mom’. . . ’

With a largely denotational, rather than referential annotation practice, see section 7.1.2, this is a direct consequence.

[185] The verb in example (180) occurs with the particle *under* ‘under’ and one might argue that it should be represented as a separate verb lemma alltogether. A special treatment of particle verbs was not, however, pursued further in the present context.

Threshold	Types	Tokens
1.00	1181	7523
0.95	1205	9482
0.90	1245	14866
0.85	1278	17195
0.80	1312	18695
0.75	1343	26948
0.70	1354	27951
0.65	1388	29674
0.60	1393	30271
0.55	1397	30314

Table 9.18: Absolute number of classified verb types and corresponding treebank tokens under various selectional association thresholds t, where $A(v,c) = t$ and $c \in \{\texttt{AnimSS}, \texttt{InanSS}, \texttt{AnimOO}, \texttt{InanOO}\}$.

Table 9.18 illustrates the number of types and tokens that receive at least one selectional restriction class under various thresholds for selectional association. Clearly, lowering the threshold allows for more variation in argument selection, hence provides wider coverage. With a threshold of 0.95 we find for instance that the verb *skriva* 'write', which did not receive a class earlier, now receives the class `AnimSS`. With a lowered threshold to 0.90 it also receives the class `InanOO`. Above, we also examined the verb *visa* 'show' which exhibited quite a bit of variation in terms of selectional restrictions. With a threshold of 0.95 it receives the class `InanOO` encoding a preference for inanimate objects. The threshold has to be lowered to 0.70, however, for the verb to receive a subject class (`InanSS`), indicating the lower degree of selectional constraint which this predicate enforces on its subject argument, as exemplified by (179) above.

9.2.8.4 Experiments with selectional restrictions

In a set of parse experiments, the information on selectional restrictions for verbs (SR), extracted as detailed above, is included as an additional feature. All experiments except one ($SR_{1.0}$) also include information on animacy as this is the semantic class relevant for the selectional restrictions. In evaluating the results we will therefore compare the results both to the general baseline (NoFeats), as well as the experiment employing only animacy information (Anim). In the experiments we furthermore vary the selectional association threshold for the verb classes expressing selectional restrictions.

	Unlabeled	Labeled
NoFeats	89.87	84.92
Anim	89.93	85.10
$SR_{1.0}$	89.80	84.91
$SR_{1.0}$&Anim	89.89	85.06
$SR_{0.95}$&Anim	89.90	85.05
$SR_{0.90}$&Anim	89.92	85.04
$SR_{0.85}$&Anim	89.93	85.04

Table 9.19: Overall results for experiments with selectional restrictions, expressed as unlabeled and labeled attachment scores.

The third row of table 9.19 shows the overall parse results with information on categorical selectional restriction class for verbs only ($SR_{1.0}$), i.e. *without* the corresponding semantic information for arguments. This experiment was performed in order to observe the effect of the selectional restriction classes in isolation. As we can see the addition has a slightly detrimental, but not statistically significant, effect on overall labeled results and displays a significant dip in unlabeled accuracy ($p<0.01$). The fourth row of table 9.19 shows the results in the parse experiments employing categorical selectional restrictions along with information on animacy ($SR_{1.0}$&Anim), and we do not find any significant improvements compared to the Anim experiment. It is clear that the problems caused by the addition of selectional restriction information in isolation is not countered by any effects obtained in combination with the semantic information for animacy.

Since the addition of information only on selectional restrictions ($Verb_{SR}$) clearly has some unexpected side-effects which carry over to the other experiments, we perform an error analysis of these results. We focus largely on the unlabeled results, which is where we observed a clear deterioration of results. Recall from section 8.2.1 that the set of unlabeled errors is defined solely by error in attachment. We sort these errors into error types based on the part-of-speech of the dependent, as well as the correct and erroneously assigned head. We find that the additional information causes a rise in the number of unlabeled attachment errors for all major types of parts-of-speech except for verbal dependents. Table 9.20 shows the total number of attachment errors for the error types which increase the most with the added information.

For nominal elements like nouns (N) and pronouns (PO), we note an increase in attachment to the superficial root of the dependency graph (ROOT). As we noted in section 8.3.2, erroneous attachment to the artificial root means that an element has not been attached at all. An increase in root attachment

POS_{Dep}	POS_{Gold}	POS_{Sys}	NoFeats	$Verb_{SR}$	
PR	V	N	828	860	-3.9
PO	V	ROOT	189	212	-12.2
AB	V	V	324	336	-3.7
N	PR	ROOT	75	87	-16.0
N	N	PR	218	229	-5.0
N	V	ROOT	288	297	-3.1
N	N	ROOT	199	207	-4.0
AB	V	ROOT	104	112	-7.7
++	V	N	119	127	-6.7
AJ	N	V	99	106	-7.1

Table 9.20: Total number of errors in the $SR_{1.0}$ experiment compared to the NoFeats baseline, along with relative deterioration compared to the baseline (%); sorted by error type ($POS_{dep}\_POS_{head}\_POS_{err}$) and ranked by total difference of deterioration.

with the added information on verbal classes thus indicates more restrictive attachments, resulting in a more fragmented analysis. The newly added attachment errors for prepositions (PR) and adverbs (AB) are characterized by being adverbial in nature; instead of attachment to the correct verb, an alternative, erroneous site is chosen (noun, another verb etc.). What we observe then, both for the argumental and adverbial elements, is a general resistance to attach to verbal elements, clearly caused by the added information.

The assumption that selectional restrictions are categorical has clearly been shown to be too strong. As we saw above, it results in missed generalizations as well as a poor coverage for the enriched annotation. Parse experiments were therefore performed testing three different thresholds ($t \in \{0.95, 0.90, 0.85\}$) for selectional association between a verb and an animacy class, expressing an increased gradience within the selectional restriction classes. In the last three rows of table 9.19, we show the results from these three experiments. With a threshold of 0.95 we now have a coverage of 55.9% of the total verb tokens in Talbanken. But even with an increased coverage, we find that the results do not improve significantly compared to the Anim-experiment.

9.2.8.5 Summary

The above sections have detailed a strategy for extracting selectional restrictions for verbs from Talbanken, based around the semantic dimension of animacy for which we have annotated data. The ensuing experiments have tested

the effect of selectional restrictions for verbs during parsing.

It has proven difficult to obtain any strong effects from the selectional restrictions. There are several interdependent factors which contribute to the situation. First of all, it is well known that argument realization is characterized by massive *variation*. We have looked at variation in semantic argument characteristics – verbs select arguments of different classes and do so to differing degrees. We employed a probabilistic measure of association strength between a verb and a semantic argument class where variation is interpreted as degrees of gradience and experimented with both categorical and gradient selectional restrictions. One complicating factor is that functional verbs such as auxiliaries, which are highly frequent, do not constrain their arguments and followingly are not assigned a selectional restriction class.

It seems fair to assume that we are dealing with a *sparse data* situation at more than one level. We have a coverage problem in the treebank data due to variation in the realization the nominal semantic classes of verbal arguments. However, even with complete coverage of verbal lemmas in the treebank data, it might still be problematic to make the assumption that the treebank contains sufficient data for assignment of selectional restrictions. This might simply be wrong and it is most likely the case that substantially more data is needed in order to make the right kinds of predictions. We will return to this issue in section 9.4.2 below where we will employ a considerably larger, automatically annotated data set. Moreover, the assumption that selectional restrictions may be reduced to a binary notion of animacy may also be debated. It might be that this is too coarse a distinction to enable interesting generalizations regarding verbal semantics.

9.3 Features of the parser

The above experiments have focused on variations in the linguistic input to parsing, whereas properties of the parser have been kept constant. This section will investigate variations over different features of the parser on the input data employed above. In particular, we will examine parser generalizability and feature locality. These will in different ways elucidate further the nature of the effects of our set of linguistic features.

9.3.1 Parser comparison

In the experiments in section 9.2, we have exclusively employed the MaltParser system for data-driven dependency parsing. The aim of this section is to com-

pare the effect of the linguistic features investigated above when employing a different parser. In particular, we will investigate the stability of the effects across parsers and to what extent they are dependent on certain properties of the parser.

As we mentioned in section 8.1.3, we may distinguish roughly between two current approaches to data-driven dependency parsing: the graph-based approaches and the transition-based approaches (McDonald and Nivre 2007). The main characteristics of the two approaches may be summarized as follows:

Graph-based locate the highest-scoring dependency graph given an induced scoring function

- global training
- exhaustive search/inference

Transition-based locate the optimal transition sequence given an induced parse guide

- local training
- greedy search/inference

The graph-based approaches typically employ global training and induce a scoring function for dependency graphs. Parsing is thus construed as search through all possible dependency graphs for a sentence to locate the highest-scoring graph. Transition-based approaches in contrast employ local training in the induction of a parse guide which in combination with a greedy search algorithm optimizes the parse transitions. As a practical consequence of the differences in induction and search, the feature models employed also differ. Graph-based approaches typically employ a rather limited feature model in the representation of dependency graphs, whereas transition-based approaches operate with a richer feature history in order to compensate for the decomposition into transitions. These two approaches thus differ with respect to a number of properties, however, achieve comparable overall results in dependency parsing (Buchholz and Marsi 2006). It might therefore be interesting to compare the effect of the linguistic features studied above.

9.3.1.1 MSTParser

MSTParser (McDonald, Crammer and Pereira 2005; McDonald et al. 2005) is an instance of a graph-based data-driven dependency parser.[186] As we mentioned above, parsing with MSTParser consists in locating the highest scoring

[186] MSTparser is freely available from http://mstparser.sourceforge.net

	MaltParser		MSTParser	
	Unlabeled	Labeled	Unlabeled	Labeled
NoFeats	89.87	84.92	89.67	82.91
Anim	89.93	85.10	89.82	83.04
ADPC	90.17	85.45	90.00	83.22
ADPCV	90.42	85.73	90.21	83.41

Table 9.21: Overall results for experiments comparing MaltParser and MSTParser, expressed as unlabeled and labeled attachment scores.

dependency graph according to a scoring function. The scoring function is induced through global training with features of the head, dependent, as well as elements occurring in the vicinity of these (before/after/between). Global training is training based on the global dependency graph and features are therefore not limited to previous parse decisions. However, the vast space of possibilities, in theory all possible subgraphs, in practice limits the expressiveness of the feature model. The feature model in MSTParser is hard-coded, hence may not be modified as easily. Our additional linguistic features (FEATS) are employed more restrictively in the scoring of edges than the part-of-speech and lexical features and represent only the head, dependent and conjunctions of these. Furthermore, dependency labels are not used as features during parsing.

9.3.1.2 Comparative experiments

In a set of experiments we run MSTparser on the NoFeats, Anim, ADPC and ADPCV data sets from Talbanken05 employed in our earlier experiments. Apart from the choice of parser, the experimental setting and evaluation measures are identical to the earlier experiments and MSTparser is run with default settings.

Table 9.21 shows the overall results in the MSTParser experiments and contrasts them with the corresponding MaltParser results. We compare the results for both parsers individually with a baseline employing no additional linguistic features (NoFeats) and experiments testing the addition of information regarding animacy (Anim), animacy, definiteness, pronoun type, and case (ADPC), as well as verbal features (ADPCV). These are the equivalents of the experiments detailed in sections 9.2.2 and 9.2.7 above.

First of all, we may note that the general results, with or without added features, are lower than the corresponding results obtained with MaltParser. The most notable discrepancy is in the labeled results. Table 9.22 shows an

Gold	System	#
SS	OO	718
++	++	674
AA	OA	557
ET	OA	489
AA	AA	481
ET	ET	469
OA	ET	465
OO	SS	461
AA	RA	451
UK	UK	435

Table 9.22: 10 overall most frequent error types for MSTParser on the NoFeats data set, where SS=subject, OO=object, ++=conjunction, AA=other adverbial, OA=object adverbial, ET=nominal post-modifier, RA=spatial adverbial, UK=subjunction.

overview of the ten most common labeled error types for the MSTParser baseline. It exhibits some differences from the most common error types for the MaltParser baseline, as presented in table 8.1 in section 8.2.3. In general, we find that attachment errors, such as ++_++, AA_AA, ET_ET, UK_UK, are more common among the MSTParser errors. These are errors where the dependent is labeled correctly, but where the attachment is incorrect. In section 8.2.3 we noted that confusion of subjects and objects, as well as various adverbial relations constituted the most common errors in the MaltParser baseline results. We find the same error types in the results for MSTParser as well, and the SS_OO error type is in fact the most common error type made by the baseline parser. The SS_OO and OO_SS errors show very similar properties to the errors analyzed in section 8.3.2. There, we found that the distribution of errors differed from the overall distribution of subjects and objects with respect to word order and morphological marking. In other words, the subjects and objects which deviated from the norm were overrepresented among the errors. We find the same, clear pattern in the baseline results for MSTParser. 83.1% of the SS_OO errors are postverbal and 94.3% are realized by a noun or case ambiguous pronoun. For the OO_SS errors, we find that 35.4% are preverbal and 96.7% are nouns or case ambiguous pronouns.

As table 9.21 shows, the added features have a positive effect on overall results also when employing MSTParser and we find that all differences are significant compared to the NoFeats baseline. The addition of information on animacy causes a clear improvement in unlabeled results ($p<0.0004$) and a

	NoFeats$_{MST}$	Anim$_{MST}$
DT	94.14	94.49
SS	87.32	87.60
ET	71.38	71.81
++	90.50	90.75
AN	30.88	32.43
ROOT	94.60	94.72
CC	76.95	77.15
TA	62.39	62.62
PA	93.77	93.82
VG	92.86	93.01

Table 9.23: 10 most improved dependency relations for the MSTParser with added information on animacy, ranked by their weighted difference of balanced F-scores.

smaller improvement ($p<0.003$) in labeled results. This is the converse situation from the results for MaltParser, where the observed improvement was largely in terms of labeled results. If we examine the sets of attachment errors, we find the most notable improvement for cases of pronouns, nouns and verbs. These are mainly errors in attachment to verbs (argument attachment), as well as attachment of nominal elements to other nominal elements (phrasal attachment).

Table 9.23 shows a ranked list of the dependency relations which show the largest improvement in the Anim$_{MST}$-experiment. Ranked at the top of the list are the determiner (DT) and subject relations (SS). A closer look at the results shows that the improvement in labeled results is largely due to the improved attachment. As a consequence, we observe a notable reduction in total number of errors for error types involving ambiguities between a phrasal and clausal reading, such as SS_DT, DT_SS. We may note that the performance for the DT relation is in fact identical for MaltParser and MSTParser with a baseline F-score of 94.14 and the effect of the animacy information causes a nearly identical improvement for this relation to 94.48 and 94.49, respectively. A clear difference between the two systems, however, is found in improvement in terms of labeling only. In contrast to the MaltParser results, performance does not improve for the argument relations of OO and SP. The total number of SS_OO and OO_SS errors, representing largely a labeling error, also does not decrease notably.[187]

[187] The results show a small improvement of 2.6% for the OO_SS error type, however, these are all due to corrected attachment errors of the SS_DT type in the immediately preceding context,

	Freq	NoFeats$_{MST}$	ADPCV$_{MST}$
DT	0.1081	94.14	95.22
SS	0.1105	87.32	88.10
PA	0.1043	93.77	94.21
ET	0.0523	71.38	72.13
OO	0.0632	79.92	80.39
++	0.0422	90.50	91.13
ROOT	0.0649	94.60	94.99
CC	0.0343	76.95	77.61
AA	0.0537	63.80	64.18
+F	0.0099	45.38	46.70

Table 9.24: 10 most improved dependency relations for MSTParser with added information on ADPCV, ranked by their weighted difference of balanced F-scores.

In the ADPC and ADPCV experiment we find a significant improvement of overall results ($p<0.0001$) both in terms of labeled, as well as unlabeled results. Table 9.24 shows the performance per dependency relation for the ADPCV experiment and we find an improved analysis for all argument relations. We may conclude that the added information has a general positive effect also with a parser which is radically different from that of MaltParser. We found that the largest effect is in terms of unlabeled results, hence an increased attachment accuracy both for clausal and phrasal constituents. Even so, we generally observe a less notable improvement in terms of labeled results compared to the results in the experiments with MaltParser. McDonald and Nivre (2007) show that MaltParser cross-linguistically has a better performance for core argument relations like subjects and objects than MSTParser and suggest that a possible reason for this is the fact that MSTParser does not condition on the previously assigned dependency relations during parsing. The results from our experiments corroborate this and indicate that the improvement in terms of argument analysis is partially dependent on properties of the preceding analysis during parsing. We have also noted that the additional linguistic features are employed highly locally in MSTParser. In the following section, we will investigate the influence of feature locality in argument disambiguation further.

see section 8.3.2 and examples (163)–(164).

	Unlabeled	Labeled
NoFeats	89.87	84.92
Anim	89.93	85.10
ADPC	90.17	85.45
Anim_{Local}	89.93	85.08
ADPC_{Local}	90.16	85.39

Table 9.25: Overall results for experiments with feature locality in MaltParser.

9.3.2 Feature locality

In the experiments in section 9.2 above, we have employed the same feature model whilst varying the input of the linguistic features. The experiments with MSTParser discussed above suggested that conditioning on properties of the preceding linguistic context is important in argument disambiguation. Varying the feature model of the parser provides a manner of testing the influence of our features further.

The feature model employed by the parse guide in MaltParser provides a rich history for each transition. The feature model in figure 13 shows that the additional linguistic features, represented by the attribute FEATS, are employed for highly local tokens in a candidate head-dependent relation, as well as tokens which are further removed in the dependency graph, such as siblings and grandparents.

We perform a set of experiments where additional linguistic features are limited to the token on top of the stack and the next input token, i.e., *top* and *next*. The FEATS-features are thus limited to a highly local context, whereas features for the remaining attributes, FORM, POS, DEP are kept constant. Table 9.25 shows the overall results for two experiments employing this local feature model with various argument features: Anim_{Local}, where animacy information is included, and ADPC_{Local}, which employs information on animacy, definiteness, pronominal type and case. We chose to focus on argument features, and not include verbal features, in order to enable isolation of the effects on arguments.

The results indicate that the observed effect of the argument features is largely local. Both experiments (Anim_{Local}, ADPC_{Local}) show slightly, but not significantly, lower overall results compared to the counterparts employing a full feature model (Anim, ADPC). This means that the added information regarding candidate head and dependent is responsible for a majority of the improvement observed with the added features.

	Freq	NoFeats	Anim$_{Local}$
DT	0.1081	94.14	94.55
SS	0.1105	90.25	90.65
OO	0.0632	84.53	84.83
AA	0.0537	68.70	69.05
PA	0.1043	94.69	94.78
TA	0.0249	70.29	70.64
OA	0.0305	70.63	70.90
AT	0.0441	95.76	95.92
UK	0.0305	93.17	93.39
SP	0.0297	84.82	85.02

	Freq	NoFeats	ADPC$_{Local}$
DT	0.1081	94.14	95.14
SS	0.1105	90.25	91.12
PA	0.1043	94.69	95.31
OO	0.0632	84.53	85.47
ET	0.0523	76.46	77.09
AA	0.0537	68.70	69.21
SP	0.0297	84.82	85.54
FS	0.0050	71.31	74.08
OA	0.0305	70.63	71.00
IO	0.0024	76.14	79.68

Table 9.26: 10 most improved dependency relations with the local feature model and animacy features (left) and animacy, definiteness, pronoun type and case features (right), ranked by their weighted difference of balanced F-scores.

In table 9.26 we see a ranked list of the most improved dependency relations, compared to the NoFeats baseline, in the experiments with a local feature model. Compared to the full feature model counterparts we observe somewhat lower results for the argument relations. In parallel with the observations made in the MSTParser experiments, we find that the DT relation is the relation for which we find the largest improvement, clearly indicating a local effect of the features. The determiner-head relation constitutes a phrasal context where two nominal elements must be disambiguated. On several occasions, we have noted the animacy effect in genitive constructions. Determiners are typically not animate, unless marked with genitive case, and these are inherent properties of the nominal in question. Differentiating features of the two nominals clearly benefit disambiguation, and we observe improved attachment for determiners, as well as labeled improvement for the DT_SS and SS_DT error types. Unlike the results obtained with MSTParser, however, we also observe improved performance for a range of argument relations in the Anim$_{Local}$ experiment, such as objects (OO) and subject predicatives (SP). The fact that MaltParser conditions on preceding dependency relations is a factor which still distinguishes the two parsers. For the OO and SP relations, which are predominantly found in postverbal position, knowledge regarding assignment of preverbal dependents is clearly important for correct analysis.

The overall results presented above cover some interesting differences between the local and full feature model parsers which a further error analysis reveals. Table 9.27 shows relevant excerpts from the confusion matrices for the SS, OO, SP, FS and IO argument relations. It provides the total number of

		NoFeats		Anim_{Local}		ADPC_{Local}		Anim		ADPC	
Gold	System	#	% tot.	#	%	#	%	#	%	#	%
SS	OO	446	25.9	419	6.1	393	11.9	388	13.0	361	19.1
OO	SS	309	23.8	299	3.2	274	11.3	263	14.9	251	18.8
SS	SP	206	12.0	202	1.9	202	1.9	203	1.5	188	8.7
SP	SS	240	31.3	231	3.8	225	6.2	231	3.8	213	11.2
FS	SS	281	93.0	279	0.7	251	10.7	272	3.2	241	14.2
IO	OO	97	67.4	95	2.1	91	6.2	95	2.1	89	8.2

Table 9.27: Total numbers of errors for error types in experiments with the local feature model (Anim_{Local}, ADPC_{Local}) compared to the full feature model baseline (NoFeats) and respective counterparts (Anim, ADPC).

errors for error types expressing confusion of argument relations in the local experiments, compared with the full feature model baseline and the respective counterparts. The first two rows show errors of the types SS_OO and OO_SS and we find an improvement with local features for both error types in both experiments. A comparison with the results using a full feature model furthermore shows that errors of these types further improve with the use of less local features. We observe the same pattern for other error types involving the confusion of argument relations. For instance, we find a reduction in total numbers of errors for the SS_SP and SP_SS error types in the ADPC_{Local}-experiment, stemming from the addition of the definiteness feature. We also observe a further improvement for both of these with the full feature model. Error types expressing confusion of other argument relations, such as FS_SS and IO_OO, presented in the last rows of table 9.27, further corroborate the general effect.

9.3.3 Features of argument differentiation

It is clear that the addition of argument features have an effect on the analysis of arguments, both in a highly local setting with no structural context, as in the MSTParser experiments, in a local setting with a structural context, as in the local MaltParser experiments, and clearly also in the experiments with a full feature model in combination a structural context.

The results of the experiments performed in the current section highlight the *relative* aspect of argument disambiguation and argument differentiation in general. Arguments are differentiated not only by inherent properties, but also by their properties relative to other arguments. As we discussed in chapter 3, the linguistic dimensions by which arguments tend to differ incur soft, rather than hard effects. As we know, knowledge that an element is inanimate

Feature	Application
Definiteness	POS-tagger
Case	POS-tagger
Animacy - NN	Animacy classifier
Animacy - PN	Named Entity Tagger
Animacy - PO	Majority class
Tense, voice	POS-tagger

Table 9.28: Overview of applications employed for automatic feature acquisition.

or indefinite, is often, in itself, not sufficient evidence for interpretation as, say, object. Additional knowledge with respect to the linguistic context of the element provides further knowledge. For instance, knowledge that an animate, definite nominal has been assigned subject status in preverbal position clearly makes an object relation for the inanimate element more likely.

9.4 Automatically acquired features

A possible objection to the general applicability of the results presented above is that the added information consists of gold standard annotation from a treebank. However, the morphosyntactic features examined here are for the most part straightforwardly derived (definiteness, case, tense, voice) and represent standard output from most part-of-speech taggers. In chapters 6 and 7, we showed that the property of animacy could be fairly robustly acquired for common nouns by means of distributional features from an automatically parsed corpus. In this section we investigate parsing with automatically acquired linguistic features.

9.4.1 Acquiring the features

The linguistic features may be acquired through the use of different NLP-applications and table 9.28 shows an overview of the applications employed for the automatic acquisition of our linguistic features. For part-of-speech tagging, we chose to employ MaltTagger – a HMM part-of-speech tagger for Swedish (Hall 2003). The pretrained model for Swedish employs the SUC tagset (Gustafson-Capková and Hartmann 2006), exemplified by the part-of-speech tagged version of (182) in (183) below.

(182) *Några har valts ut och med dem skall man nu*
some have choose-PASS out and with them shall one now
börja slutförhandlingen
start negotiation-DEF
'Some have been chosen and we will now commence negotiations with them'

(183) Example (182); part-of-speech tagged with SUC tagset:

`Några`	`dt.utr/neu.plu.ind`	
`har`	`vb.`**`prs`**`.`**`akt`**`.aux`	⇒ tense, voice
`valts`	`vb.`**`sup`**`.`**`sfo`**	⇒ tense, voice
`ut`	`pl`	
`och`	`kn`	
`med`	`pp`	
`dem`	`pn.utr/neu.plu.def.`**`obj`**	⇒ case
`skall`	`vb.`**`prs`**`.`**`akt`**`.mod`	⇒ tense, voice
`man`	`pn.utr.sin.ind.`**`sub`**	⇒ case
`nu`	`ab`	
`börja`	`vb.`**`inf`**`.`**`akt`**	⇒ tense, voice
`slutförhandlingen`	`nn.utr.sin.`**`def`**`.nom`	⇒ definiteness, case

The SUC part-of-speech tag set distinguishes tense and voice for verbs, nominative and accusative case for pronouns, as well as definiteness and genitive case for nouns. The experiments with the individual verbal features described in section 9.2.6 clearly showed the benefit of mapping the tense values to a binary set of finiteness features and this mapping was performed directly for the acquired features.[188]

The experiments with features expressing pronoun type, described in section 9.2.4 above, showed that the effect of this feature was largely due to the treatment of non-referential pronouns. Acquisition of non-referentiality is not a trivial task, although it has recently been approached with machine-learning (Boyd, Gegg-Harrison and Byron 2005). Given the fairly modest impact of this feature, however, acquisition of non-referentiality is not pursued further in the present context.

9.4.1.1 Animacy

The feature of animacy is clearly the most challenging feature to acquire automatically. Recall that Talbanken05 distinguishes person reference for all nominal constituents, and as shown in section 7.1.2, 97.8% of the nominal treebank

[188] Present, past, imperative and subjunctive forms are mapped to the finite feature (FV), all other forms are mapped to the non-finite feature (Ø).

instances annotated as animate are nouns and pronouns. Hence we will in the following focus on automatic animacy annotation for nouns and pronouns.

Common nouns

The animacy classifier developed in chapter 7 classifies common nouns based on their syntactic distribution in the Swedish Parole corpus. Whereas the gold standard classes are employed for training of the classifier, the distributional data is taken from the considerably larger, automatically parsed Parole corpus. The common nouns in Talbanken05 are classified for animacy following a leave-one-out training and testing scheme where each of the n nouns in Talbanken05 are classified with a classifier trained on $n-1$ instances.[189] This ensures that the training and test instances are disjoint at all times. Moreover, the fact that the distributional data is taken from a separate data set ensures non-circularity since we are not basing the classification on gold standard parses.

Proper nouns

In the task of named entity recognition (NER) (Tjong Kim Sang 2002b), proper nouns are classified according to a set of semantic categories (see, e.g., Chinchor et al. 1999). For the annotation of proper nouns, we make use of a named entity tagger for Swedish (Kokkinakis 2004), which is a rule-based tagger based on finite-state rules, supplied with name lists, so-called "gazetteers". The tagger distinguishes the category 'Person' for human referring proper nouns and we extract information on this category.

Pronouns

In section 6.2.3 we extracted information on pronominal reference to nouns based on simple heuristics with respect to a set of pronouns and syntactic position (the ANAAN/ANAIN features). Recall that a subset of the personal pronouns in Scandinavian, as in English, clearly distinguish their referent with regard to animacy, e.g. *han, det* 'he, it'. There is, however, a quite large group of third person plural pronouns which are ambiguous with regards to the animacy of their referent. The ambiguous pronouns include the personal pronouns, e.g., *de, dem, deras* 'they, them, theirs' , demonstrative pronouns, e.g, *dessa* 'these', as well as quantifying pronouns like *bägge, alla, många* 'both, all, many'.

[189] We employ the MBL_{opt} classifier described in section 7.3.3.

Dimension	Features	Instances	Correct	Accuracy
Definiteness	DD, Ø	40832	40010	98.0
Case	GG, AA, Ø	68313	67289	98.5
Animacy$_{NNPNPO}$	HH, Ø	68313	61295	89.7
Animacy$_{NN}$	HH, Ø	40832	37952	92.9
Animacy$_{PN}$	HH, Ø	2078	1902	91.5
Animacy$_{PO}$	HH, Ø	25403	21441	84.4
Finiteness	FV, Ø	30767	30035	97.6
Voice	PA, Ø	30767	29805	96.9

Table 9.29: Accuracy for automatically acquired linguistic features.

The pronominal part-of-speech tags from the part-of-speech tagger distinguish number and gender and in the animacy classification of the personal pronouns we classify based on these tags only. We employ a simple heuristic where the pronominal tags which had more than 85% human instances in the gold standard are annotated as human.[190] This gives us the personal non-neuter pronouns, like *vi, oss, han, du, man* 'we, us, he, you-SG, one', as well as the set of genitive pronouns, like *din, min, sina* 'your, mine, theirs', as animate (HH).[191] The pronouns which are ambiguous with respect to animacy are not annotated as animate (Ø).

In table 9.29 we see an overview of the accuracy of the acquired features, i.e., the percentage of correct instances out of all instances. Note that we adhere to the general annotation strategy in Talbanken05, where each dimension (definiteness, case etc.) contains a null category Ø, which expresses the lack of a certain property.

Many of the dimensions exhibit quite skewed distributions, hence in table 9.30, we present the class-based measures of precision and recall for each of the non-null features. Acquisition of morphological definiteness for common nouns is clearly reliable, with an overall accuracy of 98.0, despite a skewed distribution of classes. Precision and recall for the definite feature (DD) is 97.7 and 96.0, respectively. With respect to case, a property of nouns and pronouns, we find an overall accuracy of 98.5, as table 9.29 shows. However, the genitive and accusative case features are seriously outnumbered by the set of null, or

[190] A manual classification of the individual pronoun lemmas was also considered. However, the treebank has a total of 324 different pronoun forms, hence we opted for a heuristic classification of the part-of-speech tags instead.

[191] We manually excluded the third person non-neuter pronoun *den* 'it' from this group of human-referring pronouns.

Feature	Gold	Automatic	Correct	Precision	Recall
DD	14094	13924	13598	97.7	96.5
GG	3756	3414	3321	97.3	88.4
AA	1745	2180	1707	78.3	97.8
HH	16875	10777	10317	95.7	61.1
HH_{NN}	6010	3538	3334	94.2	55.5
HH_{PN}	1056	920	900	97.8	85.2
HH_{PO}	9809	6319	6083	96.3	62.0
FV	20818	20560	20371	99.1	97.9
PA	2413	3067	2259	74.0	93.6

Table 9.30: Class precision and recall for automatically acquired linguistic features compared to gold standard.

nominative, instances. As table 9.30 shows, acquisition of genitive case shows a somewhat lower recall of 88.4. For accusative case we observe the opposite situation where, the part-of-speech tagger is overgenerating compared to the gold standard.

It is not surprising that we observe the largest discrepancies from the gold standard annotation in the automatic animacy annotation. In general, the annotation of animate nominals exhibits a decent precision (95.7) and a lower recall (61.3). The automatic classification of human common nouns also has a quite high precision (94.2) in combination with a lower recall (55.5). As we noted in chapter 7, this is an advantage provided the skewed distribution of the classes in the corpus, since it indicates that the classifier is conservative in terms of class assignment to the minority class. The named-entity recognizer shows more balanced results with a precision of 97.8 and a recall of 85.2 and the heuristic classification of the pronominal part-of-speech tags gives us high precision (96.3) combined with lower recall (62.0) for the animate class.

Just as for the other morphological features, the acquisition of the verbal features of finiteness and voice from the part-of-speech tagger is very reliable, with accuracies of 97.6 and 96.9, respectively. The passive feature is infrequent and shows a quite low precision (74.0) due to syncretism in the *s*-suffix which is employed for both passives and deponent verbs.

9.4.2 Experiments

The experiments assess the extent to which we may obtain the same effect from the linguistic information with automatically acquired features. This is an important part of assessing the scalability of the results discussed above.

	Gold standard		Automatic	
	Unlabeled	Labeled	Unlabeled	Labeled
NoFeats	89.87	84.92	89.87	84.92
Def	89.87	85.02	89.88	85.03
Case	89.99	85.13	89.95	85.11
Finite	90.24	85.33	90.15	85.23
Voice	89.81	84.97	89.83	85.00
Anim	89.93	85.10	89.86	85.01
Anim_{NN}	89.81	84.94	89.86	84.99
Anim_{NNPN}	89.85	84.98	89.85	84.97
ADC	90.13	85.35	90.01	85.21
ADCV	90.40	85.68	90.27	85.54

Table 9.31: Overall results in experiments with automatic features compared to gold standard features, expressed as unlabeled and labeled attachment scores.

9.4.2.1 Experimental methodology

The experimental methodology is identical to the one described in 9.2.1 above, the only difference being that the linguistic features are acquired automatically, rather than being gold standard. As before, all experiments are performed using 10-fold cross-validation on the written part of Talbanken05 and the feature model is the extended feature model in figure 13. In order to enable a direct comparison with the results from the earlier experiments, we employ the gold standard part-of-speech tags, as before. This means that the set for which the various linguistic features are defined is identical, whereas the feature values may differ.

9.4.2.2 Results

Table 9.31 presents the overall results with automatic features, compared to the gold standard results.[192] As expected, we find that the effect of the automatic features is generally less pronounced compared to the gold standard counterparts. However, all automatic features improve significantly on the NoFeats baseline. In the error analysis we find the same tendencies in terms of improvement for specific dependency relations and error types.

[192]The results for the gold standard combined experiments ADC and ADCV in table 9.31, are somewhat lower than the combined results presented in section 9.2.7, since the former experiments do not include the pronoun type feature.

Morphological features

The morphological argument features from the POS-tagger are reliable, as we saw above, and we observe almost identical results to the gold standard results. The addition of information on definiteness causes a significant improvement ($p<.01$), and so does the addition of information on case ($p<.0001$). The improvement in terms of performance for specific dependency relations is also almost identical, with only small, non-significant variations. As before, the addition of information on definiteness causes the largest effect in terms of performance for the SP relation, as well as improvement for the SS and OO relations. Case information benefits the analysis for determiners and prepositional complements, but also argument relations such as SS, OO, SP and IO.

In parallel with the gold standard results, we find that the single feature which has the most notable effect on performance is the feature of finiteness ($p<.0001$). It influences the analysis of the argument relations, as well as the verbal relations.

Animacy

The addition of the automatically acquired information on animacy shows some interesting results. First of all, the addition of all acquired animacy information for nouns and pronouns (Anim) causes a significant improvement ($p<.03$), even though it is smaller than in the gold standard experiment. We find that the OO and SS relations are the dependency relations which exhibit the largest improvement. A clear difference from the gold standard experiment, however, resides in the performance for the DT-relation, where performance actually deteriorates slightly.[193] This is largely due to the set of plural pronouns mentioned above, which are ambiguous with respect to the animacy of their referent. In the gold standard, however, their animacy in the specific context has been manually determined. These pronouns may furthermore function as determiners, in which case they are never annotated as animate. Consequently, animacy serves as an indicator of clausal as opposed to phrasal argument status which is not provided with the automatic annotation.

We may examine the effect of the different sources of animacy information, i.e. the animacy information supplied for common nouns, proper nouns and pronouns, by examining their effect on parse results. As the results in table 9.31 indicate, it is the information supplied by the animacy classifier for common nouns which largely accounts for the improvement observed with

[193] F-scores for the DT relation go from 94.14 in the baseline to 94.09 in the Anim experiment with automatic features.

the addition of this feature. This is surprising since the recall for this feature is quite low. The addition of information only for common nouns in the Anim$_{NN}$-experiment causes a significant improvement in overall results ($p<.04$). In the corresponding gold standard experiment, the results are not significantly better than the baseline and the main, overall, improvement clearly stems from the animacy annotation of pronouns. This indicates that the animacy information for common nouns, which has been automatically acquired from a considerably larger corpus, captures distributional distinctions which are important for the general effect of animacy and furthermore that the differences from the gold standard annotation prove beneficial for the results.

An error analysis shows that the performance of the two parsers with respect to argument relations is very similar and we observe an improved analysis for the SS, OO, SP, IO with only minor variations.[194] This in itself is remarkable, since the covered set of animate instances is notably smaller in the automatically annotated data set, as shown by table 9.30 above. We furthermore find that the main difference between the gold standard and automatic Anim$_{NN}$ experiments does not reside in the analysis of arguments, but rather of non-arguments. One relation for which performance deteriorated with the added information in the gold Anim$_{NN}$ experiment is the nominal postmodifier relation (ET) which is employed for relative clauses and nominal PP-attachment. With the automatically assigned feature, in contrast, we observe an improvement in the performance for the ET relation, compared to the gold standard experiment, from a F-score in the latter of 76.14 to 76.40 in the former. Since this is a quite common relation, with a frequency of 5% in the treebank as a whole, the improvement has a clear effect on the results.

The analysis of postnominal modification is influenced by the differences in the added animacy annotation for the nominal head, as well as the internal dependent. If we examine the corrected errors in the automatic experiment, compared to the gold standard experiment, we find elements with differing annotation. In general, the relation of postnominal modification disprefers attachment to animate nominals. Consider (184)–(185) below which illustrate corrected errors of the types ET_OA and ET_AA, respectively. The nominal heads in these constructions, *vän* 'friend' and *kandidater* 'candidates', are instances which are annotated as animate in the gold standard, but inanimate in the automatically classified data set. The automatic annotation as inanimate results in a correct attachment and labeling of the modifiers, a relative clause in (184) and the head preposition *till* 'to' in (185).

[194] The gold standard Anim$_{NN}$ results exhibit slightly better performance for the SS and SP relations, whereas the automatic Anim$_{NN}$ results show slightly better performance for the OO and IO relations.

(184) *De flesta vill trots allt ha en riktig* **vän** *att hålla*
the most want after all have a real friend to hold
ihop med
together with
'Most people, after all, want a real friend to be together with'

(185) *För* **kandidater** *till landsting och kommunfullmäktige gäller*
for candidates to municipals and boards holds
fortfarande bostadsbandet
still residence-restriction-DEF
'With respect to candidates for municipal boards, the restriction on residence still holds'

We also observe an effect of differing annotation for the nominal dependent in prepositional ET constructions. Preferences with respect to animacy of prepositional complements vary, as we noted in section 7.1.2 and illustrated with table 7.5 on page 137. In (186), the automatic annotation of the noun *djur* 'animal' as animate results in correct assignment of the ET relation to the preposition *hos* 'among', as well as correct nominal, as opposed to verbal, attachment. This preposition, as we recall, is one of the few with a preference for animate complements. In contrast, the example in (187) illustrates a ET_OA error, where the automatic classification of *barn* 'children' as inanimate causes a correct analysis of the head preposition *om* 'about'.

(186) . . . *mer permanenta samhällsbildningar hos olika* **djur**
. . . more permanent societies at different animals
'. . . more permanent social organizations among different animals'

(187) *Föräldrar har vårdnaden om sina* **barn**
parents have custody-DEF of their children
'Parents have the custody of their children'

A more thorough analysis of the different factors involved in PP-attachment is a complex task which is clearly beyond the scope of the present study. We may note, however, that the distinctions induced by the animacy classifier based purely on linguistic evidence proves useful for the analysis of both arguments and non-arguments.

Selectional restrictions revisited

In section 9.2.8, we investigated enrichment of the treebank annotation by extension to verbal classes of selectional restrictions centered around the notion

of animacy. We concluded that the added information caused more damage than good, in spite of modest improvements for argument relations. In particular, we found deterioration in attachment caused by the addition of verbal class information.

We noted earlier that a larger data set might provide more reliable generalizations regarding verbal semantics. We therefore extracted selectional restrictions from the automatically tagged and parsed version of the Parole corpus, where animacy was assigned automatically, as detailed in section 9.4.1 above. The selectional restrictions were otherwise extracted and selectional association was calculated in a manner identical to the one described in section 9.2.8. The only restriction placed on the extraction was a frequency threshold of 10 overall instances in the Parole corpus. Clearly the data employed is considerably more noisy, relying on fully automatic annotation. An inspection of the resulting classification shows that the noise in the data influences the selectional associations. We examined the selectional association scores acquired for the verbs *skriva* 'write' and *visa* 'show' which we discussed in section 9.2.8. These indicate that associations with the animate class in general are considerably lower than under treebank acquisition. This is not surprising since we rely on automatic animacy classification with a quite low recall. However, we also know that the classifier has fairly good precision, so we may assume that the quality of these restrictions is reasonable despite the variation. For the subject argument of the verb *skriva* 'write', we find that the association with the animate class is 0.76, compared to 0.95 in the gold standard experiment.

For the parse experiments we set a selectional association threshold of 0.75 in order to take into account the noise in the data. This gives us a very high coverage of the treebank verbs, unlike the previous experiments. With a threshold of 0.75, as many as 91.7% of the verb tokens receive a class. A closer look at the classes shows that several reasonable distinctions are captured, exemplified by (188)–(191) which show Talbanken05 verbs from the different classes with a threshold of 0.75:

(188) `AnimSS`: *lära* 'learn', *berätta* 'tell', *hitta* 'find', *märka* 'notice', *jobba* 'work'

(189) `InanSS`: *gälla* 'concern', *händer* 'happen', *kosta* 'cost', *betyder* 'mean', *minska* 'lessen'

(190) `AnimOO`: *bry* 'bother', *gifta* 'marry', *förlåta* 'forgive', *älska* 'love', *umgås* 'socialize'[195]

[195]The class of `AnimOO` include a group of so-called *deponent* verbs, characterized by a passive *s*-suffix, but which have an agentive semantics. Examples include *hoppas* 'hope', *trivas* 'enjoy'. These have been part-of-speech tagged as passives, hence their subject has been recorded as an object in terms of selectional restrictions.

	Unlabeled	Labeled
NoFeats	89.87	84.92
Anim	89.86	85.01
$SR_{0.75}$	89.89	84.91
$SR_{0.75}$&Anim	89.92	84.95

Table 9.32: Overall results for experiments with selectional restrictions acquired from the Parole corpus with automatically acquired animacy information.

(191) `InanOO`: *utföra* 'execute', *göra* 'do', *söka* 'seek', *ge* 'give', *veta* 'know'

The `InanOO` class is in clear majority and is assigned to 84.8% of the tokens. Inaccuracy in the annotation for the inanimate class along with overgeneration of passives for verbs, as discussed above, is the cause of this overgeneration.

We perform two parse experiments with the acquired selectional restrictions, one with only verbal classes ($SR_{0.75}$) and one with the acquired animacy information as well ($SR_{0.75}$&Anim). Table 9.32 shows the results which do not differ significantly from the baseline. We may note, however, that unlike the gold standard experiments, we observe an improved, rather than deteriorated, attachment accuracy, given by the unlabeled attachment score. This is most likely a result of the increased coverage of classification.

Feature combinations

In parallel with the results achieved with gold standard features, we observe an improvement of overall results compared to the baseline ($p<.0001$) and each of the individual features when we combine the features of the arguments (ADC; $p<.01$) and the argument and verbal features (ADCV; $p<.0001$).

Table 9.33 shows the dependency relations which improve the most in the ADCV-experiment and table 9.33 shows the ranked list of argument relations only. We may compare the results here with the corresponding information for the gold standard experiment ADPCV presented in table 9.12 above. We find that the ranked lists are nearly identical, but with overall somewhat lower results in the experiment with automatic features. We thus observe the same tendencies with the automatically acquired features. With respect to argument relations, we find improvement for all relations except the FS relation. This difference is clearly due to the fact that our set of automatic features does not include information on referentiality for pronouns.

	Freq	NoFeats	ADCV
SS	0.1105	90.25	91.32
OO	0.0632	84.53	86.10
DT	0.1081	94.14	94.67
VG	0.0302	94.65	96.44
PA	0.1043	94.69	95.06
ROOT	0.0649	86.71	87.26
+F	0.0099	52.07	55.27
SP	0.0297	84.82	85.80
MS	0.0096	63.35	66.06
AA	0.0537	68.70	69.04

	Freq	NoFeats	ADCV
SS	0.1105	90.25	91.32
OO	0.0632	84.53	86.10
SP	0.0297	84.82	85.80
AG	0.0019	73.56	81.02
FO	0.0009	56.68	65.38
VO	0.0007	72.10	83.12
VS	0.0006	58.75	68.75
ES	0.0050	71.82	72.60
IO	0.0024	76.14	76.29
OP	0.0011	27.91	30.77

Table 9.33: 10 most improved dependency relations with automatic ADCV features (left) and improved argument relations with automatic ADCV features (right), ranked by their weighted difference of balanced F-scores.

9.5 Summary of main results

The error analysis presented in chapter 8 revealed consistent errors in syntactic analysis, namely the confusion of argument functions, resulting from word order ambiguity and lack of case marking. In the current chapter, a set of experiments have been reported which examine the effect of various linguistically motivated grammatical features hypothesized to target these errors.

A set of *linguistic features* were formulated which capture different aspects of argument relations. The features provided approximations of linguistic dimensions shown to be involved in argument differentiation in a range of languages, as well as more language-specific properties of Scandinavian argument realization. An extended feature model enabled us to experiment with the addition of lexical information for arguments through features expressing animacy, definiteness, pronoun type and case. The experiments showed that each feature individually causes a significant improvement in terms of overall labeled accuracy, performance for argument relations, and error reduction for the specific types of errors performed by the baseline parser. Error analysis comparing the baseline parser with new parsers trained with individual features reveal the influence of these features on argument disambiguation. We find that animacy influences the disambiguation of subjects from objects, objects from indirect objects as well as the general distinction of arguments from non-arguments. Definiteness has a notable effect on the disambiguation of subjects and subject predicatives, and pronoun type distinguishes between referential and non-referential subjects. Information on morphological case shows a clear effect

in distinguishing between arguments and non-arguments, and in particular, in distinguishing nominal modifiers with genitive case. Experiments with features of the verb included information on tense and voice, and furthermore established the importance of the property of finiteness in parsing of Scandinavian. The final experiments combining all features, exhibited a cumulative effect of the linguistic features and also served to validate the choice of these features as important factors in argument disambiguation. The ADPCV experiment which combined information on animacy, definiteness, pronoun type, case and verbal features showed results which differed significantly from the baseline, as well as each of the individual experiments ($p<.0001$). We found clear improvements for the analysis of all argument relations and clear error reduction in terms of argument disambiguation. For the error types confusing subjects and objects (SS_OO, OO_SS), for instance, we observe a 44.6% and 46.0% error reduction compared to the baseline.

In section 9.2.8, we furthermore enriched the treebank annotation with selectional restrictions, a relational category expressed as a lexical semantic property of the verb which determines the animacy of its arguments. The study discussed in section 9.2.8, showed the importance of dealing with variation in restrictions and a probabilistic measure of selectional association allowed us to experiment with various levels of gradience for the selectional restriction classes. Experiments testing the effect of selectional restrictions acquired from the treebank, as well as restrictions acquired from an automatically annotated and considerably larger corpus, gave inconclusive results and we found no significant improvements compared to the simple addition of information on animacy. The experiments indicate that information on argument animacy can, and should, be utilized independently of selectional restrictions from the verb.

In section 9.3 we examined the effect of variations over properties of the *parser* on argument disambiguation. The application of a graph-based, data-driven dependency parser to the same data sets as earlier enabled a contrastive study of argument disambiguation. We observed significant improvements with the added information, however, the error analysis for argument relations highlighted the importance of conditioning on a rich linguistic context. Experiments with a local feature model for MaltParser further elucidated the relative influence of our features in argument disambiguation.

The *scalability* of the results was addressed in section 9.4. In contrast to the earlier experiments, the linguistic features employed during parsing were acquired automatically. We found that the results may largely be replicated with automatic features and a generic part-of-speech tagger. All added features gave significant improvements over the baseline and the tendencies in terms of error reduction for specific dependency relations were highly similar. We further-

more employed annotation from the animacy classifier developed in chapter 7 during parsing, and in this way externally evaluated the lexical information acquired there. The application of animacy information based purely on linguistic, distributional data proved to capture important distinctions which gave a performance which was as good as, and even slightly better than, the gold standard counterpart experiment.

10 CONCLUDING REMARKS

In the introduction to this thesis we set out to study linguistic factors involved in argument differentiation and made the initial assumption that these may be studied using data-driven methods which generalize over language data. A unifying theme in the work presented here has been the induction of linguistic generalizations from differential properties of syntactic arguments in language data. Consistent correlations between the morphosyntactic and semantic realization of arguments have been exploited in the lexical acquisition of animacy, which was the topic of chapters 6–7 and in argument disambiguation in syntactic parsing, which was the focus in chapters 8–9.

In this final chapter, we conclude the thesis by outlining its main contributions and directions for future work. We will in particular describe the main findings which unite the thesis, as well as more specific contributions internal to its two main parts.

10.1 Main contributions

The underlying methodological conviction expressed throughout this thesis has been an empiricist one, focusing on the essential role of language data in linguistic investigations. We have shown how data-driven, computational models of language can be employed for linguistic investigations and in turn how linguistic generalizations can improve on computational models.

The main contributions of the thesis are found in its attempt to unify insights from different subfields of linguistics, in particular theoretical and computational approaches. We have seen how the study of soft constraints and gradience in language can be carried out using data-driven models and have argued that these provide a controlled setting where different factors may be evaluated and their influence quantified. By focusing on empirical evaluation, we have come to a better understanding of the results and implications of data-driven models and we have shown how linguistic motivation in turn can lead to improved computational models. Data-driven models clearly benefit from linguistically informed feature selection and error analysis.

10.1.1 Lexical acquisition

The initial assumption made at the beginning of Part II is that there is a close relation between the syntactic distribution of nouns and their semantic properties; so close, in fact, that we may approach the latter by generalizing over the former. A corpus study of the distribution of human and inanimate nominal elements, detailed in section 7.1.2, confirms the central role of animacy in argument differentiation and shows significant distributional differences between the distinctions in argumenthood established in chapter 3.1: subject and object, core and non-core arguments, as well as argument and non-argument. We approach the task of animacy classification for nouns through data-driven, lexical acquisition based on morphosyntactic distributional data which capture exactly the tendencies in argument differentiation discussed above.

The task of animacy classification is not a widely studied one in computational linguistics, allthough it resembles other semantic classification tasks like named-entity recognition or verb classification. A main contribution of Part II is thus found in the definition of the classification task and the identification of several factors central to its performance. We have shown that classification performance is influenced by several factors, such as data representation and sparsity, the size of the data sets and their class distribution. Obtaining animacy data is another topic which has been dealt with extensively. We have reviewed and evaluated annotation schemes for animacy and addressed the inventory of classes through empirical investigations into the dimension of animacy and its gradience.

In chapters 6–7 we identified several factors which individually and in combination influence the classification results. We varied the feature representation of the nouns, from a small set of theoretically motivated features to a more general feature space. An accuracy of 95% was obtained on a small set of high frequency nouns with only seven morphosyntactic features, and we ascertained that backing off to a smaller set of the three most frequent features allowed us to maintain similar performance for nouns with considerably lower frequencies ($\sim$50). The scaling of the classification task to a larger data set extracted from an annotated treebank, highlighted the skewed distribution of the classes of animate and inanimate, showing an approximate 10-90 split in the data. An important part of dealing with the skewed class distribution was found in extending the feature space to include more information on each individual noun. We also saw how the size of the data set influenced the performance of the classifier, however, notwithstanding the influence of data sparsity. We obtain results for animacy classification, ranging from 97.3% accuracy to 94.0% depending on the sparsity of the data. With an absolute frequency threshold of 10, we obtain an accuracy of 95.4%, which constitutes a 50% reduction of

error rate. The classifier is conservative with respect to the minority class of animate instances and, with a frequency threshold of 1, it exhibits a precision of 79.1 and a recall of 40.5 for the animate class. The corresponding results for the majority class of inanimate elements is 94.6 and 99.0.

We initially defined the classification task as a binary one with the classes of 'animate' and 'inanimate'. Gradience in the animacy dimension was established through experiments with varying class granularity as well as a comparison of human and automatic annotation for the dimension of animacy. In section 6.6, we included a set of collective nouns denoting organizations in our data set. Results from a three-way classification experiment showed that these constitute a distinct group based on linguistic distribution due to their potential for both highly agentive, as well as mass-like readings. Clustering experiments with the same data set clearly supports a main distinction between animate and inanimate entities, with a gradience of the animate category which extends to the aforementioned group of organization nouns. A comparison of nouns which show gradient properties in the human annotation for animacy and in the automatic classification underline the fact that animacy classification strictly deals with animacy as a linguistic category. We find that 53.1% of inanimate entities which are misclassified as animate by the automatic classifier are elements of gradient animacy, such as animals, collective nouns and abstract or vague nouns. The treatment of animacy as a denotational property based only on linguistic evidence has lead to a consistent annotation which captures relevant information in the task of argument disambiguation.

Distributional features and a denotational treatment of animacy were argued to constitute prerequisites for acquisition of lexical preferences based on soft, probabilistic constraints on arguments. We have seen that proposed distinctions relevant to the animacy dimension may be explored employing machine learning. The extensive feature analysis performed throughout Part II has clearly shown the acquisition of these functional, distributional preferences. We conclude that statistical tendencies in argument differentiation with respect to the dimension of animacy supports automatic classification of unseen nouns and has been shown to be robust, generalizable across machine learning algorithms – both supervised and unsupervised – as well as scalable to larger data sets.

10.1.2 Parsing

Part III of this thesis was devoted to the study of argument disambiguation in data-driven syntactic parsing. The main goal of this part of the thesis was to investigate the influence of various linguistic features on argument disambiguation. We motivated the choice of a data-driven parser by the direct relationship

to frequency in language use which alleviates the need for explicit formulation of constraints on arguments. Moreover, a dependency representation enables acquisition of generalizations at the level of grammatical functions, abstracting away from specific, structural realizations, whilst limiting the structural assumptions to the minimal syntactic relation between a head and its dependent. For our experiments we employed MaltParser, a language-independent system for data-driven dependency parsing.

In order to enable a detailed study of the influence of different features, we developed an explicit methodology for error analysis of the results, whereby we manipulate sets of errors and in this way quantify improvement and deterioration of results. The results from a baseline parser were analyzed in chapter 8, where we employed only a limited set of features to represent tokens in the parse configuration: part-of-speech (POS), lexical form (FORM) and previously assigned dependency relations (DEPREL). We noted the acquisition of a range of generalizations regarding syntactic arguments based only on the distribution of these features in the data, such as the canonical ordering of arguments, categorial and lexical preferences with respect to argument realization and a reasonably good distinction of arguments from non-arguments. The error analysis also revealed consistent errors in argument assignment, and we determined that properties common to Scandinavian type languages, namely word order variation combined with little morphological marking, were largely responsible for these errors.

Following the initial error analysis presented in chapter 8, we performed a set of experiments with an extended feature model and linguistically motivated features. The features of animacy, definiteness and referentiality were motivated by linguistic studies employing typological, theoretical and psycholinguistic data and found to be important in argument differentiation, as presented in chapter 3. Furthermore, features representing case, tense and voice were features which approximated defining properties of argument realization in Scandinavian type languages, as presented in chapter 4. Each feature individually caused a significant improvement in terms of overall labeled accuracy, performance for argument relations, and error reduction for the specific types of errors performed by the baseline parser. We furthermore established that the replacement of verbal tense with the property of finiteness significantly improved the effect of verbal features. We also achieved a cumulative effect in the combination of the features which differed significantly from the baseline, as well as each of the individual experiments ($p<.0001$). Moreover, resulting error analyses revealed the acquisition of functional preferences for a range of argument relations and linguistic features in line with the observations in chapter 3.

Comparative experiments on identical data sets with a graph-based dependency parsed, MSTParser, gave significant improvements in results also here. Moreover, the results highlighted the importance of conditioning on the previous linguistic context for improved argument disambiguation, and lead us to experiment with feature locality in MaltParser. With a highly local feature model, where additional linguistic features were limited to candidate head and dependent during parsing, we found that our features gave clear improvements, however, lower effects in terms of argument disambiguation.

The features employed initially were gold standard features taken from the treebank annotation. The scalability of the results achieved with the gold standard annotation was addressed and largely confirmed in section 9.4. Similar, although slightly lower, results in terms of parse performance were achieved with a set of automatically acquired features taken largely from a generic part-of-speech tagger. We applied the animacy classifier developed in chapter 7 and found that it captured linguistic distinctions which proved important for the disambiguation of arguments. The addition of automatically acquired animacy information for common nouns resulted in a significant improvement of overall parse results and was shown to give as good, or even slightly better, results than the gold standard counterpart. Error analysis revealed the influence of gradient animacy categories.

10.1.3 Argument differentiation

The ability to distinguish between different types of arguments is central to syntactic analysis, and the way in which this is done is dependent on a range of interacting factors. In this thesis, we have approached the topic of argument differentiation by establishing a set of argument distinctions and a set of linguistic dimensions which we hypothesize to be correlated. We have furthermore argued that data-driven models can provide an elucidating perspective on the coupling of arguments and linguistic properties without an explicit expression of a set of constraints. Generalization over language data has shown consistent statistical tendencies in argument realization and we have seen how these may be employed to acquire linguistic categories.

In the initial chapter we posed two main research questions, repeated below:

1. How are syntactic arguments differentiated?
    - Which linguistic properties differentiate arguments?
    - How do linguistic properties interact to differentiate an argument?

2. How may we capture argument differentiation in data-driven models of language? What are the effects?

More than anything, we hope to have shown in this thesis that these are interesting and worthwhile questions. We have also provided some possible answers which highlight different aspects of argument differentiation.

First of all, we may conclude that differentiation of arguments takes place relative to several distinctions in argumenthood, such as the distinction between the two arguments of a transitive construction, the subject and object, the core and non-core arguments, as well as the main distinction between arguments and non-arguments. Our results emphasize this relative nature of argument differentiation. The most important distributional features employed for animacy classification, such as the SUBJ, OBJ and GEN features, provide information regarding important distinctions in argumenthood, hence environments which tend to exhibit differential properties. The parse experiments highlight this point further, and we find that error reduction in terms of argument disambiguation increases in line with conditioning on the linguistic context in terms of grammatical relations and linguistic features.

The formulation of question 1, which asks how arguments are differentiated, furthermore points to part of the answer, namely: through a set of linguistic properties. In this thesis we have identified several such properties, and, perhaps more importantly, attempted to explicate and evaluate the conditions under which these properties affect syntactic argumenthood. We have examined linguistic properties, such as animacy, definiteness and referentiality, which are relevant to a range of languages, as well as more language-specific properties relating to morphological and structural properties, such as case and finiteness. With respect to levels of linguistic analysis these properties represent a mixture, ranging from semantic and discourse-oriented properties, as well as morphosyntactic ones. We propose that the interaction of the linguistic properties is probabilistic and that the frequency distribution of linguistic properties relative to different distinctions in argumenthood in language data directly determines their importance in argument differentiation. These results are clearly compatible with a view of argumenthood as determined by a set of soft constraints, as suggested by theoretical and psycholinguistic work.

As mentioned above, one of the main contributions of this thesis is methodological and a large portion of this thesis has therefore been concerned with the second question posed initially, i.e., how argument differentiation may be captured in data-driven models. In the introduction, we proposed to employ theoretical proposals regarding argument differentiation to motivate the definition of data-driven learning problems and thereby to guide generalization from language data. In order to capture argument differentiation we have formulated

features which generalize over syntactic arguments as well as their linguistic properties.

In general, we have seen that a separation of a notion of argumenthood from specific morphosyntactic realization has been a key component in the data-driven modeling of argument differentiation. This was achieved through distributional features in Part II, which generalize over specific structural realization, such as word order. This representation was shown to capture statistical tendencies in argumenthood with respect to the linguistic property of animacy. In Part III, we studied argument differentiation in context through the study of argument disambiguation in a data-driven dependency parser for Swedish. The error analysis showed that further improvement of argument analysis was partly dependent on properties of argument realization other than word order and morphology. The separation of functional arguments from structural position which characterizes dependency analysis enabled the acquisition of functional generalizations irrespective of structural realization. For Scandinavian type languages, which are characterized by considerable word order variation and lack of morphological marking, the separation of function from structural realization constitutes an important property. The acquisition of soft, functional constraints is furthermore clear from the type of improvement which the added information incurred. We found improvement largely in labeled results caused by disambiguation of grammatical functions, rather than structural positions (attachment). For instance, for the errors confusing subjects for objects and vice versa, which were largely errors in labeling, we observed an error reduction of 44–46% in the experiments combining all features. We found that a majority of the improved errors were arguments which were non-canonical in some sense, i.e., departing from the most frequent structural and morphological properties. Improvement thus relied on other properties of argument relations and the abstraction over specific realization in terms of dependency relations.

The representation of the linguistic properties introduced in chapters 3–4 has also constituted an important part of the data-driven modeling of argument differentiation. We have throughout the thesis explicitly stated that data-driven modeling relies largely on approximation. The formulation and evaluation of the features proposed to approximate linguistic properties of arguments has therefore constituted a central part of the work described above.

10.2 Future work

A natural next step is to extend the studies performed here to other languages. The linguistic tendencies in argument differentiation presented in chapter 3

have been attested in a range of different languages. The work on acquisition of animacy described in Part II is based on linguistic generalizations relevant to a wide range of languages, for instance English (Zaenen et al. 2004). It would therefore be interesting to experiment with animacy classification for other languages. Properties of the Scandinavian languages which may be connected with errors in argument assignment, see part III, turn out not to be isolated phenomena. A range of other languages exhibit similar properties, for instance, Italian exhibits word order variation, little case, syncretism in agreement morphology, as well as pro-drop; German exhibits a larger degree of word order variation in combination with quite a bit of syncretism in case morphology; Dutch has word order variation, little case and syncretism in agreement morphology. These are all examples of other languages for which the results described here are relevant. Work on subject-object disambiguation for Italian suggests that a very similar approach might be worth pursuing also for this language (Dell'Orletta et al. 2005, 2006). The cross-linguistic generalizability of our results can be tested empirically by data-driven dependency parsing of other languages with motivated features.

Gradience in the animacy dimension was here addressed through experiments with a more fine-grained set of classes. However, one might also draw a more radical conclusion from the notion of gradience and dispense with discrete categories alltogether, opting for a continuous animacy dimension. As a first step, application of a soft clustering algorithm might provide an exploratory overview. We have furthermore treated animacy as a lexical property of nouns and adopted a denotational treatment of this property through type-level classification. This has been shown to be a viable approach through the application of the animacy classifier during parsing. However, we have also discussed in several places, the influence of the linguistic context on referential properties of nominal elements. An interesting possibility is thus to perform token level animacy classification, where type-level classification constitutes some sort of prior probability (Brew and Lapata 2004), and in addition incorporating the specific linguistic context.

In the automatic acquisition of features for parsing, detailed in section 9.4, we did not deal with the acquisition of non-referential subjects. As we have noted already, there are a range of quite different constructions which include non-referential subjects and the extent to which this is dependent on the argument structure of the verb also varies. This makes classification of non-referential subjects a challenging, but also interesting task since it provides a more fine-grained picture of argument properties.

Finally, the work described in this thesis has been largely concerned with syntactic arguments which are per definition subcategorized for by the verb. We have also discussed differentiation of arguments from non-arguments. The

error analysis for the baseline parser showed that confusion of different kinds of adverbials is a frequent error type in the parsing of Swedish, in addition to confusion of arguments. The case of adverbial disambiguation constitutes an area with several commonalities to that of argument disambiguation. First of all, adverbial placement is characterized by variation, in particular following the finite verb in Swedish (Andréasson 2007). Hence, an approach which operates with a separate level of functional analysis, like dependency grammar, may capture regularities irrespective of structural realization, much like the case of argument disambiguation.

REFERENCES

Aarts, Bas 2004. Conceptions of gradience in the history of linguistics. *Language Sciences* 26 (4): 343–389.

Abeillé, Anne (ed.) 2003. *Treebanks: Building and using parsed corpora*. Dordrecht: Kluwer Academic Publishers.

Ahrenberg, Lars 1990. A grammar combining phrase structure and field structure. *Proceedings of COLING-90*, Volume 2, 1–6.

Aissen, Judith 1999. Markedness and Subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17 (4): 673–711.

Aissen, Judith 2003. Differential Object Marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21 (3): 435–483.

Andréasson, Maia 2007. *Satsadverbial, ledföljd och informationsdynamik i svenskan*. Göteborgsstudier i Nordisk Språkvetenskap. Göteborg University.

Anttila, Arto 1997. Deriving variation from grammar: A study of Finnish genitives. Frans Hinskens, Roeland van Hout and W. Leo Wetzels (eds), *Variation, change and phonological theory*, 35–68. Amsterdam: John Benjamins.

Baker, Mark 1983. Objects, themes and lexical rules in Italian. Lori Levin (ed.), *Papers in Lexical-Functional Grammar*, 1–45. Indiana Univ. Ling. Club.

Baker, Mark 1997. Thematic roles and syntactic structure. Liliane Haegeman (ed.), *Elements of grammar: Handbook of generative syntax*, 73–137. Dordrecht: Kluwer Academic Publishers.

Baldwin, Timothy 2006. Data-driven methods for acquiring lexical semantics. Lecture notes from the ESSLLI 2006 Course on Data-Driven Methods for Acquiring Linguistic Information, Málaga, Spain.

Bard, Ellen Gurman, Dan Robertson and Antonella Sorace 1996. Magnitude Estimation of linguistic acceptability. *Language* 72 (1): 32–68.

Barlow, Michael and Suzanne Kemmer (eds) 2000. *Usage based models of language*. Stanford, CA: CSLI Publications.

Baroni, Marco 2007. Distributions in text. Anke Lüdeling and Merja

Kytö (eds), *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter.

Beaver, David I. and Hanjung Lee 2004. Input-output mismatches in OT. Reinhard Blutner and Hank Zeevat (eds), *Optimality theory and pragmatics*, 112–153. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.

Bikel, Daniel M. 2004. Intricacies of Collins' parsing model. *Computational Linguistics* 30 (4): 479–511.

Blaheta, Don and Eugene Charniak 2000. Assigning function tags to parsed text. *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL)*, 234–240.

Bock, J. Kathryn and Richard K. Warren 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21 (1): 47–67.

Bod, Rens 1995. Enriching linguistics with statistics: Performance models of natural language. Ph.D. diss., University of Amsterdam.

Bod, Rens 1998. *Beyond grammar: An experience-based theory of language*. Stanford, CA: CSLI Publications.

Bod, Rens, Jennifer Hay and Stefanie Jannedy 2003. Introduction. Rens Bod, Jennifer Hay and Stefanie Jannedy (eds), *Probabilistic linguistics*, 289–341. Cambridge, MA: MIT Press.

Boersma, Paul and Bruce Hayes 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32 (1): 45–86.

Boleda, Gemma 2007. Automatic acquisition of semantic classes for adjectives. Ph.D. diss., Pompeu Fabra University.

Boleda, Gemma, Toni Badia and Eloi Batlle 2004. Acquisition of semantic classes for adjectives from distributional evidence. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, 1119–1125.

Börjars, Kersti 1998. *Feature distribution in Swedish noun phrases*. Oxford: Blackwell Publishers.

Börjars, Kersti, Elisabet Engdahl and Maia Andréasson 2003. Subject and object positions in Swedish. Miriam Butt and Tracy Holloway King (eds), *Proceedings of the LFG03 conference*. Stanford, CA: CSLI Publications.

Bouma, Gerlof 2008. Starting a sentence in Dutch: A corpus study of subject- and object-fronting. Ph.D. diss., Groningen University.

Boyd, Adriane, Whitney Gegg-Harrison and Donna Byron 2005. Identifying non-referential *it*: A machine learning approach incorporating linguistically motivated features. *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, 40–47.

Branigan, Holly P., Martin J. Pickering and Mikihiro Tanaka 2008. Contributions of animacy to grammatical function assignment and word order production. *Lingua* 118 (2): 172–189.

Bresnan, Joan 2001. *Lexical-Functional Syntax*. Malden, Mass.: Blackwell Publishers.

Bresnan, Joan 2006. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. Sam Featherston and Wolfgang Sternefeld (eds), *Roots: Linguistics in search of its evidential base*. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and Harald Baayen 2005. Predicting the dative alternation. Gosse Bouma, Irene Kraemer and Joost Zwarts (eds), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, Shipra Dingare and Christopher D. Manning 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. Miriam Butt and Tracy Holloway King (eds), *Proceedings of the LFG01 conference*. Stanford, CA: CSLI Publications.

Bresnan, Joan and Jonni M. Kanerva 1989. Locative inversion in Chicheŵa: A case study of factorization in grammar. *Linguistic Inquiry* 20 (1): 1–50.

Bresnan, Joan and Tatiana Nikitina 2007. The gradience of the dative alternation. Linda Uyechl and Lian Hee Wee (eds), *Reality exploration and discovery: Pattern interaction in language and life*. Stanford, CA: CSLI Publications.

Brew, Chris and Mirella Lapata 2004. Verb class disambiguation using informative priors. *Computational Linguistics* 30 (1): 45–73.

Bröker, Norbert 1998. How to define a context-free backbone for DGs: Implementing a DG in the LFG formalism. *Proceedings of the COLING-ACL workshop on Processing of Dependency Grammars*, 29–38.

Buchholz, Sabine 2002. Memory-based grammatical relation finding. Ph.D. diss., Tilburg University.

Buchholz, Sabine and Erwin Marsi 2006. CoNLL-X shared task on multilingual dependency parsing. *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, 149–164.

Buch-Kromann, Matthias 2006. Discontinuous Grammar: A model of human parsing and language acquisition. Ph.D. diss., Copenhagen Business School.

Bybee, Joan and Paul Hopper (eds) 2001. *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Carlson, Greg 1980. *Reference to kinds in English*. New York: Garland Press.

Carreras, Xavier and Lluís Màrquez 2005. Introduction to the CoNLL-2005 shared task: Semantic Role Labeling. *Proceedings of CoNLL-2005*, 89–97.

Carroll, Glenn and Mats Rooth 1998. Valence induction with a head-lexicalized PCFG. *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 36–45.

Carroll, John 2000. Statistical parsing. Robert Dale, Hermann Moisl and Harold Somers (eds), *Handbook of natural language processing*, 525–543. New York/Basel: Marcel Dekker.

Carroll, John and Edward Briscoe 2002. High precision extraction of grammatical relations. *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, 134–140.

Chang, Chih-Chung and Chih-Jen Lin 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Charniak, Eugene 1996. Treebank grammars. *Proceedings of the 13th national conference on Artificial Intelligence (AAAI)*, 1031–1036.

Charniak, Eugene 1997. Statistical parsing with a context-free grammar and word statistics. *Proceedings of the 14th national conference on Artificial Intelligence (AAAI)*, 598–603.

Charniak, Eugene 2000. A maximum-entropy-inspired parser. *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL)*, 132–139.

Charniak, Eugene and Mark Johnson 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL)*, 173–180.

Chinchor, Nancy, Erica Brown, Liza Ferro and Patty Robinson 1999. 1999 Named Entity Recognition Task Definition. MITRE and SAIC. Version 1.4.

Choi, Hye-Won 2001. Phrase structure, information structure, and resolution of mismatch. Peter Sells (ed.), *Formal and empirical issues in Optimality Theoretic syntax*, 17–62. Stanford, CA: CSLI Publications.

Chomsky, Noam 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam 1975. *The logical structure of linguistic theory*. New York: Springer.

Chomsky, Noam 1981. *Lectures on government and binding*. Holland: Foris Publications.

Chomsky, Noam 1995. *The minimalist program*. Cambridge, MA: MIT Press.

Chrupała, Grzegorz and Josef van Genabith 2006. Using machine-learning to assign function labels to parser output for Spanish. *Proceedings of the COLING/ACL main conference poster session*, 136–143.

Collins, Michael 1996. A new statistical parser based on bigram lexical dependencies. *Proceedings of the 34th annual meeting of the Association for Computational Linguistics (ACL)*, 184–191.

Collins, Michael 1999. Head-driven statistical models for natural language parsing. Ph.D. diss., University of Pennsylvania.

Comrie, Bernard 1989. *Language universals and linguistic typology*. Chicago, IL: University of Chicago Press.

Covington, Michael A. 2001. A fundamental algorithm for dependency parsing. *Proceedings of the 39th annual ACM southeast conference*, 95–102.

Croft, William 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Croft, William 2003. *Typology and universals*. 2nd edition. Cambridge: Cambridge University Press.

Crystal, David 1967. English. *Lingua* 17 (1): 24–56.

Daelemans, Walter 1999. Memory-based language processing. *Journal for Experimental and Theoretical Artificial Intelligence* 11 (3): 287–467.

Daelemans, Walter and Antal van den Bosch 2005. *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, Walter, Antal van den Bosch and Jakub Zavrel 1999. Forgetting exceptions is harmful in language learning. *Machine Learning* 34 (1): 11–43.

Daelemans, Walter, Jakub Zavrel, Ko Van der Sloot and Antal Van den Bosch 2004. TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. Technical Report, ILK Technical Report Series 04-02.

Dahl, Östen 2000. Egophoricity in discourse and syntax. *Functions of Language* 7 (1): 37–77.

Dahl, Östen 2008. Animacy and egophoricity: Grammar, ontology and phylogeny. *Lingua* 118 (2): 141–150.

Dahl, Östen and Kari Fraurud 1996. Animacy in grammar and discourse. Thorstein Fretheim and Jeanette K. Gundel (eds), *Reference and referent accessibility*, 47–65. Amsterdam: John Benjamins.

Dalrymple, Mary 2001. *Lexical Functional Grammar*. New York: Academic Press.

Dell'Orletta, Felice, Alessandro Lenci, Simonetta Montemagni and Vito Pirrelli 2005. Climbing the path to grammar: A maximum entropy model of subject/object learning. *Proceedings of the 2nd Workshop on Psychocomputational Models of Human Language Acquisition*, 72–81.

Dell'Orletta, Felice, Alessandro Lenci, Simonetta Montemagni and Vito Pirrelli 2006. Probing the space of grammatical variation: Induction of cross-lingual grammatical constraints from treebanks. *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*, 21–28.

Diderichsen, Paul 1957. *Elementær dansk grammatik*. København: Gyldendal.

Dietterich, Thomas G. 1998. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation* 10 (7): 1895–1923.

Dik, Simon C. 1989. *The theory of functional grammar*. Dordrecht: Foris.

Dowty, David 1982. Grammatical relations and Montague grammar. Pauline Jacobson and Geoffrey K. Pullum (eds), *The nature of syntactic representations*, 79–130. Dordrecht: Reidel.

Dowty, David 1991. Thematic proto-roles and argument selection. *Language* 67 (3): 547–619.

Dryer, Matthew S. 1986. Primary objects, secondary objects and antidative. *Language* 62 (4): 808–845.

Eide, Kristin Mehlum 2008. Finiteness and inflection: The syntax your morphology can afford. Downloaded from http://ling.auf.net/lingBuzz on Dec 10, 2008.

Einarsson, Jan 1976a. *Talbankens skriftspråkskonkordans*. Dept. of Scandinavian languages, Lund University.

Einarsson, Jan 1976b. *Talbankens talspråkskonkordans*. Dept. of Scandinavian languages, Lund University.

Engdahl, Elisabet, Maia Andréasson and Kersti Börjars 2004. Word order in the Swedish midfield – an OT approach. Fred Karlsson (ed.), *Proceedings of the 20th Scandinavian Conference of Linguistics*. http://www.ling.helsinki.fi/kielitiede/20scl/proceedings.shtml.

Erk, Katrin 2007. A simple, similarity-based model for selectional preferences. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 216–223.

Faarlund, Jan Terje, Svein Lie and Kjell Ivar Vannebo 1997. *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Falk, Cecilia 1993. Non-referential subjects in the history of Swedish. Ph.D. diss., Dept. of Scandinavian languages, Lund University.

Fass, Dan 1988. Metonymy and metaphor: What's the difference? *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*, 177–181.

Fellbaum, Christiane (ed.) 1998. *Wordnet: an electronic lexical database*. Cambridge, MA: MIT Press.

Fillmore, Charles J. 1968. The case for case. Emmon Bach and Robert Thomas Harms (eds), *Universals in linguistic theory*, 1–88. New York: Holt, Rinehart and Winston.

Fraurud, Kari 1996. Cognitive ontology and NP form. Thorstein Fretheim and Jeanette K. Gundel (eds), *Reference and referent accessibility*, 65–88. Amsterdam: John Benjamins.

Frazier, Lyn 1985. Syntactic complexity. David Dowty, Lauri Karttunen and Arnold M. Zwicky (eds), *Natural language parsing*, 129–189. Cambridge: Cambridge University Press.

Garretson, Gregory, M. Catherine O'Connor, Barbora Skarabela and Marjorie Hogan 2004. *Optimal typology of determiner phrases coding manual*. Version 3.2. Boston University. Downloaded from http://people.bu.edu/depot/coding_manual.html on 02/15/2006.

Givón, Talmy 1984. *Syntax: A functional-typological introduction*. Amsterdam: John Benjamins.

Goldwater, Sharon and Mark Johnson 2003. Learning OT constraint rankings using a Maximum Entropy model. Jennifer Spenader, Anders Eriksson and Östen Dahl (eds), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–120.

Grimshaw, Jane 1990. *Argument structure*. Cambridge, MA: MIT Press.

Gundel, Jeanette K., Nancy Hedberg and Ron Zacharski 1993. Cognitive status and the form of referring expressions. *Language* 69 (2): 274–307.

Gustafson-Capková, Sofia and Britt Hartmann 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Dept. of Linguistics, Stockholm University.

Hagen, Kristin, Janne Bondi Johannessen and Anders Nøklestad 2000. A constraint-based tagger for Norwegian. Carl-Erik Lindberg and Steffen Nordahl Lund (eds), *Proceedings of the 17th Scandinavian Conference of Linguistics*.

Hale, John and Eugene Charniak 1998. Getting useful gender statistics from English text. Technical Report, Comp. Sci. Dept. at Brown University, Providence, Rhode Island.

Hall, Johan 2003. A probabilistic part-of-speech tagger with suffix probabilities. Master's thesis, Växjö University, Sweden.

Haspelmath, Martin 2006. Against markedness (and what to replace it with). *Journal of Linguistics* 42 (1): 25–70.

von Heusinger, Klaus 2002. Specificity and definiteness in sentence and discourse structure. *Journal of Semantics* 19 (3): 245–274.

Hobbs, Jerry R. 1976. Pronoun resolution. Technical Report, City College of New York.

Holmberg, Ander 1986. Word order and syntactic features in the Scandinavian languages and English. Ph.D. diss., Dept. of General linguistics, University of Stockholm.

Holmberg, Anders and Christer Platzack 1995. *The role of inflection in Scandinavian syntax*. New York/Oxford: Oxford University Press.

Holmes, Philp and Ian Hinchliffe 2003. *Swedish: A comprehensive grammar*. London: Routledge.

de Hoop, Helen and Monique Lamers 2006. Incremental distinguishability of subject and object. Leonid Kulikov, Andrej Malchukov and Peter de Swart (eds), *Case, valency and transitivity*. Amsterdam: John Benjamins.

Hopper, Paul J. and Sandra A. Thompson 1980. Transitivity in grammar and discourse. *Language* 56 (2): 251–299.

Hoste, Véronique 2005. Optimization issues in machine learning of coreference resolution. Ph.D. diss., University of Antwerp.

Hudson, Richard 1990. *English word grammar*. Oxford: Blackwell Publishers.

Hundt, Marianne 2004. Animacy, agentivity and the spread of the progressive in Modern English. *English Language and Linguistics* 8 (1): 47–69.

Jäger, Gerhard 2004. Learning constraint sub-hierarchies: The bidirectional Gradual Learning Algorithm. Reinhard Blutner and Hank Zeevat (eds), *Optimality theory and pragmatics*, 251–287. Houndmills, Basingstoke, Hampshire: Palgrave/Macmillan.

Jäger, Gerhard and Anette Rosenbach 2006. The winner takes it all – almost: Cumulativity in grammatical variation. *Linguistics* 44 (5): 937–971.

Joanis, Eric and Suzanne Stevenson 2003. A general feature space for automatic verb classification. *Proceedings of the 10th Conference of the European Association for Computational Linguistics (EACL)*, 163–70.

Johannessen, Janne Bondi 1998. Tagging and the case of pronouns. *Computers and the Humanities* 32 (1): 1–38.

Johansson, Richard and Pierre Nugues 2007. Extended constituent-to-dependency conversion for English. Joakim Nivre, Heiki-Jaan Kaalep and Mare Koit (eds), *Proceedings of NODALIDA 2007*, 105–112.

Johnson, Mark 1998. PCFG models of linguistic tree representations. *Computational Linguistics* 24 (4): 613–632.

Jurafsky, Dan 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. Rens Bod, Jennifer Hay and Stefanie Jannedy (eds), *Probabilistic linguistics*, 289–341. Cambridge, MA: MIT Press.

Kager, René 1999. *Optimality Theory*. Cambridge: Cambridge University Press.

Kaplan, Ronald M. and Joan Bresnan 1982. Lexical-Functional Grammar: A formal system for grammatical representation. Joan Bresnan (ed.), *The mental representation of grammatical relations*, 173–281. Cambridge, MA: MIT Press.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä and Atro Anttila (eds) 1995. *Constraint Grammar: A language-independent system for parsing unrestricted text*. Berlin: Mouton de Gruyer.

Karypis, George 2002. Cluto: A clustering toolkit. Technical Report, Dept. of Computer Science, Univ. of Minnesota. Technical Report #02-017.

Katz, Jerrold J. and Jerry A. Fodor 1963. The structure of semantic theory. *Language* 39 (2): 170–210.

Keenan, Edward L. 1976. Towards a universal definition of "subject". Charles N. Li (ed.), *Subject and topic*, 303–333. Cambridge, MA: Academic Press.

Keenan, Edward L. and Bernard Comrie 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8 (1): 63–99.

Keller, Frank 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Ph.D. diss., University of Edinburgh.

Klein, Dan and Christopher D. Manning 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 423–430.

Kokkinakis, Dimitrios 2001. A framework for the acquisition of lexical knowledge: Description and application. Ph.D. diss., Department of Swedish Language, Göteborg University.

Kokkinakis, Dimitrios 2004. Reducing the effect of name explosion. *Proceedings of the LREC Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks*.

Kübler, Sandra 2004. *Memory-based parsing*. Amsterdam: John Benjamins.

Kübler, Sandra and Jelena Prokić 2006. Why is German dependency parsing

more reliable than constituent parsing? *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT)*, 7–18.

Kuhn, Jonas 2001. Generation and parsing in Optimality Theoretic syntax: Issues in the formalization of OT-LFG. Peter Sells (ed.), *Formal and empirical issues in Optimality-theoretic syntax*, 313–366. Stanford, CA: CSLI Publications.

Kuno, S. and E. Kaburaki 1977. Empathy and syntax. *Linguistic Inquiry* 8: 627–672.

Lakoff, George 1987. *Women, fire and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

Lakoff, George and Mark Johnson 1980. *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Lappin, Shalom and Stuart Shieber 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics* 43 (2): 393–427.

Levin, Beth 1993. *English verb classes and alternations*. Chicago, IL: University of Chicago Press.

Lin, Dekang 1998. Automatic retrieval and clustering of similar words. *Proceedings of the 17th International Conference on Computational Linguistics (COLING)*, Volume 2, 768–774.

Lyons, Cristopher 1999. *Definiteness*. Cambridge: Cambridge University Press.

Lyons, John 1977. *Semantics*. Cambridge: Cambridge University Press.

MacDonald, Maryellen C., Neal J. Pearlmutter and Mark S. Seidenberg 1994. Lexical nature of syntactic ambiguity resolution. *Psychological Review* 101 (4): 676–703.

Magerman, David M. 1995. Statistical decision-tree models for parsing. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 276–283.

Mak, Willem M., Wietske Vonk and Herbert Schriefers 2006. Animacy in processing relative clauses: The hikers that rocks crush. *Journal of Memory and Language* 54 (4): 466–490.

Manning, Christopher D. 2003. Probabilistic syntax. Rens Bod, Jennifer Hay and Stefanie Jannedy (eds), *Probabilistic linguistics*, 289–341. Cambridge, MA: MIT Press.

Manning, Christopher D. and Hinrich Schütze 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, M. P., B. Santorini and M. A. Marcinkiewicz 1993. Building a large

annotated corpus for English: The Penn treebank. *Computational Linguistics* 19 (2): 313–330.

Markert, Katja and Malvina Nissim 2006. Metonymic proper names: A corpus-based account. Anatol Stefanowitsch and Stefan Th. Gries (eds), *Corpus-based approaches to metaphor and metonymy*, 152–174. Berlin: Mouton de Gruyter.

Maruyama, Hiroshi 1990. Structural disambiguation with constraint propagation. *Proceedings of the 28th meeting of the Association for Computational Linguistics (ACL)*, 31–38.

McDonald, Ryan, Koby Crammer and Fernando Pereira 2005. Online large-margin training of dependency parsers. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 91–98.

McDonald, Ryan and Joakim Nivre 2007. Characterizing the errors of data-driven dependency parsing. *Proceedings of the Eleventh Conference on Computational Natural Language Learning (CoNLL)*, 122–131.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov and Jan Hajič 2005. Non-projective dependency parsing using spanning tree algorithms. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 525–530.

McEnery, Tony and Andrew Wilson 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Megyesi, Beáta 2002. Shallow parsing with PoS taggers and linguistic features. *Journal of Machine Learning Research* 2: 639–668.

Mel'čuk, Igor 1988. *Dependency syntax: Theory and practice*. Albany: State University of New York Press.

Merlo, Paola and Gabriele Musillo 2005. Accurate function parsing. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 620–627.

Merlo, Paola and Suzanne Stevenson 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics* 27 (3): 373–408.

Merlo, Paola and Suzanne Stevenson 2004. Structure and frequency in verb classification. *Proceedings of Incontro di Grammatica Generativa XXX*.

Mihalcea, Rada 2002. Instance based learning with automatic feature selection applied to word sense disambiguation. *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*.

Mihalcea, Rada 2006. Word sense disambiguation. Lecture notes from the ESSLLI 2006 Course on Word Sense Disambiguation, Málaga, Spain.

Mikkelsen, Line Hove 2002. Reanalyzing the definiteness effect: Evidence from Danish. *Working Papers in Scandinavian Syntax* 69: 1–75.

Mitchell, Tom M. 1997. *Machine learning*. New York: McGraw-Hill.

Morante, R. and B. Busser 2007. ILK2: Semantic role labeling of Catalan and Spanish using TiMBL. *Proceedings of the 4th International Workshop on Semantic Evaluations (SEMEVAL)*, 183–186.

Nilsson, Jens and Johan Hall 2005. Reconstruction of the Swedish treebank Talbanken. MSI report 05067, School of Mathematics and Systems Engineering, Växjö University.

Nilsson, Jens, Johan Hall and Joakim Nivre 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. *Proceedings of the NODALIDA Special Session on Treebanks*, 119–132.

Nivre, Joakim 2003. An efficient algorithm for projective dependency parsing. *Proceedings of the Eighth International Workshop on Parsing Technologies*, 149–160.

Nivre, Joakim 2004. Incrementality in deterministic dependency parsing. *Incremental parsing: Bringing engineering and cognition together. Workshop at ACL-2004*, 50–57.

Nivre, Joakim 2006. *Inductive dependency parsing*. Dordrecht: Springer.

Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel and Deniz Yuret 2007. CoNLL 2007 Shared Task on Dependency Parsing. *Proceedings of the CoNLL Shared Task session of EMNLP-CoNLL 2007*, 915–932.

Nivre, Joakim, Johan Hall and Jens Nilsson 2004. Memory-based dependency parsing. *Proceedings of the Eleventh Conference on Computational Natural Language Learning (CoNLL)*, 49–56.

Nivre, Joakim, Jens Nilsson and Johan Hall 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC)*, 1392–1395.

Nivre, Joakim, Jens Nilsson, Johan Hall, Gülşen Eryiğit and Svetoslav Marinov 2006. Labeled pseudo-projective dependency parsing with Support Vector Machines. *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

van Noord, Gertjan 2004. Error mining for wide-coverage grammar engineering. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 446–453.

Nunberg, Geoffrey 1979. The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy* 3: 143–184.

Orăsan, Constantin and Richard Evans 2001. Learning to identify animate references. *Proceedings of the Workshop on Computational Natural Language Learning*, 1–8.

Orăsan, Constantin and Richard Evans 2007. NP animacy resolution for anaphora resolution. *Journal of Artificial Intelligence Research* 29: 79–103.

Øvrelid, Lilja 2004. Disambiguation of syntactic functions in Norwegian: Modeling variation in word order interpretations conditioned by animacy and definiteness. Fred Karlsson (ed.), *Proceedings of the 20th Scandinavian Conference of Linguistics*. http://www.ling.helsinki.fi/kielitiede/20scl/proceedings.shtml.

Øvrelid, Lilja 2005. Animacy classification based on morphosyntactic corpus frequencies: Some experiments with Norwegian nouns. Kiril Simov, Dimitar Kazakov and Petya Osenova (eds), *Proceedings of the Workshop on Exploring Syntactically Annotated Corpora*, 24–34.

Øvrelid, Lilja 2006. Towards robust animacy classification using morphosyntactic distributional features. *Proceedings of the EACL 2006 Student Research Workshop*, 47–54.

Øvrelid, Lilja and Joakim Nivre 2007. When word order and part-of-speech tags are not enough – Swedish dependency parsing with rich linguistic features. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 447–451.

Platzack, Christer 1987. Huvudsatsordföljd och bisatsordföljd. Ulf Teleman (ed.), *Grammatik på villovägar*, 87–96. Solna: Esselte Studium.

Pollard, Carl and Ivan A. Sag 1994. *Head-driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Prat-Sala, Mercè and Holly P. Branigan 2000. Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language* 42 (2): 168–182.

Pustejovsky, James 1991. The generative lexicon. *Computational Linguistics* 17 (4): 409–441.

Quinlan, J. Ross 1993. *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Rahkonen, Matti 2006. Some aspects of topicalization in active Swedish declaratives: A quantitative corpus study. *Linguistics* 44 (1): 23–55.

Resnik, Philip 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61 (1): 127–159.

Rosenbach, Anette 2002. *Genitive variation in english: Conceptual factors in synchronic and diachronic studies*. Berlin/New York: Mouton de Gruyter.

Rosenbach, Anette 2003. Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. Günter Rohdenburg and Britta Mondorf (eds), *Determinants of grammatical variation in English*, 379–411. Berlin/New York: Mouton de Gruyter.

Rosenbach, Anette 2005. Animacy versus weight as determinants of grammatical variation in English. *Language* 81 (3): 613–644.

Rosenbach, Anette 2008. Animacy and grammatical variation - findings from English genitive variation. *Lingua* 118 (2): 151–171.

Sag, Ivan A. and Thomas Wasow 2008. Performance-compatible competence grammar. Robert D. Borsley and Kersti Börjars (eds), *Non-transformational theories of syntax*. Oxford: Blackwell Publishers.

Sag, Ivan A., Thomas Wasow and Emily M. Bender 2003. *Syntactic theory: A formal introduction*. 2nd edition. Stanford, CA: CSLI Publications.

Sahlgren, Magnus 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Ph.D. diss., Stockholm University.

Schröder, Ingo 2002. Natural language parsing with graded constraints. Ph.D. diss., Dept. of Computer Science, University of Hamburg.

Schröder, Ingo, Horia F. Pop, Wolfgang Menzel and Kilian A. Foth 2001. Learning grammar weights using genetic algorithms. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Schütze, Hinrich 1998. Automatic word sense discrimination. *Computational Linguistics* 24 (1): 97–122.

Seidenberg, Mark S. and Maryellen C. MacDonald 1989. A probabilistic constraints approach to language acquisition and processing. *Cognitive Linguistics* 23 (4): 569–588.

Sgall, Peter, Eva Hajicŏvá and Jarmila Panevová 1986. *The meaning of the sentence in its pragmatic aspects*. Dordrecht: Reidel.

Siewierska, Anna 1988. *Word order rules*. London: Croom Helm.

Silverstein, Michael 1976. Hierarchy of features and ergativity. Robert M.W. Dixon (ed.), *Grammatical categories in Australian languages*, 112–171. Canberra: Australian Institute of Aboriginal Studies.

Stevenson, Suzanne and Eric Joanis 2003. Semi-supervised verb class discovery using noisy features. *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 71–78.

Stevenson, Suzanne and Paul Smolensky 2005. Optimality in sentence pro-

cessing. Paul Smolensky and Geraldine Legendre (eds), *The harmonic mind*, 827–860. Cambridge, MA: MIT Press.

Sveen, Andreas 1996. Norwegian impersonal constructions and the unaccusative hypothesis. Ph.D. diss., University of Oslo.

de Swart, Peter 2007. Cross-linguistic variation in object marking. Ph.D. diss., Netherlands Graduate School of Linguistics.

de Swart, Peter, Monique Lamers and Sander Lestrade 2008. Animacy, argument structure and argument encoding: Introduction to the special issue on animacy. *Lingua* 118 (2): 131–140.

Teleman, Ulf 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Lund: Studentlitteratur.

Teleman, Ulf, Staffan Hellberg and Erik Andersson 1999. *Svenska Akademiens Grammatikk*. Stockholm: Nordstedts.

Tjong Kim Sang, Erik 2002a. Memory-based named entity recognition. *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 203–206.

Tjong Kim Sang, Erik F. 2002b. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 155–158.

Tjong Kim Sang, Erik F. and Sabine Buchholz 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*.

Schulte im Walde, Sabine 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32 (2): 159–194.

Schulte im Walde, Sabine 2007. The induction of verb frames and verb classes from corpora. Anke Lüdeling and Merja Kytö (eds), *Corpus linguistics. an international handbook*. Berlin: Mouton de Gruyter.

Wasow, Thomas, Amy Perfors and David Beaver 2005. The puzzle of ambiguity. Orhan Orgun and Peter Sells (eds), *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, 265–282. Stanford, CA: CSLI Publications.

Weber, Andrea and Karin Müller 2004. Word order variation in German main clauses: A corpus analysis. *Proceedings of the 20th International Conference on Computational Linguistics*, 71–77.

Weckerly, J. and M. Kutas 1999. An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology* 36: 559–570.

Yamada, Hiroyasu and Yuji Matsumoto 2003. Statistical dependency analysis with support vector machines. Gertjan Van Noord (ed.), *Proceedings of the Eighth International Workshop on Parsing Technologies (IWPT)*, 195–206.

Yamamoto, Mutsumi 1999. *Animacy and reference: A cognitive approach to corpus linguistics*. Amsterdam: John Benjamins.

Zaenen, Annie, Jean Carletta, Gregory Garretson, Joan Bresnan, Andrew Koontz-Garboden, Tatiana Nikitina, M. Catherine O'Connor and Tom Wasow 2004. Animacy encoding in English: why and how. Donna Byron and Bonnie Webber (eds), *Proceedings of the ACL Workshop on Discourse Annotation*.

Zeevat, Hank and Gerhard Jäger 2002. A reinterpretation of syntactic alignment. D. de Jongh, H. Zeevat and M. Nilsenova (eds), *Proceedings of the 3rd and 4th International Symposium on Language, Logic and Computation*. Amsterdam.

Zhao, Yang and George Karypis 2003. Criterion functions for document clustering. Technical Report, Dept. of Computer Science, Univ. of Minnesota.