UNIVERSITÄT DES SAARLANDES
Department of Computational Linguistics
- Master in Language Science and Technology -

A Hybrid Approach to Compound Noun Disambiguation

Masters Thesis

Antske Sibelle Fokkens

Department of Computational Linguistics
Saarbrücken 2006-2007

UNIVERSITÄT DES SAARLANDES MASTER'S THESIS ABSTRACT

Department of Computational Linguistics
- Master in Language Science and Technology -

Author:	Antske Sibelle Fokkens	
Title of thesis:		
A Hybrid Approach to Compound Noun Disambiguation		
Date:	June 14 2007	**Pages:** 104
Supervisor:	PD. Dr Evagelia Kordoni	
Corrector:	PD. Dr. Evagelia Kordoni Prof. Dr. Hans Uszkoreit	

This thesis investigates the automatic detection and identification of nominalisations. It extends research on using statistical methods for the interpretation of the grammatical relation existing between the modifier and head in compound nominalisations. Our basic method achieves an accuracy of 71% on a subject-object classification task, and 58% on a three-fold disambiguation task distinguishing subject, object and prepositional complement relations.

Using a database created from NOMLEX, CELEX and CATVAR, our detection method achieves recall of 94.5% and 99.5% on nominalisations detection in a set of compound nouns retrieved from open text.

We implement three methods that identify nominalisations from sets containing compounds. Best performance is achieved when only compounds whose modifier is attested as an argument of the head noun's root verb are identified as nominalisations. We achieve an accuracy of 70% and 76% using this method, over a baselines of 42% and 64%, respectively.

Keywords: Compound Nouns, Nominalisations
Hybrid methods, Statistics, WordNet, VerbNet
Language: English

Acknowledgements

We would like to thank Su Nam Kim and Timothy Baldwin for generously sending us their set of “seed verbs”.
Also, many thanks to Mirella Lapata and Jeremy Nicholson who kindly contributed their data to our study.

A lot of credit goes to the three annotators Rebecca Didran, Jeremy Nicholson and Mat Wilson, who were willing to annotate 1,000 (or 400) compounds during busy times. Without their contribution, the research presented in this thesis could not have been carried out.

Last but not least, we thank Valia, for her advice, useful comments and support.

Saarbrücken, Germany June 14th 2007

Antske Fokkens

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Compound Noun Disambiguation

This thesis proposes a hybrid approach to compound noun disambiguation. But what does *compound noun disambiguation* mean? In order to explain this, let us look at the term *compound noun disambiguation* itself. The meaning of this noun phrase is more than just the sum of the three individual nouns.

To begin with, the compound has a syntactic structure: we can either interpret the term as [compound][noun disambiguation], where first *noun* and *disambiguation* form a compound noun, and then this term forms a new compound with the noun *compound.* Under this interpretation, *compound noun disambiguation* could be rephrased as “noun disambiguation using compounds”.

The second possible structure, which leads to the interpretation meant here, is [compound noun][disambiguation]: *compound* and *noun* form a compound, which modifies *disambiguation.* In this case we are referring to the “disambiguation of compound nouns”. In this particular example, the intended meaning is clear, but a compound consisting of three nouns can have either structure in English.

When two or more nouns form a compound, an implicit semantic relation exists between the individual nouns composing the compound. Next to the syntactic structure, the meaning of a compound depends on this semantic relation. In the example of *compound noun disambiguation*, *compound noun* refers to a noun that forms a compound. The whole expression refers to the disambiguation of these nouns: *compound noun disambiguation* refers to *compound nouns* being disambiguated. When we talk about the implicit semantic relation in a compound noun, we refer to a description indicating how the modifying noun relates to the head of the compound. We will present a more detailed description of semantic relations in section 1.4.

The ambiguous nature of compound nouns has been shown to form a "stumbling block" in NLP tasks such as machine translation and information retrieval (Isabelle (1984)). Their disambiguation entails detecting the implicit semantic relation that exists between the compound's nouns, and, in case the compound consists of more than two nouns, finding the syntactic structure of the group.

Automatic detection of the implicit relations between English compounds is the topic of this thesis, while the disambiguation of the syntactic structure will not be addressed in this work.[1] We limit our research to compounds consisting of two nouns. This restriction can be justified by the fact that 88.4% of the compounds in the Wall Street Journal corpus and 90.6% of the compounds in the British National Corpus consist of two nouns only (Kim and Baldwin (2006)). Moreover, we assume that our probabilistic approach can easily be extended to find the syntactic structure.

The restriction to binary compounds is not the only scope limitation in this work. Disambiguating compounds is a challenging task for several reasons. The first challenge lies in the fact that the formation of compounds is highly productive. This is not only true for English, but for most languages in which compounds are found. Secondly, the disambiguation of compounds is particularly tricky in English, because there are no syntactic and hardly any morphological cues indicating the relation between the nouns: as mentioned above, the nouns are connected by an *implicit* semantic relation. Third, the interpretation of compound nouns may be influenced by pragmatic and con-

[1] For a detailed study on compound noun bracketing, see Lauer (1995).

textual factors.

Finally, compound noun disambiguation forms a challenging task because researchers do not agree on the kind of semantic relations that may be found in compound nouns. Since Downing (1977), it is assumed that the number of relations that may occur is theoretically infinite.

In short, disambiguating compound nouns is far from trivial. Therefore we limit the scope of our investigation and focus on compounds containing a grammatical relation, i.e. SUBJECT, OBJECT and PREPOSITIONAL COMPLEMENT. This subset of compounds is known by the name of nominalisations. Levi (1978) defines nominalisations as compound nouns where the head is derived from a verb, and the modifying noun corresponds to one of the arguments of this verb.[2]

Our work is based on statistical approaches proposed by Lapata (2002) and Nicholson (2005). We propose extensions to their systems which try to identify nominalisations. These extensions include integration of linguistic resources such as WordNet (Fellbaum (1998)) and VerbNet (Kipper et al. (2000)), and the integration of a second disambiguation system. The methodology used in this work is thus a hybrid approach to compound noun disambiguation.

Our thesis is structured as follows: this introductory chapter will present work in theoretical linguistics that has influenced later classification systems proposed in NLP.

The second chapter provides an overview of previous work in computational linguistics. We present work in a variety of disambiguation approaches, but focus on statistical approaches and nominalisations because our research is centred around these two topics.

The third chapter presents our disambiguation systems, and the experiments we carried out detecting and disambiguating nominalisations. The evaluation of these systems is presented in chapter 4, which is followed by the conclusion and discussion of future research.

[2] A more detailed description of nominalisations will be provided below.

1.2 Theoretical background

The subject of compound nouns has long been discussed in linguistics (e.g. Brugmann (1900) (for German), Bergsten (1911) and Jespersen (1942) (for English)). Due to space limitations, we have restricted ourselves to work that has been taken up in computational proposals.

Though many aspects of compound nouns have been examined, three discussion points may be considered to be the central questions addressed in theoretical linguistics on this subject (Levi (1978)). They are the following:

1 **The definition:** How can compound nouns be defined? What are their properties?

2 **The relations:** What are the relations that may exist between two nouns forming a compound? Is there a fixed number of possible relations?

3 **Their creation:** How are compound nouns formed? That is, what is their underlying structure?

More recently, computational linguistics and psycholinguistics have addressed a fourth essential question related to compound nouns:

4 **The interpretation:** How are compound nouns interpreted?

The purpose of the first two questions is clear. Answers to both can also be found in computational linguistic work, though they are generally not discussed in detail. The third question has mostly been of theoretical importance and was mainly used in the discussion between lexicalists and generative semantics. Generative semantics assumed that all nominalisations are derived from a verb and are desentential (Lees (1960), Lakoff (1970)). This idea has been contested by Chomsky's lexicalist hypothesis (Chomsky (1970)), claiming that — except for gerunds — nominalisations are listed in the lexicon and "stay" nouns throughout the derivation. This point of view has again been criticised by (McCawley (1975), Newmeyer (1971) and Ross (1974)).

Even though the lexicalist/generative question has occupied linguists from the transformational tradition for over 35 years (a recent discussion can be found in Roeper (2005), and Newmeyer (2005)), we will not address this issue in this thesis. We relate nominalisations to verbs in order to find an appropriate description of the meaning of the compound. We do not make any claims that nominalisations are derived from sentences in a "deeper" derivation level. Instead of the third question, we will address the fourth question related to the interpretation of compounds in the next chapter.

In this chapter we present some theoretical insights on the first two questions. The next section will discuss properties and the definition of compound nouns. As we will see, none of the criteria proposed are suited for automatic disambiguation. For that reason, we will adopt a definition coming from the computational side.

Section 1.4 presents proposals concerning existing semantic relations. We will present the classification systems proposed by Levi (1978) and Warren (1978), and discuss the approaches proposed by Finin (1980), Isabelle (1984) and Downing (1977), respectively. Finally, we present twenty classes based on the classification proposed by Barker and Szpakowicz (1998), which we will use in our own system.

1.3 Definition

In the introduction, we explained that compound nouns are sequences of two or more nouns linked by an implicit semantic relation. Most researchers explain compound nouns in similar terms (Finin (1980), Lapata (2002)), though some have tried to find a more clear-cut definition. An elaborate discussion on the difficulties of finding good criteria to define compounds can be found in Jespersen (1942). Lees (1960), Levi (1978) and Warren (1978) also address this question. Their definitions focus on the distinction between a compound noun and an adjective-noun sequence that looks like a compound on the sur-face. Some do not consider all noun-noun sequences compound nouns, and impose other restrictions clearer than Finin's "an implicit semantic relation existing between the two" (Finin (1980), p.1).

Three criteria have been proposed to be essential properties of compound nouns: *fronted stress*, *permanent aspect* and *semantic specialisation* (Levi (1978)). However, is has been known since Jespersen (1942) that none of these criteria form a necessary or sufficient condition for a definition of *compound nouns*. Levi shows that it is even difficult to find a compound that fits all three criteria. Moreover, none of the criteria is based on information that is present in the written text samples we use for our analysis. Therefore, we will not take these criteria in consideration, but adopt the following definition of compound nouns:

I. Definition of Compound Nouns:

> **A sequence of two or more nouns that are connected through an implicit modification relation and together function syntactically as a noun.**

Our definition can be seen as a combination of the definition given by Buckeridge and Sutcliffe (2002), namely:

> ...a sequence of two or more nouns which together function syntactically as a noun.

(Buckeridge and Sutcliffe (2002), p.1)

and the definition given by Finin (1980):

> ...sequences of two or more nouns related through modification.

(Finin (1980), p.1)

The first part of the definition includes almost any sequence of two nouns. The second part of the definition restricts our investigation to *endocentric* compounds, i.e., compounds that refer to a subset of the head noun's denotation (Levi (1978)). This excludes two categories of noun-noun sequences. The first category we will not treat in this work is that of noun pairs such as *speaker-listener* and *player-trainer*, where neither of the nouns forms the head. Secondly, metaphorical expressions are not considered endocentric compounds: *battle fatigue* does not refer to a kind of *fatigue*. Note that this means that we exclude (almost) all compound nouns according to Allerton's semantic specialisation definition mentioned above.

An additional category that we do not include in our study are noun-noun sequences in which the first noun functions as a name of the second noun. Thus, expressions such as *Centre Street* and *Amstel Hotel* will not be considered compounds.

1.4 Semantic Relations

The previous section provided a definition of compound nouns, according to which compound nouns are sequences of nouns related through modification. In this section we will examine the kind of modification relations that may exist between nouns forming a compound.

First, we will introduce the distinction between "grammatical" and "pure semantic" relations (Levi (1978)). This distinction is essential for our work which focuses on nominalisations. Next, we present the (finite) classifications proposed by Levi (1978) and Warren (1978), respectively. Two rather different proposals coming from early work in computational linguistics by Finin (1980) and Isabelle (1984) will also be discussed.

Nominalisations

Levi (1978) distinguishes nominalisations from other compounds. This distinction has been adopted by many researchers (see, for instance, Lapata

(2002)). The relation between nouns forming a nominalisation corresponds to a *grammatical relation*: the head noun is derived from a verb (its root verb) and the modifier can be seen as one of the root verb's arguments. The examples below illustrate the difference between nominalisations and other compounds. The examples in (1) are nominalisations accompanied by the grammatical relation suggested, (2) provides some examples of other compounds:[3]

(1) a. Career damage (OBJECT)
b. Blood circulation (SUBJECT)
c. Drug test (PREPOSITIONAL COMPLEMENT)

(2) a. Rehabilitation technique
b. Retirement age
c. Health economists

In some cases the head noun of a compound is derived from a verb, but the relation to its modifier is not grammatical. For instance, the head nouns in *antique-style frame* and *school forecast* can be related to verbs (*to frame*, and *to forecast*, respectively), but, unless the school is forecasting or something is being framed in an antique-style, these compounds cannot be interpreted by a grammatical relation.

In this thesis, the term *nominalisation* will only refer to compounds where the relation between modifier and head is a grammatical relation, thus compounds in which the modifier corresponds to the subject, object or prepositional complement of the head noun's root verb. Our terminology differs from that of Lapata (2002), who uses the term for any compound with a deverbal head.

In order to distinguish the relation within a nominalisation from those in other compounds, the term *semantic relation* will henceforth only refer to implicit relations that are not grammatical relations. The rest of this section presents a discussion on the nature of these relations.

[3] Examples taken from the British National Corpus.

Semantic Relations

Nominalisations form a well-defined class: there are subjects, objects and prepositional complements. Deciding on the number of possible semantic relations is not as straight-forward. In fact, Selkirk (1982) suggests that any effort to classify semantic relations outside grammatical relations is lost. Researchers do not even agree on the possibility to define all semantic relations in the first place. Jespersen (1942) is one of the first to express the idea that the number of possible relations is infinite. Levi (1978) and Warren (1978), on the other hand, claim that a small set of possible relations allows the description of all compounds. Both provide a (different) finite set that attempts to cover all possible semantic relations, based on a corpus study. However, even those researchers who support the idea that there is a fixed set of relations, do not seem to agree on the number and nature of relations. Levi and Warren's finite sets of relations have been criticised by (among others) Downing (1977), Spärck Jones (1983) and Bauer (1979). Bauer (1979) presents an extensive description of pragmatic factors that influence compound noun interpretation. Downing (1977) shows in her psychological study that people use pragmatic factors to create and interpret new semantic relations spontaneously. Her example *apple juice seat* (meaning the place at a table where the apple juice is standing) has been used by many afterwards to support the claim that the number of semantic relations is infinite (Selkirk (1982), Hobbs et al. (1993), Lapata (2002) and Copestake and Lascarides (1997) *inter alia*).

Many researchers have accepted the idea that the semantic relations that may occur in a compound form an open class, but for practical reasons, most computational approaches follow the intuition expressed by Copestake and Lascarides (1997) that most compounds can be interpreted by a finite set of nouns, though pragmatics can always introduce new relations and override the standard interpretation. In our work, we will rely on this assumption as well.

Most classifications used in NLP are based on theoretical work by Levi (1978), Warren (1978), and early computational approaches by Finin (1980) and Isabelle (1984). Due to the impact these works have made on computational analyses of compounds, we will briefly present them below.

Levi (1978) takes a generative point of view and claims that compound nouns are formed by two processes: change of category from verb to nouns (nominalisations) and deletions of predicates (nominal compounds). As mentioned above, she contradicts the claim made by (among others) Jespersen (1942) that the number of possible relations is endless. Levi proposes nine recoverable deletable predicates, that would allow to interpret any endocentric compound noun that is not a nominalisation. Most of the categories have also been used by other authors (for instance Koziol (1937), Jespersen (1942) and Brekle (1970) in English, Motsch (1970) in German), though none of these classifications corresponds exactly to the one proposed by Levi. Her nine deletable predicates are presented in table 1.1. According to Levi, all compounds that are not nominalisations should be interpretable with one of

Table 1.1: Levi's Deletable Predicates

Relation	**Description**	**Examples**
CAUSE	**Causative** (Direct and Indirect) Modifier is Cause or Causee	*drug death*, *malaria mosquito*
HAVE	**Possessive** Modifier is Possessor or Possessed	*student power*, *apple cake*
MAKE	**Productive, Compositional** Modifier is Maker or Made	*honeybee*, *chocolate bar*
USE	**Instrumental** Modifier is Used	*steam iron*, *voice vote*
BE	**Essive, Appositional** Modifier is Subset of Head	*pine tree*, *murder charge*
IN	**Locative** (Spatial or Temporal) Modifier is Location	*office friend*, *summer breeze*
FOR	**Purpose, Benefactive** Modifier is Purpose	*picture album*, *headache pills*
FROM	**Source, Ablative** Modifier is Source	*wood ash*, *sea breeze*
ABOUT	**Topic** Modifier is Topic	*abortion law*, *love story*

the predicates of this list. As we have seen above, this has been demonstrated to be wrong by Downing (1977) and others. Nevertheless, the classification proposed by Levi has been valuable to computational linguists. It manages to classify most compound nouns, and the reasons of membership of a specific class are well defined. Her predicates are found in several classifications, though these generally contain more than the originally proposed nine classes. One of the problems with Levi's proposal is that some relations work in two directions: in CAUSE, HAVE and MAKE, the modifier can either be seen as an agent or patient of the description. This leaves room of ambiguity in the interpretation. Splitting these deletable predicates up in two relations (one for each direction) is one of the reasons why later classifications tend to contain more than nine classes.

Around the same time, Warren (1978) presented a classification which differs mostly in organisation from Levi's proposal.
In her corpus study, Warren investigates whether there is a fixed number of semantic relations that may occur between two nouns forming a compound. She also addresses the number and kind of relations that can be found. She proposes a division with six major types of semantic relations, subdivided in more specific relations. Her six main classes and their direct subclasses are presented in table 1.2. The strength of Warren's approach is that it is based on a corpus study. Due to the fact that compounds can be created spontaneously, it can never be proven that a classification manages to cover the full range of possible data. However, it is clear that a classification that can identify all relations of 4,557 different noun-noun compounds found in a corpus provides enough coverage for general disambiguation tasks.
The problem with the classification in Warren (1978) is that the membership of a specific compound noun to a class is not always motivated. Explanations of why a certain noun is included in a specific class are hardly present in this work, and is often far from intuitive. For instance, Warren classifies *waterline* as POSSESSION. Naturally, classification depends on the intended meaning of this ambiguous word. The meaning that seems to come closest to a posses-

Table 1.2: Warren's Participant Roles

Relation	Mod. description	Subrelation	Examples
CONSTITUTE	Constitutes Head or Vice Versa	SOURCE-RESULT RESULT-SOURCE COPULA	*student group* *sheet metal* *girl friend*
POSSESSION	Possessor/Possessed	PART-WHOLE WHOLE-PART	*armchair* *eggshell*
LOCATION	Origin/Location, in Time or Space	PLACE-OBJECT TIME-OBJECT ORIGIN-OBJECT GOAL-OBJECT	*law degree* *Sunday paper* *hay fever* *moon rocket*
PURPOSE	Purpose of Head	GOAL-INSTRUMENT	*fireplace*
ACTIVITY-ACTOR	Activity/Interest	ACTIVITY-OBJECT	*cowboy*
RESEMBLANCE	Modifier Resembles Head	COMPARANT-COMPARED	*clubfoot*

sive relation is "line used to transport water". However, PURPOSE seems to be a more appropriate interpretation than PART-WHOLE. Classifying other meanings of *waterline* as possession relations seems even more far-fetched.
Another example is the classification of *football* as a PURPOSE relation. This seems better motivated than the *waterline* as possession, but still one could ask if the ball was really made for feet?
The absence of clear criteria deciding whether a compound exhibits a specific relation or not could explain why there is full coverage of the data: given that class membership may be as contrived as in the examples above, a compound can always be placed in some class.

Finin (1980) presents a classification system that refers to both Levi and Warren's proposals. At first sight, the classification proposed by Finin reveals similarities to these proposals. The main three relation types he distinguishes are idiomatic or exocentric relations, semantic relations, and structural relations. A similar distinction is found in Levi (1978), who sets idiomatic expressions apart from other compounds (leaving them out of scope of her work), and proposes a main distinction between nominalisations and deletable predicates.

However, most of the correspondences end here. Finin does respect the idea expressed by both Warren and Levi that there exists regularity in occurring relations, but rejects the idea that it would be possible to define them all. Notably, some of the semantic relations he proposes (such as *location* and *part-of*) correspond to relations proposed by Levi and Warren, but because he considers that semantic relations form an open class, Finin's interpretations of compounds can get rather specific.

Another important difference is that Finin describes relations in both directions. Both Levi and Warren sought to define how the modifier relates to the head. Finin does accept the last noun in the compound as the head noun, and preceding nouns as modifier, but does not use this fact when he defines his relations for compounds.

Both of these properties can be found when looking at his description of *salt water*. Finin assigns the relation "dissolved in" to it. First, this relation is only used to define the connection between a substance and a liquid (which is preferably water). Second, the relation reflects the perspective of the salt, rather than the perspective of the water. The explanation of the term *salt water*, becomes "salt dissolved in water". This may describe a necessary state of affairs for salt water, but when talking about *salt water*, we talk about the water rather than the salt.

Leaving the class of possible rules open may lead to more precise definitions, but in this case, it does not seem to lead to better representations of the meaning of the compound. It is surprising that Finin describes the meaning of *salt water* in such an awkward way, because in his introduction, he states that the problem of interpreting compounds is to detect "the underlying relation that the utterer intends to hold" (Finin (1980), p.1). It seems unlikely that an utterer of the expression *salt water* intends to refer to the dissolution process of the salt.

Finin's structural relations form a more regularly structured and better defined class than the semantic relations described above. He therefore claims that structural relations are the most interesting for linguistic research.

As already mentioned above, structural relations reveal similarities to Levi's nominalisations. However, Finin's structural relations cover many more compound nouns than just nominalisations, and include more relations than regular argument-verb relations.

The main distinction within Finin's structural relations is the distinction between RoleValues and RoleNominals. In a RoleValue-concept compound, the head noun refers to a concept that represents an action. The RoleValue plays a "role" in this action. The typical roles are those related to the subject and object, but RoleValues are not restricted to these arguments. This brings us to the first expansion of the structural class compared to nominalisations. Finin provides examples such as *January flight* (with time), *F4 flight* (with vehicle) and *engine housing* (with super-part). In other words, every modification that could be related to the event expressed by the verbs can be considered a role. In this aspect, Finin's idea of RoleValues shows much similarity to Davidson's event semantics (Davidson (1967)). It includes relations such as TIME, CAUSE and INSTRUMENT, which have been considered "deletable predicates" by Levi.

The term *RoleNominals* also refers to participant roles, but RoleNominals do not play a role in the concept expressed by their fellow compound noun. They play a role to another concept, included in their own meaning. A nominalisation example of a RoleNominal is *driver* referring to the agent of *drive*. If the modifier can be related to the same action in a similar way as RoleValues, then the compound has a "concept-RoleNominal" relation. For instance, in the expression *truck driver*, the *truck* refers to the object in the event.

Finin (1980) classifies *truck driver* as a concept-RoleNominal relation, but note that *truck* can also be seen as a RoleValue according to the definition given above. Concepts preceding RoleNominals can also include Davidsonian arguments such as LOCATION and TIME. The difference between a RoleValue-concept compound and a concept-RoleNominal lies in the fact that in the case of a RoleNominal both the concept and the RoleNominal are roles of the action described. In the case of RoleValue-concept, on the other hand, the concept refers to the action itself, not to one of its roles.

In addition to the extensions regarding acceptable roles, there are two more properties of Finin's structural relations that cause it to cover a larger class than nominalisations. First, just like semantic relations, structural relations can be defined in both directions. Next to RoleValue-concept and concept-RoleNominal sequences, where, in correspondence to nominalisations, the head noun refers to an event, compound in which the modifier refers to the action are considered structural relations as well.

The last extension lies in the fact that in nominalisations, according to the definition given above, the noun had to be derived from the verb. In Finin's classification any noun that can be related to an event can induce a structural relation. For instance, *pilot* relates to *fly* in the same way that *driver* relates to *drive*. The noun *food* can be seen as a RoleNominal of *eat*: *cat food* can be seen as a concept-RoleNominal, where *cat* is interpreted as the subject, and *food* as the object.

Because Finin includes these expressions in his structural relations, many of Levi and Warren's classes can be treated as structural compounds. It is, however, doubtful that this shift really improves the quality of the classification. In a sense, one could say that Finin removed the distinction between grammatical and semantic relations, including most general semantic relations in his structural relation group. His semantic group represents the idea that an infinite number of relations can occur (*salt water* is the only example he provides from this class), and seems to contain only those compounds where the regular relations do not apply.

Isabelle (1984) proposes to expand the interpretation through "argument relations" (i.e., cases where one noun can be seen as an argument of the other) even further than Finin's structural relations.

In addition to Finin's role nominals, Isabelle discusses two classes of nouns, which, when forming the head of a compound noun, could have an argument relation with their modifier. The first group of nouns he discusses are the so-called root nouns. This is a set of nouns that can take an argument of their own. They include among others measure nouns (e.g. *speed (of)*), area nouns (e.g. *side of*), collective nouns (e.g. *top of*) and some additional examples such as *brother (of)* and *location (of)*.

The next extension Isabelle proposes is toward de-adjectival nominalisations. Adjectives can be predicative, and therefore a noun derived from an adjective may have an argument. Grammatical relations within de-adjectival nominalisations can be found in compounds such as *uranium scarcity*.

The main motivation Isabelle gives for the assumption that compounds generally contain grammatical relations, is that grammatical relations are well-defined. He criticises Finin for lack of organisation, motivation and clear definitions of relations in his work. As mentioned above, this criticism seems justified. The extensions described above seem to improve Isabelle's classification on these points, but the improvements are limited to these two groups of compounds.

It is true that the system presented in Isabelle (1984) is well organised, and that described relations are motivated and defined clearly. We must note, though, that Isabelle only discusses compounds that fall under Finin's structural relations. This part of Finin's classification did have clear structure and relations were as well-defined as they are in Isabelle (1984).

Isabelle mentions that he restricts his classification to cover the specific domain of aircraft maintenance manuals, and that he does not aim for full coverage, for this would be too difficult. However, his paper does not reveal what percentage of compounds in these manuals can be explained by his argument relations. In fact, other relations are not addressed at all in this work. It is not clear whether he excluded them because he considers these relations unnecessary, or because they cannot be defined properly, or whether they are out of scope because it is not feasible to cover all relations.[4]

Therefore, Isabelle's classification does not, in our opinion, provide a significant improvement to Finin's classification. Finin's interpretation of *salt water* may be awkward, but Isabelle's system simply does not seem to provide a possible class for such a compound. In his defence, aircraft maintenance manuals will presumably not contain this compound very frequently. However, it is likely that Isabelle's data set will have other compounds with relations that cannot be seen as argument-head relations.

[4] Isabelle does mention genitive compounds, which are out of scope of his study.

Another issue addressed in Isabelle (1984) is the direction in which relations apply. He cites three authors on the matter: Finin (1980), claiming that either modifier or head can be an argument of the other, Selkirk (1982) who claims that the modifier can be an argument of the head, but not vice versa, and McDonald (1982), saying that both directions are possible, but that cases where the modifier is the argument are much more frequent. Isabelle follows McDonald's point of view, which, in fact, had already been expressed by Finin (1980).

In our study, we follow Selkirk (1982), and reject the idea that in a compound, the head noun should be interpreted as an argument of the modifier. We do not contradict the claim that the referent of the head noun may play a role in the event depicted by the modifier, but we believe that compound nouns should be interpreted from the point of view of the head noun. For instance, in the compound *repair man*, *man* refers to the agent of *repair*. However, we use the expression to talk about the man, not about the repairing. A paraphrase of this compound could be "a man who repairs", but not "the reparation by a man". In other words, when interpreting a compound, we address the question of how the head is modified by the preceding noun, rather than the question of how the head may relate to its modifier.

1.5 The nature of semantic relations

The semantic classes of Levi (1978) and Warren (1978) are both the result of an extensive study of data. However, neither of the two authors addresses the question why these classes occur in compounds, and how they came up with the appropriate name for a class. The classes they propose allow classification of all observed data, according to their intuition.

We have already seen that Warren (1978)'s data includes some examples where this intuition is hard to follow. Costello et al. (2006) present an annotation task showing that it is not trivial to classify compounds according to Levi (1978)'s classes either. They start with a classification of 14 semantic relations (Gagné and Shoben (1997)) including all nine deletable predicates.

Their annotators were instructed to assign one of the 14 classes if they could, but invent new relations when they thought this was necessary. The resulting classification included 139 semantic relations, observed in a set of 2,500 noun-noun compounds.

Gagné and Shoben (1997) present a psychological study on compound nouns arguing that a set of 14 relations (corresponding to Levi (1978)'s deletable predicates, where five may apply in both directions) can account for all compounds in their annotated corpus. They refer to these relations as "thematic roles". The idea that thematic roles are logical candidates for semantic relations in compounds can also be found in Rosario and Hearst (2001), who introduce a "linguistically motivated" classification, including —among others— standard case roles, and directive descriptions (Talmy (1985)). This brings us back to the descriptions of role nominals provided by Finin (1980) and Isabelle (1984).

Finin (1980) claims that structural relations form the most interesting class of compounds, because they are relatively easy to define, whereas general semantic classes are problematic because of the intuition involved in their interpretation.[5] We have already mentioned above that Finin has a rather Davidsonian point of view on argument slots of verbs. As a result, his structural rules include several of Levi and Warren's semantic classes. Isabelle (1984) takes Finin's broader view on arguments, including role nominals as arguments. He links his role nominals to Fillmore's case theory (Fillmore (1971)) and *lexical functions* introduced by Zholkovskij and Mel'čuk (1972). In Finin's classification, a compound does not need to contain a term that is directly derived from a verb in order for it to contain a structural relation. If we combine this idea with Isabelle's observation that role nominals are similar to thematic roles, the structural class seems to include most of Levi and Warren's semantic classes. In other words, semantic compound relations that have been proposed reveal strong similarities to suggested thematic roles for verbs. Hence Gagné and Shoben (1997)'s terminology for their compound classes.

[5] Thinking of Finin's intuition on *salt water* mentioned above, we completely agree that this is problematic.

When looking at Levi's deletable predicates (*cause, have, make, use, be, in, for, from* and *about*), strong resemblances to classifications of verb semantics can be observed. The set of necessary thematic roles is as controversial as the set of possible semantic relations in compounds. But, when looking at thematic roles that are found in most approaches (*Agent, Experiencer, Theme, Instrument, Cause, Location, Source, Goal*), the role of *Experiencer* is the only one that is not regularly found in descriptions of compound nouns. On the other hand, when looking at the classification based on Barker and Szpakowicz (1998), presented in table 1.3, the only class that is not regularly discussed in work on thematic roles, is *Container*.

The resemblance between thematic roles does not lead to a clearer insight on the nature of semantic classes. Some other similarities they show in relation to thematic roles is that scientists cannot agree on the number and names for existing classes necessary to classify compounds in a satisfying way. Most scientists only agree on the idea that possible relations form an open set, and that a classification system of the most typical relations can, at best, cover a majority of compounds. For practical reasons, most computational approaches do not make use of a possibly open set of relations, as was done in Costello et al. (2006).

The intuitive character of classes becomes apparent again, when looking at the variety of classifications proposed: most computational linguists introduce their own set of classes. The only exceptions we are aware of is Turney (2006), who uses a set of classes proposed by Nastase and Szpakowicz (2003). Kim and Baldwin (2006) use a classification that is almost identical to the one proposed by Barker and Szpakowicz (1998): only the class PROPERTY is removed. According to Kim and Baldwin (2006), all instances of this class could also be defined by one of the other nineteen classes. In our work, we use this classification system with only one minor change.

Our work focuses on nominalisations, as defined by Levi (1978). The main classes of our system are thus SUBJECT, OBJECT and PREPOSITIONAL OBJECT. In one of our experiments we will combine our nominalisation disambiguation system with a system using Kim and Baldwin's methodology. Therefore, we use the same classification as Kim and Baldwin with PREPO-

SITIONAL OBJECT as an additional class. The classification is presented in table 1.3.

Note that we do not use this system because we believe it to be superior to those proposed by other authors.[6] As we have argued and demonstrated above, the interpretation of compound nouns heavily relies on intuition, and, in this case, there seems to be no reason why we should assume that the intuition of some scientists is superior to others.

1.6 Summary

This chapter presented basic assumptions on compound nouns. Work we presented discussed properties and interpretations between possible relations in compounds. We have seen that attempts to find necessary and sufficient properties to identify compound nouns have failed, except for the definition coming from computational linguistics according to which a compound is a sequence of two or more nouns syntactically functioning as a compound. This is a necessary condition, but we have argued that this condition is not sufficient. We do not consider sequences such as *speaker-listener*, *Amstel Hotel* or *cottontail* (referring to an animal) to be compounds. Therefore, we added the condition that there must be an implicit semantic relation between the components of the compound. Our final definition is repeated below:

I. Definition of Compound Nouns:

> **A sequence of two or more nouns that are connected through an implicit modification relation and together function syntactically as a noun.**

Next to the definition, we have presented four classification proposals that have inspired other authors. Levi (1978) and Warren (1978)'s systems focus on semantic relations between head noun and modifier. Finin (1980) and Isabelle (1984), on the other hand, focus on compounds where the modifier can

[6] In chapter 4 we will point out a few problems with this classification.

be seen as an argument of the head noun or vice versa. These “structural” or “grammatical” relations have the advantage that their definitions are relatively straight-forward. On the other hand, the scope of these relations are rather limited. We have seen that Finin’s descriptions of compounds that are not “structurally” related lacks organisation, and that the only example he presents (*salt water*) has a rather dubious description. Isabelle avoids this controversy, by simply ignoring compounds other than those containing argument-related nouns.

Despite the rather different approach taken by Levi and Warren on the one hand, and by Finin and Isabelle on the other hand, their work may have more in common than one would readily notice. This is mainly due to the resemblance between classes defined in Levi (1978) and Warren (1978) and thematic roles, and the extension that Finin introduces to structural relations, causing them to include most thematic roles. Because neither Finin nor Isabelle present a complete set of definitions for their structural relations, it is difficult to determine the exact overlap.

Therefore, it may not be surprising that most classification systems used in computational linguistics reveal more similarities with the classifications proposed by Levi and Warren than to Finin and Isabelle. Though, to our knowledge, none of the proposed classifications correspond exactly to either Levi’s or Warren’s classification. Levi’s classes are present in most classifications; her work is by far the most cited in papers on compound nouns.

Finally, we introduced a classification system that will be used in our classification. We have chosen this classification system because we use a methodology that is based on that of a classifier using this set of classes.

As has been shown in this chapter, the definition of compounds and of possible relations existing in them is far from trivial, and has led to many discussions in the theoretical field. Despite these difficulties, computational linguists have attempted to find appropriate interpretations automatically. The next chapter will present work on compound nouns in NLP approaches.

Table 1.3: Our Classification System

Relation	**Description**	**Examples**
SUBJECT	C *is performed by* M	*student protest*, *band concert* *military assault*
OBJECT	M *is acted on by* C	*engine repair*, *blood transfusion* *car lover*
PREP OBJECT	M *is a prep. comp of root verb of* C	*side show*, *health education*
BENEFICIARY	M *benefits from* C	*student price*, *charity donation*
CAUSE	M *causes* C	*exam tray*, *overdue fine*
CONTAINER	M *contains* C	*printer tray*, *flood water*, *film music*, *story idea*
CONTENT	M *is contained in* C	*paper tray*, *oil pen*, *eviction notice*
DESTINATION	M *is destination of* C	*game bus*, *exit route*, *entrance stairs*
EQUATIVE	M *is also* Head	*composer arranger*, *player coach*
INSTRUMENT	M *is used in* C	*diesel engine*, *laser printer*, *electron microscope*
LOCATED	M *is located at* C	*building site*, *home town*, *solar system*
LOCATION	M *is location of* C	*lab printer*, *desert storm*
MATERIAL	C *is made of* M	*gingerbread man*, *water vapour*, *carbon deposit*
POSSESSOR	M *has* C	*national debt*, *student loan*, *company car*
PRODUCT	M *is product of* C	*automobile factory*, *light bulb*, *colour printer*
PURPOSE	C *is meant for* M	*concert hall*, *soup pot* *grinding abrasive*
RESULT	M *is result of* C	*storm cloud*,*cold virus*, *death penalty*
SOURCE	M *is source of* C	*foreign capital chest pain*, *north wind*
TIME	M *is time of* C	*winter semester*, *late supper*, *morning class*
TOPIC	C *is concerned with* M	*computer expert*, *horror novel*

Chapter 2

Background Research

Compound nouns have received moderate, but continuous attention in computational linguistics since the early 80s. As a result, a rich literature exists on the topic, starting with early rule-based approaches developing to more recent statistical and corpus based work. This chapter provides an overview of methodologies proposed for compound noun disambiguation.
First, we will present purely hand-crafted rule based approaches, as well as Vanderwende (1994)'s approach extending coverage of rules by using a linguistic resource. The second section of this chapter, present machine learning approaches for the identification of semantic classes in compounds. Research on nominalisations is presented in the last section of this chapter, where we will provide a detailed description of work by Lapata (2002) and Nicholson (2005), which forms the foundation of our own experiments.

2.1 Rule Based Approaches

In the previous chapter, we have seen classification systems for compound nouns proposed by Finin (1980) and Isabelle (1984). Both of these classifications have been designed for automatic disambiguation of compound nouns. Finin (1980) proposes a disambiguation system that works in two steps. First, the nouns occurring in the compound are linked to concepts. Then, a set of hand-crafted rules applies and decides what the right interpretation may be. To take up the example from the first chapter, in the case of *salt water*, *salt* is related to the concept of substance, and *water* to liquid. One of Finin's rules indicates that the relation between a substance and a liquid (preferably water) is "dissolved in".

There are several difficulties for the implementation of Finin's approach. First, Finin insists on respecting Downing's claim that the number of possible relations is an open set. If possible relations form an open set, the rules that lead to the interpretation of these relations are also an open set. In other words, it is not possible to define a set of rules that would allow the interpretation of all compounds.

In addition, Finin's approach was particularly ambitious in the early eighties, because linguistic resources and ontologies were limited. The rules used in the system apply for a pair of concepts. Therefore, the system needs an ontology that can map all nouns that occur in compounds to the right concept, or concepts. Due to the productivity of the phenomenon this means that an ontology containing all existing nouns would be needed.

The last challenge for this approach lies in ambiguity. Even if we would make the simplifying assumption that two compounds containing the same concepts will have the same interpretation, many nouns will relate to different concepts, and for different pairs, different rules can apply. Finin addresses this problem by assigning scores to rules that would find the "best" interpretation when several rules apply. Finding the right score for each rule forms another highly demanding task. In order to get a reliable indication of the most likely rule in different cases, a large corpus study would be necessary.

Isabelle (1984) does not go into the details of the implementation of his system. The only concrete aspect he refers to is a necessary extension of the lexicon, encoding the interpretation of role nominals. For instance, the lexical entry of *pilot* would as follows (Isabelle (1984), p. 513, example (18)):

(3) pilot : x such that FLY(x, y, p) <aircraft(y)>

where x refers to the pilot, and p to the path.

Because of the large amount of hand-coding that this approach requires, Isabelle limits the scope of his research to the specific domain of aircraft maintenance manuals. Note though, that Isabelle's method requires less manual effort than Finin's, because he avoids Finin's problem of needing an endless set of rules. The main effort in Isabelle's approach lies in the hand-coding of the lexicon so that specific nouns can be recognised by the rules.

The same observation can be made for the methodology proposed by Leonard (1984). Leonard defines a fixed set of rules that would cover data included in classification systems such as the ones proposed by Levi (1978) and Warren (1978). It avoids one of Finin's problems by only using a limited set of rules, but, just like in Isabelle (1984), these rules can only apply when the nouns occurring in the compound are known by the system.

Leonard's proposal has the advantage over Isabelle's work that she does not limit her research to a specific domain. Moreover, her treatment of ambiguity is straight-forward: the interpretation rules are ordered, and the first rule that can apply, applies. Leonard's methodology thus has a clear advantage over Finin's approach, because the implementation is simpler.

On the other hand, we have argued above that finding the right ordering of rules is a challenging task. Vanderwende (1994) provides examples of compounds with the head noun *pen* which show that ordering rules may be too simple. The three following rules, forming only a part of the system designed by Leonard, are to be applied in the order presented here:

1 'Search for material modifier'

2 'Search for a related verb'

3 'Search for a locative'

Now, consider the expressions *silver pen*, *ink pen* and *bull pen*. The first two expressions both have a material modifier, and the first rule will apply. However, for *ink pen*, this is clearly the wrong interpretation. Instead, the second rule should have applied. In the case of *bull pen*, which refers to a pen for bulls (thus referring to a different meaning of *pen*), the compound

contains a locative relation. Leonard's classification, however, will apply the second rule, where *bull* will be considered an agent using a pen.
In order to avoid such problems, a more fine-grained taxonomy would be necessary. This would lead to an even greater manual effort, in an approach where the large amount of necessary hand-coding already was problematic.

For all three approaches above, getting good coverage seems to be the main challenge. Despite this problem, they were good efforts considering the possibilities that existed in the early eighties. Moreover, today coverage and ambiguity are still major problems for compound noun disambiguation. Unfortunately, it is not known to what extend results have improved, because none of the early rule-based approaches provides a detailed account of their results. Possibilities have improved these days, thanks to the presence of resources such as on-line dictionaries (e.g. Vanderwende (1994)), FrameNet (e.g. Moldovan et al. (2004)) and WordNet (e.g. Kim and Baldwin (2005)). If these resources were used in Isabelle (1984) or Leonard (1984), they could solve a significant part of their coverage problems. In Finin's approach on the other hand, the main complication lies in his rules. An ontology could improve the coverage of individual rules, but the costs of hand-coding a complex set of rules would remain problematic.

Vanderwende (1994) presents a study proving that linguistic resources can be used successfully to improve coverage of a rule-based system. She is one of the first researchers to explore usage of lexical resources in compound disambiguation.
Vanderwende proposes to use an on-line dictionary (The Longman Dictionary of Contemporary English) as a resource to find relations. She creates rules and uses the dictionary as an ontology. Her goals are to interpret compound nouns in unrestricted text, without taking context into account, which is out of scope of her study.
Vanderwende designed a classification system, containing thirteen classes of possible interpretations for compounds, that should cover most of the possible occurrences of compounds. These classes were based on descriptive work

by Downing (1977), Jespersen (1942), Lees (1960) and Levi (1978).

The system consists of a set of general rules, which can be seen as conditions. Conditions either apply to the head or to the modifier of the compound noun. Each rule is associated with a weight that influences the likelihood of the interpretation.

In order to see whether a condition holds, the system analyses the definitions of the components of the compound. For instance, when looking for the interpretation of *bird sanctuary*, the system finds that a sanctuary is "an area for birds." The condition for the semantic attribute PURPOSE is automatically derived from this definition.

In this case, the relation is rather trivial: one of the nouns of the compound (*birds*) is present in the definition of the other noun. In most cases, mapping between the two nouns is not that straight-forward. The system can also detect relations that are not present in the dictionary in such a direct way. Had the compound been *owl sanctuary* for instance, the system would find that sanctuaries are for birds, and that an owl is a kind of bird. An expression such as *cat scratch* can be classified correctly, because the system will detect that scratching is done with claws and that a cat has claws from the definitions of these words in the dictionary.

Next to these concept relation rules, the system can relate nouns to verbs, and check whether the other noun present in the compound can be one of its arguments. Vanderwende notes that rules for nominalisations are triggered too often, because many nouns are homographic to verbs. Therefore she sets priority on purely semantic rules in her disambiguation algorithm. The algorithm steps are given below:

1 Apply head-based rules to all senses of the head noun and to the lemma of the modifier.

2 Apply modifier-based rules to all senses of the modifying noun and the lemma of head noun.

3 If no interpretation is found with a weight over a threshold θ, then

apply deverbal-head-based rules to all verb senses of the verb and to the lemma of the modifier.

4 Order possible interpretations by comparing the assigned weights, and list them according to likelihood.

In order to link relations to likelihood, Vanderwende assigns rule weights and matching weights. Rule weights reflect how likely it is that a specific rule applies. Match weights reflect how similar the nouns in the compound are to the nouns that triggered the rules. Training data is used to define and weight rules, but it is not clear whether this is done in a systematic matter, or whether she used intuition.
She tested her system both on the training set that she had used to define the rules, and on a test set of randomly selected compound nouns. As expected, the system performed well on her training set: 79% was classified correctly. The test data had 51% classified correctly, and for 21% the best interpretion was ranked second.
These results were based on a relatively small set of data (100 training samples, and 97 test samples). Results on the test data are surprisingly high, considering the small size of the training data, and the complexity of the task.

Despite the promising results achieved by Vanderwende, recent approaches do not make use of hand-crafted rules. Efforts are directed towards using machine learning and using statistical information. For both approaches, data sparseness and the absence of indicative features form the main challenges in the disambiguation task. To sum up the difficulties compounds pose for machine learning: many compounds are singletons, there are hardly any morphological or syntactic clues that may be used for their interpretation, nor do annotated corpora contain information on the relation existing in compounds. In the next section, we present some of the proposal that have been made to overcome these difficulties.

2.2 Machine Learning

Two different strategies to face these challenges can be distinguished. In one approach, researchers retrieve linguistic information from corpora that may provide an indication of the relation between the nouns forming a compound. Researchers using machine learning are mostly troubled by the fact that the overall majority of their test data has not occurred in their training data. Therefore, several methods focus on methods that allow to relate test examples to training examples through semantic similarity. In this section, we will first present two "corpus based approaches", which are directed towards retrieving paraphrases of compounds from corpora. Next, we will present an approach introduced by Rosario and Hearst (2001), where lexical resources are used to relate test data to training data.

The main idea behind corpus based approaches is that if two nouns form a compound, the implicit relation that exists within the compound occurs frequently between these two nouns. This means that, even if the relation is implicit in the compound, there may be other co-occurrences of the two nouns in which the relation is expressed explicitly. This intuition is clearly expressed by Lauer in the following sentence (Lauer (1995), p. 158):

> The key assumption here is that semantically related pairs of concepts are expressed equally often by means of a compound noun and by means of a prepositional phrase.

Lauer (1995) provides the first notable attempt to use corpus statistics for compound noun disambiguation. The quote above reveals two properties of Lauer's approach. First, he uses information from prepositional phrases in order to interpret compound nouns. Second, the probability of an interpretation relates to the frequency **concepts** are expressed through a prepositional phrase, and not the individual nouns themselves.

Prepositional phrases are used in two ways in Lauer's study: first, the goal of his experiments is to find the best prepositional paraphrase for a compound noun. He defines a set of prepositions based on Warren (1978), namely {OF,

FOR, IN, ABOUT, WITH, FROM, ON, AT}. Nominalisations and equative compounds are excluded from his methodology.
The probability of a preposition providing the right interpretation of a compound depends on both the likelihood of the head noun being modified by this preposition, and the likelihood of the modifier being the complement of the preposition. As mentioned above, this likelihood is approximated using frequencies of the head concept and the modifying concept occurring in these positions. Lauer uses Roget's thesaurus to relate words to their concepts.
Lauer trains his system on 23,851 compounds, and achieves an accuracy of 47% on a set of 400 unseen tokens. Considering the complexity of the task, and the fact that his examples were retrieved from a domain-independent setting this is a good result. On the other hand, a relatively large set of training data was used in his experiment (about 20 times the size of sets found in other approaches). Lauer only reports results using this training and test set. It is possible that the system performs comparably when a smaller training set is used.

Another method using corpus statistics to find the most likely paraphrase of compounds has been proposed by Kim and Baldwin (2006). Though instead of evidence from prepositions, Kim and Baldwin use verb semantics to get an indication of the compound's meaning. They use a classification system consisting of 19 possible semantic relations (as described by Barker and Szpakowicz (1998)). 17 of the semantic classes are related to "seed verbs" that can be seen as paraphrases of the semantic classes (for the two classes TIME and EQUATIVE this is not possible). A set of approximately 60 seed verbs is created and enriched with synonyms. Rules are created relating root verbs to semantic classes. For instance, the rule for POSSESSOR is:

$S(\{have, own, possess\}^V, M^{SUBJ}), H^{OBJ}$
$S(\{belong\_to\}^V, H^{SUBJ}, M^{OBJ})$

where V is the set of seed verbs, M the modifier and H the head noun.
Three corpora[1] were parsed and for each verb, its voice, subject, object and PP attachment (including head noun of the PP) were identified. The implicit relation within a compound was identified by the extraction of sentential patterns including both nouns. If the verb within a pattern was not one of the seed verbs or its synonyms, it was mapped to one or more seed verbs according to similarity (based on WordNet and Moby's thesaurus).
In addition, the TIME relation was defined for each compound whose modifier was included in the *TIME* class of CoreLex (Buitelaar (1998)). Coordination structures were used as evidence for inclusion in the EQUATIVE class.
In the classification only verbs present in WordNet were included. If the head and modifier co-occurred less than five times together, they were not included in the classification. Due to this restriction, only 20% of the data set could be used to test the system. On this dataset, an accuracy of 52.6% was achieved.
The results from this experiment are good. It is, however, problematic that only 20% of the data can be classified in this way. The authors experimented using a WordNet based nearest neighbour classification method described in Kim and Baldwin (2005) to classify the rest of the data. The overall performance was slightly inferior to the performance on the smaller data set.
Overall, these are promising results: the effort of hand-coding is kept small by using a limited set of seed verbs. The main problem, as for most approaches, is the data sparseness. Better results may be achieved when instead of searching for individual nouns, the authors would integrate Lauer's idea and retrieve noun classes or concepts to find the most likely interpretation.

Using machine learning in compound noun disambiguation is challenging due to the lack of useful features and the productivity of the phenomenon. Even with a large set of training samples, a test set will contain many unseen tokens. Moreover, obtaining training data is extremely time consuming, because it requires a difficult annotation task. Despite these difficulties, several efforts to use machine learning for compound noun disambiguation have

[1] The BNC, the Brown Corpus and the Wall Street Journal corpus.

been made. Notably, researchers have experimented with methods to relate unseen test samples to training samples that are semantically similarity

Rosario and Hearst (2001) describe a disambiguation method for compound nouns in the biomedical domain. Their system uses a neural network trained on labelled data. The authors experiment with two approaches: a lexical approach and a general approach. In the lexical approach, lexical items are directly used as features to train the system. In the general method, lexical items are mapped to concepts, which serve as features. The Unified Medical Language System (UMLS), a language resource specialised in biomedics, was used to map lexical items to concepts. These concepts can be mapped easily to the MeSH lexical hierarchy (Lowe and Barnett (1994)), equally specialised for the biomedics domain.

Rosario and Hearst train and test their system on compounds retrieved from MedLine, a resource created from biomedical journals. They select 2,245 compounds of which both nouns can be mapped to the MeSH hierarchy. These compounds were annotated, according to a classification system based on general thematic roles, and specific information relevant for information retrieval. The former are partially inspired by Talmy (1985), Levi (1978) and Warren (1978), the latter are the result of their own observations in the domain of biomedics. The resulting classification contains thirty-eight classes of possible interpretations. The classification task is performed with a neural network with eighteen output layers, corresponding to the most frequent classes. The hidden layers were variated up to best performing settings.

Using lexical items directly as features leads to the best results: 62% of the compounds is classified correctly. Though, as noted by the authors, only a fraction of their test data contains nouns that were not present in the training data (75 out of 805). This is a rather unrealistic scenario, and therefore the more general method using concepts provides a better indication of the performance of the system. It achieves an accuracy of 52%, which is still high for an eighteen-fold classification.

Overall, Rosario and Hearst achieve good results, but there are a few problems with their method. The first problem lies in the selection of their data. By only including data from the MeSH hierarchy, their already specialised domain becomes even further restricted. This selection process probably explains the unrealistic situation of their experiment, namely that the test set contains hardly any unseen examples.

Another problem is formed by the thirty-eight observed classes and the eighteen output layers of the neural network. The authors mention in their introduction that only eighteen of the thirty-eight classes occur often enough to be used in their machine learning method. However, this is not addressed in the explanation of their methodology. It is not clear what happened to the examples belonging to the twenty classes that were not taken in consideration. Later in their paper the authors mention that they use 50% for training, and 50% for testing, namely 855 training points, and 805 test points. Ignoring the fact that 855-805 is not exactly a 50-50 split, these numbers reveal another interesting point: the total set of data seems to contain 1,660 compounds instead of 2,245. The 555 compounds that have been removed from the corpus may be compounds belonging to the other 20 classes, but one can only guess.

In the end, it is difficult to judge the system developed by Rosario and Hearst. On the one hand, results are impressive. On the other hand, it is not clear whether these results are related to their disambiguation method, or their data-selection method. At first sight, the latter seems more likely.

Rosario and Hearst's idea to use concepts from an ontology to approximate unseen instances has been followed by other researchers. Both in Moldovan et al. (2004) and in Kim and Baldwin (2005), WordNet is used to define similarity between test and training samples. Moldovan et al. do not present an evaluation of their automatic labelling task. The research presented by Kim and Baldwin on the other hand provides a good indication of the potential of this approach. They test different similarity measures calculated using WordNet :: Similarity (S. Patwardhan (2003)), an open source software package which implements different similarity measures. Kim and Baldwin use a training corpus of 1,088 tokens annotated according to a 20-fold classification defined by Barker and Szpakowicz (1998). Their best results on a test set of 1,081 tokens is 53.3%.

2.3 Classification according to Semantic Classes

The previous sections presented rule based approaches and machine learning approaches to compound noun disambiguation. Not all researchers provide a detailed evaluation of their system. The approaches that are evaluated all reveal similar result, achieving accuracies of approximately 50%.

Vanderwende (1994) already managed to build a system which classified 51% correctly, using her rule base approach consisting of a limited set of rules. Twelve years later, the state of the art in compound noun disambiguation still seems to be that about half of the compounds can be classified correctly. However, the similar results do not mean that the methodologies presented above reveal equal performance. They are all designed to classify compound nouns according to their semantic relation, but this is about all they have in common. For several reasons, different approaches to compound noun disambiguation are difficult to compare.

First, researchers tend to use different sets of classes. The complexity of a classification task directly depends on the number of classes considered. Second, in several approaches only a limited set of data is considered. Rosario and Hearst (2001) only use data from the biomedical domain, Kim and Baldwin (2006) reduce their data so that it only includes compounds whose components co-occur in their corpus, Lauer (1995) excludes nominalisations and equative relations. These differences between datasets makes direct comparison of results an unreliable indication for the performance of the systems. An additional hindrance for comparing are differences in the size of training sets: Lauer (1995) uses over 20,000 compounds to train his data, Vanderwende (1994) only uses 100 examples to define her rules.

In short, from the observations presented above we can derive to observation on compound noun disambiguation. First, the controversy on the number and nature of semantic classes makes it impossible to compare the results from different approaches. Second, compound noun disambiguation remains a difficult task, where no approach seems to do much better than classifying half of the data correctly.[2]

Both of these observations have inspired several authors to limit the scope of their research to nominalisation, i.e. compounds containing a grammatical relation. This limitation has the advantage that it reduces the complexity of the task, and that only a small set of well-defined classes is considered. As mentioned in our introduction, we also limit the scope of our research to nominalisations.

The next section presents three studies, which use a statistical approach for the disambiguation of nominalisations, namely Lapata (2002), Grover et al. (2003) and Nicholson (2005). The methodology proposed by Lapata and extended by Nicholson forms the basis of our research, and will therefore be described in detail.

2.4 Nominalisations

Lapata (2002) introduces a statistical approach for the disambiguation of nominalisations containing a subject or object relation. Her methodology can also be found in Grover et al. (2003) and Nicholson (2005). We will present these three works, focusing on the research carried out be Lapata and Nicholson.

There are two main obstacles in using a statistical method for nominalisation disambiguation. The first obstacle is that —due to the productivity of the phenomenon— most compound nouns will not occur more than once, even in a large corpus. The second, even more problematic, obstacle is that information about the relation between two nouns is not present in corpora.

Lapata solves this problem with the assumption that the probability of a relation *rel* between two nouns (n_1 and n_2) can be approximated by linking the head noun n_2 of the nominalisation to the verb from which it was derived (henceforth "root verb" or v_{n_2}). If the modifying noun occurs more often as the subject of the verb than as its object, the compound noun would be

[2] The only exception we have found are the results from Rosario and Hearst (2001), who achieve an accuracy of 63%. However, the authors note themselves that due to the selection process of the data, their experiment had an unrealistic setting.

more likely to contain a subject-relation. On the other hand, if the modifying noun is more likely to be an object of the root verb, this would point to an object relation between the two nouns. The probability of a specific relation *rel* is thus approximated using the following formula:

$$P(rel|n_1, n_2) \approx \frac{f(v_{n_2}, rel, n_1)}{\sum_i f(v_{n_2}, rel_i, n_1)} \tag{2.1}$$

where $P(rel|n_1,\ n_2)$ stands for the probability of a certain relation *rel* existing between the modifier noun n_1 and the head noun n_2, $f(v_{n_2},\ rel,\ n_1)$ stands for the frequency of relation *rel* occurring between the verb v_{n_2} and n_1. This frequency is divided by the total number of times n_1 occurs with the verb v_{n_2}, in any relation.[3]

Lapata detects nominalisations by selecting noun-noun sequences (using an algorithm proposed by Lauer (1995)) of which the head noun was listed in NOMLEX (Macleod et al. (1998)) or CELEX Baayen et al. (1995). NOMLEX and CELEX are manually created resources containing deverbal nouns and their root verbs.[4]

Using this algorithm, Lapata retrieves 1,027 nominalisation compounds from the BNC. After manual inspection, where compounds that are interpreted either as one of Levi's deletable predicates or as a prepositional complement are removed, 796 compounds containing either a subject relation or an object relation remain.

Even with the approximation, data sparseness remains one of the main challenges in this experiment. A large effort in Lapata's work therefore looks into performance using different smoothing techniques. She compares a general discount method (Katz (1987)) to methods that approximate probabilities by relating the modifying noun to similar nouns. She investigates two class-based techniques, where the linguistic resources WordNet and Roget's thesaurus are used to determine similarity between nouns. In addition, Lapata experiments with distance-based similarity. She uses two calculations based

[3] In Lapata (2002), this is either "subject" or "object", since other grammatical relations are out of scope in her study.

[4] For a detailed description on NOMLEX and CELEX, see section 3.2.

on distribution similarity of nouns in corpora. Because results with different smoothing techniques were comparable, we will not provide a detailed comparison here.

Next to different smoothing techniques, Lapata investigates whether morphological affixes can be used for disambiguation. If the nominalisation has one of the agentive suffixes *-er*, *-or* or *-ant* an object relation was assigned, the passive suffix *-ee* would indicate a subject relation. Using these suffixes led to a slight improvement, but this was not significant.

Finally, Lapata examines whether direct context provides a useful indication for the relation present in the nominalisation compound. She runs experiments where lemma's and part of speech tags of words directly preceding or following the compound are used as features for classification. Because we use Lapata's methodology in our research, we provide a detailed description of her algorithm.

In the first stage, the head noun is mapped to its root verb using Nomlex and Celex. Next, frequencies of the root verb and modifier are retrieved from the BNC, parsed by Cass (Abney (1996)). Lapata uses a likelihood ratio (Mosteller and Wallace (1964), Hindle and Rooth (1993)) for the comparison between probabilities. She calls this ratio "RA-score", after "argument relation" (Lapata (2002), p.4). The RA is calculated using the following formula:

$$RA(rel, n_1, n_2) = \log_2 \frac{P(\textsc{obj}|n_1, n_2)}{P(\textsc{subj}|n_1, n_2)} \tag{2.2}$$

A positive RA-score points to an object-relation, whereas a negative RA-score, rather indicates a subject relation.

Note that the RA-score cannot be determined if both probabilities are zero. If $P(\textsc{obj}|n_1, n_2)$ or $P(\textsc{subj}|n_1, n_2)$ is zero, the value will reach positive infinity or negative infinity, respectively. Lapata avoids these values by approximating the frequency of each relation that does not occur in the corpus, using the smoothing techniques mentioned above. As already mentioned, we will not provide a detailed comparison of different smoothing techniques. We will restrict ourselves to a description of the WordNet smoothing technique, which led to overall best results, and was therefore used in our research as well.

The WordNet smoothing technique is based on the idea that semantically similar nouns will have similar probabilities to occur in a specific grammatical relation with the root verb. Lapata retrieves all semantic classes to which the modifier can belong from the WordNet. Arguments of the root verb are linked to these classes, and for each class frequencies of subject and object relations are obtained.

The difficulty of this approach is that most words have more than one sense, and that each sense can relate to several conceptual classes. Because the corpus does not contain any information on the class of the noun in a specific context, Lapata makes the simplifying assumption that all classes are equally probable and divides the frequency for each approximating noun by the number of classes it belongs to.

The sum of these relative frequencies is divided by the number of conceptual classes that are hypernyms of the modifier. These frequencies are used to approximate probabilities of the relation between the nouns of the compound.

The approach presented by Lapata performs significantly above the baseline with all techniques, except when only back-off smoothing is used. Best results are achieved using suffixes, two lemma's of preceding and following context, and WordNet smoothing. Combining these methods, an accuracy of 86.1% is reached. Compared to a baseline of 61.5%, and an "upper baseline" of 89.7% correctly labelled by human annotators, this is a very good result.

Lapata's research shows that frequencies of grammatical relations can be used successfully for the disambiguation of nominalisations. On the other hand, there are the approach presented in Lapata (2002) leaves room for improvement.

First, as explained above, Lapata applies her smoothing techniques for each relation that does not occur in the corpus. If, for instance, a given modifier occurs four times as the subject of a verb, and never as its object, the frequency of the object relation will be approximated. The value of the subject frequency will remain four. Intuitively, the fact that a modifier occurs a number of times in relation A, and never in relation B, could provide evidence

for relation A being the right interpretation. Instead of using this intuition, Lapata smooths over relation B, and compares two scores which represent different values.
The second problem with Lapata's approach is that it limits itself to subject and object relations. 9.2% of the potential nominalisation compounds she retrieved from the BNC was excluded from her data. This selection could only be done manually. This brings us to the most important short-coming in her experiment: it performs extremely well on manually selected data. It is not clear from Lapata's experiments how her system would perform on open text. In total, 37.2% of the automatically retrieved compounds had to be removed from the dataset, because they did not contain a subject or object relation. This indicates that precision in nominalisation detection lies around 60%, but there is no indication on recall. Performance on nominalisation identification, and the disambiguation of prepositional complements are both addressed in Grover et al. (2003) and Nicholson (2005).

Grover et al. (2003) investigate how techniques developed for general text can be used in the Biomedical Domain. For this purpose, they run experiments on nominalisation disambiguation. They use the same basic idea as Lapata (2002) in their approach. Their main improvement compared to Lapata is that the expand her two-fold classification to a system including thirteen classes. Next to the subject and object relation, their system identifies the prepositional relations WITH, TO, ON, FOR, IN, FROM, ABOUT, AGAINST, BY, INTO, OF.
Grover et al. retrieved sequences of two nouns where the head noun was either listed in NOMLEX or which head a verbal stem with suffix *-ing*. From these compounds, a sample of 1,000 tokens was selected at random. 79.7% of the retrieved tokens contained a grammatical relation.
The choice of the best relation was carried out in a similar way as in Lapata (2002): the head noun was related to its root verb, and the most frequent relation between the modifier and head noun was assigned. In order to cope with data sparseness problems, the followed Rosario and Hearst (2001) and

used the UMLS meta thesaurus and WordNet. The system is trained using C4.5 (Quinlan (1993)), a decision tree learner.

They tested their system on both the full set of 1,000 compounds and on the subset containing only nominalisations. When combining context and grammatical relations, they achieve an accuracy of 66.1% on the full set, and 73.6% on the set of nominalisations.

In addition, they run a small experiment on the detection of compounds. For this purpose, the authors develop an *fsgmatch* grammar, that identifies sequences of nouns, and sequences of adjectives and nouns. This grammar achieves an accuracy of 95.6% in identifying compounds.

Grover et al. (2003) improve on Lapata (2002)'s approach by including a set of prepositions and experimenting with performance on a dataset containing compounds that are not nominalisations. On the other hand, they only cover data from a restricted domain. Moreover, their approach required some linguistic pre-processing of the data in order to facilitate the retrieval of compound nouns, and fine tuning of their *fsgmatch* grammar on a large set of data (namely the first two years of the corpus). Because of the large manual effort that needed to be performed on this specific dataset, the methodology presented in Grover et al. (2003) is not suited for applications on unrestricted text.

Nicholson (2005) criticises the fact that neither Lapata's study, nor the work carried out by Grover et al. investigates the disambiguation of nominalisations on open text. His research forms an extension to Lapata's work, and improves on her system in several ways.

First, Nicholson runs experiments on the identification of nominalisations. For this purpose, he uses a chunker, a part-of-speech tagger and a database consisting of deverbal nouns. When using a database only containing entries from NOMLEX and CELEX, corresponding to the resources used in Lapata (2002), accuracy on nominalisation detection was 41.1%. The first improvement Nicholson introduces is adding deverbal nouns from CATVAR to the database.[5] With the additional data from this partially automatically cre-

[5] For a detailed description of CATVAR see section 3.2.

ated resource, the accuracy goes up to 69.8%. Note though, that CATVAR introduces noise, and precision on nominalisation detection drops from 72.3% to 57.1%. Whether the gain in recall or loss in precision is more important depends on the task.

The next improvement compared to Lapata is that Nicholson's disambiguation system takes prepositional complements into account. His approach differs from Grover et al. (2003), because he does not use a predefined set of acceptable prepositions. Nicholson shows that Lapata's methodology can be extended to include prepositional complements on open domain. He achieves an accuracy of 57.4%, which lies significantly above the majority class baseline (48.8%).

In addition to the extensions mentioned above, Nicholson integrates a few improvements in the general methodology found in Lapata (2002), the most significant of which is presented here. Nicholson uses z-scores for comparison instead of Lapata's RA-scores. The z-score of an observation represents its standardised difference from the mean. This way of determining the most likely relation is proposed to be more robust in case of sparse data (Nicholson and Baldwin (2006)).

Z-scores are calculated as follows. When comparing two relations, we start with the assumption that both relations are equally probable. Now, let $|A| = f(v_{n_2}, rel_A, n_1)$ be the frequency of v_{n_2} and n_1 occurring in relation A, and $|B| = f(v_{n_2}, rel_A, n_1)$ be the frequency relation B occurs between v_{n_2} and n_1.

The simple mean μ and the standard deviation σ are defined as follows:

$$\mu = \frac{1}{2} \cdot (|A| + |B|) \tag{2.3}$$

$$\sigma = \frac{1}{2}\sqrt{(|A| + |B|)} \tag{2.4}$$

The z-scores for relation A and B respectively are expressed in the following formulae:

$$Z_A = \frac{|A| - \mu}{\sigma} \tag{2.5}$$

$$Z_B = \frac{|B| - \mu}{\sigma} \tag{2.6}$$

If the verb can occur with a preposition, Nicholson (2005) also determines z-scores of the prepositional complement in comparison with the subject. From these four z-scores, the highest score will determine the relationship that will be assigned.

Nicholson repeats Lapata's experiment using only back-off smoothing and achieves an accuracy of 70.1%. This result is significantly better than Lapata's, when using the same technique (64.7%). However, this difference does not lie in the more robust character of z-scores. In a two-fold disambiguation where RA-scores or z-scores are compared directly, both methods will lead to the exact same results: the most frequent relation will be assigned.

There are two possible explanations why Nicholson performs better than Lapata. The first and main reason lies in the application of smoothing techniques. We have argued above, that Lapata's approach might be flawed because, when one of the relations does not occur in the corpus, this relation is approximated. Nicholson, on the other hand, only performs smoothing when z-scores are equal. In practise, this boils down to cases where none of the relations is attested. This difference is probably the main reason for his superior performance: it avoids Lapata's problem of unequal comparison.

The other difference lies in the inclusion of statistics from prepositional objects, even in the two-fold disambiguation. In all disambiguation tasks, Nicholson (2005) obtains z-scores for prepositional objects in relation to subjects, next to the object-subject z-scores, but in his two-fold classification an object relation is assigned when the prepositional object has the highest z-score. As a result, the system is biased to assigning subject relations. When, for instance, we count 30 subjects, 35 objects and 5 prepositional objects, the highest z-score would be that of the subject compared to the prepositional object. In a direct comparison between subject and object, the object relation would have been assigned. Recall on subjects is generally low in Lapata's work, and therefore this property may have had an impact on Nicholson's results.

This second difference is hardly more related to the use of z-scores than the first is: comparing RA-scores of subject-object and subject-prepositional probabilities would lead to the same bias towards subject relations. In the end, RA-scores and z-scores serve the same purpose, only the methods differ. At most, there may be some borderline cases, where one score will assign subject, while the other assigns object. It is unlikely that the choice between RA-score and z-scores will have any impact on the performance of the system when using all three relations, and even impossible that it will when only two relations are concerned.

Overall, Nicholson's work presents the state of the art in the statistical disambiguation of nominalisations. First, it investigates how well nominalisations can be detected from open text when using a database of deverbal nouns to decide whether a compound is a nominalisation. Second, it shows performance of the purely statistical method including prepositional complements on compounds retrieved from an unrestricted domain.

The three approaches presented above reveal that statistics on verb-argument relations can be used successfully to disambiguate compounds that are known to be nominalisations. In addition, Nicholson presents results on the detection of nominalisations. However, he does not investigate how nominalisations can be detected and identified. Just like Lapata and Grover et al., he assumes that a compound with deverbal head will contain a grammatical relation. All datasets mentioned above show that this is not the case. As we have seen, Lapata removes 28.0% of the data because the relation is not grammatical, Grover et al. report that 20.2% of their compounds with deverbal head were not nominalisations, in Nicholson (2005), precision on nominalisation identification was 72.3% and 57.1%, respectively.

None of the approaches above address the question of how nominalisations can be distinguished from other compounds with a deverbal head. In order to have a system that gets good recall and precision on open text, it would be necessary to incorporate a component that checks whether the deverbal compound contains indeed a grammatical relation. Such a component is missing from all three methods presented above, even though it is essential for a system meant for use on unrestricted data.

Chapter 3

A Hybrid Disambiguation System

Nominalisation Detection and Identification

In the previous chapter, we presented three statistical approaches for the disambiguation of nominalisations. All three approaches revealed promising results. However, none of the approaches addresses the question of how nominalisations may be identified. Nicholson does investigate precision and recall on open text, but does not extend the system so that it can distinguish nominalisations from other compounds with a deverbal head.

In our experiments, we focus on the automatic detection of nominalisations. More particularly, we focus on increasing the precision of nominalisation identification. Improving recall is not addressed in detail, because —as we will see below— we achieve high recall (94.5% and 99.5%, respectively) by introducing some minor changes in Nicholson's approach.

Three hybrid methods

We implement three methods that try and detect whether a relation in a compound is a grammatical relation. The first two methods test whether the modifier is likely to be an argument of the verb. In the first experiment, presented in section 3.4, we test whether the modifier occurs as an argument of the verb in the BNC. If neither the modifier, nor a noun with similar meaning occurs as an argument of the verb, we consider that the compound is not likely to contain a grammatical relation. The second experiment tests whether the modifier can be an argument of the verb according to the verbs selectional restrictions. This experiment will be described in section 3.5.

In the third experiment (see section 3.6), we disambiguate compounds using both our nominalisation disambiguator, and a second system that takes semantic classes into consideration. When the second system shows a high probability for a semantic relation, we will consider that the compound does not exhibit a grammatical relation.

All three methods make use of a basic system that uses the methodology found in Lapata (2002) and Nicholson (2005). This basic system forms the statistical component of our disambiguation system. In the first two experiments, we combine this statistical method with linguistic resources such as WordNet (Fellbaum (1998)) and VerbNet (Kipper et al. (2000)). In the third experiment we combine two systems. We thus experiment with three hybrid methods for nominalisation detection.

Before we present these three experiments in detail, we will describe the basic system we implemented.

3.1 Our Basic System

We implement our basic system according to the methodology found in Lapata (2002) and Nicholson (2005). We use a database created from Celex, Nomlex and Catvar to relate deverbal nouns to their root verbs. The creation of this database is explained in section 3.2. Note that, despite the fact that our experiments are directed towards improvement of precision, we will include the data from Catvar. Recall in Nicholson's experiment went up to 49.6% (compared to 26.4%) when this resource was included. We assume that our detection methods can filter out Catvar's non-deverbal infiltrates. The probability of grammatical relations within the nominalisation are approximated by looking at the probability of the modifier standing in a specific relation with the head noun's root verb. Frequencies on grammatical relations for each modifier root-verb pair are obtained by extracting verb-object and verb-subject tuples from the BNC, parsed by RASP (Briscoe et al. (2006)). Because our system will also cover prepositional relations, we also extract tuples of prepositional relations, including the preposition.

In Lapata, frequencies are compared using RA-scores, Nicholson uses z-scores. We have argued in the previous chapter that this choice should not have a significant impact on the results of the disambiguator. In our experiments, we only use z-scores. We determine z-scores for subject in relation to object, object in relation to subject, and prepositional complements in relation to subjects. The relation with the highest z-score is assigned.

In case of equal z-scores, we use WordNet for smoothing. Our smoothing technique slightly differs from the method used by Lapata. Recall from the previous chapter, that Lapata uses WordNet to link both the modifier and nouns occurring as the verb's argument to concepts, and that these concepts are used to approximate the probability of the relation.

The problem of this approach is that, in the end, all nouns relate to the class <entity>. Therefore, all nouns that occur as an argument of the root verb will contribute to the approximation, even if their semantics is not related to the semantics of the modifier in the slightest way. The question of whether more closely related nouns should have more impact on the approximation is not addressed in Lapata's work.

Nouns that are more closely related to the modifier will have a bigger impact on the approximation in Lapata's approach, because they will have more classes in common with the modifier. Their occurrences will thus be taken into account for more than one class in the approximation process. This aspect of Lapata's approach takes relatedness slightly into account. Nevertheless, we decided to use a different approach using the idea that more closely related nouns provide a better indication in a more direct way.

In the smoothing process, we retrieve all occurrences of the verb from the BNC. For all nouns occurring with the verb, we determine the probability of it being the subject or the object of the verb. We then use WordNet to calculate the distance between the two nouns.

We define the distance between a noun n_1 and a noun n_2 by the number of steps that are needed to get from n_1 to n_2 over hypernym chains. We detect common hypernyms of the two nouns, and detect the hypernym H_c that links the two nouns with the least steps possible. The distance between the two nouns is the sum of the distance (*dist*) from n_1 to H_c and the distance from n_2 to H_c.

$$dist(n_1, n_2) = dist(n_1, H_c) + dist(n_2, H_c) \tag{3.1}$$

Words that are placed lower in the hypernym hierarchy have a more specific meaning. Wu and Palmer (1994) introduce a measure that takes this intuitive idea into account, assuming that words having a common hypernym placed lower in the hierarchy are more similar than those having a common hypernym relatively close to the root concept.

We integrate this idea by dividing the distance between n_1 and n_2 by the distance of the two nouns to the root concept R_c:

$$d(n_1, n_2) = \frac{dist(n_1, H_c) + dist(n_2, H_c)}{dist(n_1, R_c) + dist(n_2, R_c)} \tag{3.2}$$

The resulting value d represents the dissimilarity between two nouns. Because we divide the distance between the two nouns by their distance from the root, and the latter is greater than or equal to the former, all our values are between zero and one. In order to give greater importance to the statistics coming from nouns that are more similar to the modifier, we multiply the probabilities for each approximating noun with (1-d). Note that the dissimilarity value will be exactly one when the two nouns are only related by the root concept <entity>. Their probabilities will thus not be taken into account for the approximation.

The overall approximated probability corresponds to the sum of individual probabilities, divided by the number of nouns this was based on. The calculation of the overall approximation is expressed in the formula below:

$$P(rel|n_1, n_2) \approx \frac{1}{N} \cdot \sum_{i=1}^{i=N} \left(\frac{f(v_{n2}, rel, n_i)}{\sum_j f(v_{n2}, rel_j, n_i)} \cdot (1 - d(n_1, n_i)) \right) \qquad (3.3)$$

where N refers to the number of nouns occurring as an argument of the root verb, $f(v_{n2}, rel, n_i)$ represents the frequency of a noun n_i standing in a relation rel to the root verb, divided by the number of times n_i occurs as an argument of the verb. This way, we obtain the probability $P(rel|n_i, v_{n2})$, which is multiplied by $1 - d(n_1, n_i)$, where $d(n_1, n_i)$ represents the dissimilarity between n_i and the modifier according to WordNet.

3.2 Preparatory Work

The previous section described the basic system we used in our experiments. In our description, we have mentioned several databases that form a central part of our system: Nomlex, Celex and Catvar, used to detect nominalisation candidates and the BNC parsed by RASP, used to retrieve frequencies of verbal relations.

This section describes which resources we use, and how we created databases containing relevant information for our system.

3.2.1 Nouns and their root-verbs

We retrieve our noun-verb pairs from Nomlex, Celex and Catvar. A brief description of each of these resources is given below.

Nomlex (Macleod et al. (1998)) is a manually created nominalisation dictionary. It provides the root verb, as well as possible nominal complements for 1,025 nominalisations. We retrieve noun-verb pairs from Nomlex, as well as the “nominalisation type”, which indicates whether the nominalisation refers to the subject or object of the verb, or to neither specifically. We do not use the sub-categorisation information from the database, because it is mostly directed toward prepositional complements of the noun.

Celex (Baayen et al. (1995)) is a manually created lexicon for Dutch, English and German. It contains information on syntax, phonology, orthography, morphology and frequency of English words and lemmas. We use the database containing morphological information to retrieve pairs of related nouns and verbs.

Documentation on Celex (Burnage (1990)) provides a decent motivation for the creation of the database, and for the information that was included, but it does not provide detailed information on the structure of the corpus. In a first stage, we retrieved noun-verb pairs automatically from Celex. However, it turned out that this automatically created database of noun-verb pairs contained errors. Therefore, we created our own database, this time retrieving noun-verb pairs in different stages according to their morphology. Because the derivation process is reflected in the representation of Celex entries, we managed to retrieve more data with this method. The database we created from Celex contains 6,093 nouns, compared to 2,130 nouns in Nicholson's set of automatically retrieved candidates.

Catvar is a "Categorial-Variation Database" (Habash and Dorr (2003)). The database presenting clusters of words from different categories related to the same lemma. We used the second release, containing 63,146 clusters made of 109,807 words. Catvar is made automatically, partially using algorithms, partially taking up information from resources such as WordNet and Nomlex.

Nicholson (2005) mentions that the database contains errors due to the automatic creation, but he does not attempt to solve this problem. We retrieved all noun-verb clusters from Catvar automatically, and went through this database manually to remove errors. In the filtering process, we removed pairs that were clearly not related, thus verb-noun pairs such as *dent-dentist* and *do-door/don*, and cases where the verb was metaphorically related to the noun such as *dog-dog*.[1]

In most cases where British and American spelling differs, each spelling has

[1] The verb *to dog* means "to follow like a dog".

its own entry in Catvar. While going through the corpus, we merged entries of words only differing in spelling, such as *honour* and *honor*. This way, data with British spelling can be used to interpret compounds with American spelling and vice versa. Since data sparseness is a major hindrance in our method, merging British and American spelling could have a significant impact on the performance of our system.

We combined noun-verb pairs from Nomlex, Celex and Catvar into one database. Following Nicholson (2005), our database allows for one nominalisation to be mapped to several verbs. In the disambiguation of the compound *city destruction* for instance, this allows us to use both frequencies coming from *destroy* and *destruct*.

3.2.2 Frequencies of grammatical relations

As mentioned above, we get our frequencies from the British National Corpus (Burnard (2000)). The BNC is a part-of-speech annotated corpus, made out of over 4,000 different text samples, both written and spoken (Burnard (2000)). It contains more than 100 million tokens.

Because we needed to use grammatical relations, we used a version of the BNC that was parsed by RASP (the "robust accurate statistical parser"). The parser is a "manually-developed wide-coverage tag sequence grammar" (Briscoe et al. (2006)). It aims to find a full parse of a given sentence using partial parses. Partial parsers are associated with probabilities, which allow ranking full parses according to likelihood. RASP contains a tokeniser, part-of-speech tagger, and a lemmatiser. All these modules can take either raw data or annotated text as input.

In the output of RASP, grammatical relations of the sentence are represented in a binary way: each output line contains two tokens and the relation connecting them. Object relations can be retrieved directly. Prepositions are retrieved in two steps: first the relation between verb and preposition is identified, then the relation between preposition and noun. In the case of a subject relation, we need to take the voice of the verb into account. If the verb is passive, the instance counts as an object relation in our database. The RASP output for a passive and an active sentence can be found in the appendix.

We used the RASP parsed version of the BNC to create two databases. The first database contains tuples of nouns, verbs, and frequencies of each grammatical relation (subject, object or prepositional, introducing a new entry for each preposition). The second database, which is used to map noun pairs to seed verbs, contains transitive verbs together with their subject and (prepositional) object.

3.3 Three Identification Methods

In order to identify nominalisation candidates, we propose three methods. When selecting nominalisation candidates, we check whether the head noun of the compound is present in our database of deverbal nouns. Using a database created from Nomlex, Celex and Catvar, where data from Celex is retrieved in several steps as explained above, we achieve recall of 99.5% and 95.5% on an open set. Therefore our experiments concentrate on increasing precision. We explore two main ideas for selecting those compounds that are indeed nominalisations.

The first idea is that the modifying noun must be a good candidate for being an argument of the verb. Or, in other words, when relating the head noun to its root verb, is the modifier likely to be one of its arguments? This method should be able to exclude possible candidates such as *swimming pool*, where the modifier is highly unlikely to occur as an argument of the head noun's root verb.

The second idea is to test whether the compound is a good candidate for another semantic relation. In this approach we do not try and exclude bad candidates, but we see whether the compound may be a better candidate for a semantic relation than for a grammatical relation.

We use two different approaches to see whether the modifier would be a good argument for the root verb. We have one approach to test whether the compound is more likely to constitute another semantic relation. The three identification methods we use are described in the following sections.

3.4 Experiment 1

The first way to check whether the modifier makes a good argument for the verb is based on a very simple idea: "if it does not occur, it is not good". We search for occurrences of the root verb and the modifier noun, and assume that a noun that would be a likely argument of the verb will occur with it. However, as has been mentioned before, data-sparseness is a major hindrance in this statistical approach. If we only consider modifier-verb pairs that occur in the BNC, recall is expected to be extremely low.

Therefore we use WordNet to include synonyms and closely related nouns occurring with the verb as evidence for the nominalisation to be a good candidate. Nevertheless, due to the magnitude of the data sparseness problem, this experiment will —most likely— not lead to impressive results, and can therefore serve as a baseline for the other two methods. The system works as follows:

We implement a system as described in section 3.1 above. The system takes a sequence of exactly two nouns as an input, and checks whether the head noun can be related to a verb, according to our nominalisation-verb database. If the modifier and verb occur in our database with non-zero frequencies from the BNC, we will consider the compound to be a good candidate for a nominalisation.

If the modifier does not occur with the verb, we consult WordNet to get synonyms of both the modifier. We then check our database for occurrences of the modifier's synonyms with the verb. If these turn out to be present in the database, the compound will be accepted as a nominalisation. The frequencies from the synonyms will be used to approximate the probabilities of each relation.

In order to increase recall, we also run experiments where hyponyms, or

nouns within a certain WordNet distance from the modifier are used as positive evidence for the compound containing a grammatical relation.
The smoothing method described in section 3.1 is not used in this experiment, because it generally finds occurrences allowing for approximation for all noun-verb pairs. However, in case words within a certain distance of the modifier are included, we do use the factor of (1 - d) described above.

3.5 Experiment 2

The second experiment implements another method that tests whether the modifier is a good candidate for standing in a grammatical relation to the verb. Namely, we test whether the modifier is a possible argument according to the verb's selectional restrictions. If the modifier differs significantly from the verb's semantic roles, we classify the compound as "not a nominalisation".
Again, our basic system uses the methods described in Lapata (2002) and Nicholson (2005). We only check selectional restrictions of verbs that are not found with the modifier in our frequency database, assuming that if a noun occurs as an argument of the verb, the selectional restrictions of the latter allow for this.
We tested three resources to retrieve information on the semantic roles of the verbs: PropBank (Kingsbury and Palmer (2002)), FrameNet (Baker et al. (1998)) and VerbNet (Kipper et al. (2000)). For all of these resources, coverage can form a problem. PropBank contains 3,323 verbs, FrameNet has a little over 10,000 lexical entries, and VerbNet contains 3,604 verbs. It is therefore likely that some of the nominalisations will have root verbs that are not represented in these resources. However, a test run revealed that over 96% of the verbs in Lapata's data were included in PropBank.
PropBank ("propositional bank") is a corpus of annotated verbal propositions and their arguments, created from the Penn TreeBank. Because this resource is corpus-based, it was our first choice as a resource for retrieving information on the verb's semantics. In the original design of the experi-

ment, we compared the semantics of the modifying noun to the semantic roles annotated in PropBank, using WordNet. Though, several difficulties arose when linking PropBank annotations to WordNet, and comparing them with the modifiers.

First, a part of the PropBank entries are deverbal nouns, which only seem to exist as these particular semantic roles. Nouns such as *agreer* and *praiser* are not present in WordNet. Second, many entries consist of more than one word, for instance *cause of rightness* (first role of *to square*) and *item being labelled* (second role of *to portray*).

Many object roles start with the words *thing* or *entity*. For instance, *observe* has *thing looked at* as its second role. Therefore we tried to use the first word of each entry only, taking the risk that these more general terms would lead to a decrease in precision. This lead to the discovery of a third problem of PropBank: the entries in this resource are rather inconsistent. For instance, the entry *acquire* entry has *thing being acquired* as its object role, but the verb *barb* has an object defined as *barbed thing*. If the first word of the description is the participle of the verb, this will not be found in WordNet.

Unfortunately the inconsistency does not end here: as mentioned above, some entries, such as *agree*, have rather specific roles (in this case *agreer* and *other entity agreeing*), whether others, such as *advance*, have more general roles (in this case *agent*).

Even if we would manage to link all role descriptions of PropBank to WordNet entries, we still could not find a good way to test whether the modifier is an appropriate candidate: if the role is as general as *agent*, only hyponyms should be accepted. On the other hand, if the role is as specific as *runner*, the WordNet distance between modifier and role can be quite large, and should definitely not be restricted to hyponyms.

In its present form, PropBank is completely unsuitable for this task. The only option to use this resource would be to manually correct these shortcomings, but this would be extremely time consuming, and may lead to new errors. We therefore investigated whether FrameNet or VerbNet would form better options.

FrameNet can be used to relate lexical items to semantic frames, which are defined as "schematic representations of situation types (eating, spying, removing, classifying, etc.) together with lists of the kinds of participants, props, and other conceptual roles that are seen as components of such situations."[2] It contains over 10,000 lexical units, mapping to over 825 frames. The definition of each frame as well as the lexical units relating to it are discussed and carefully checked. Thanks to this perfectionist attitude in the creation of the database, FrameNet is free of inconsistencies that PropBank contains.

However, FrameNet is not ideal for our purposes either. The label assigned to an actant does not necessarily provide information on the set of entities that can fulfil the slot. The role *Agent* will generally refer to animate entities, and has a corresponding entry in WordNet, but it is less trivial to map *Theme* to an appropriate WordNet entry. The definition of *theme* in WordNet is far too specific, and cannot be used in the same way as *agent*. When looking at the variety of verbs assigning a theme role, it is clear that almost anything can be a theme of some verb. The information provided in FrameNet is thus not specific enough for our purposes.

VerbNet is in that sense more complete: next to information on thematic roles, it presents selectional restrictions (when they are imposed by the verb). For each group of verbs, it is indicated whether the agent needs to be animate, or an organisation, whether the argument is a location or a path, whether it is concrete or abstract. This information would allow testing whether the modifier in a compound could be a good argument for the verb.

The only obstacle is that most descriptions of selectional restrictions cannot be mapped to WordNet. Fortunately, there is only a limited set of selectional restrictions used in VerbNet. For each selectional restriction, we selected a WordNet entry that is a hypernym of all semantically plausible arguments of the verb. For instance, the description *animate* mentioned above, was mapped to *living entity*. We created a database of VerbNet entries, where selectional restrictions on arguments are linked to these WordNet entries. A table presenting the selectional restrictions from VerbNet together with WordNet entries we mapped them to, can be found in the Appendix.

[2] http://framenet.icsi.berkeley.edu/index.php?option=com$_c$ontent&task=view&id=53&Itemid=49

3.6 Experiment Three

In our third experiment, we test whether the compound could be considered more likely to form another semantic relation. We combine our nominalisation disambiguation system as described above, to a more general system, which identifies semantic relations. If the relation has a high probability for another semantic relation, we will assume that the compound is not a nominalisation.

From the systems described in chapter 2, we selected the system proposed by Kim and Baldwin (2006). This system uses verb semantics to approximate probabilities of a specific semantic relation. The general method is rather similar to the approach used to find the right grammatical relation in nominalisations. For each pair of nouns forming a compound, instances are retrieved where one of the nouns is the subject and the other noun is the (prepositional) object. In the next step, the meaning of the verb is linked to one of a set of predefined relations. When this system is combined with our nominalisation disambiguator, the result is a system that tries to cover all compounds, using only verbal meaning and grammatical relations. The implementation of the system based on Kim and Baldwin (2006) is described below.

3.6.1 General Compound Noun Disambiguation

Kim and Baldwin (2006) adopt the classification system proposed by Barker and Szpakowicz (1998), but remove the class PROPERTY, because it is too general. We use the exact same classification in our system. This corresponds to the classification presented in table 1.3, except for the relation PREP OBJECT, which is only treated by our nominalisation disambiguator. The relations SUBJECT and OBJECT are thus covered in both systems.

Kim and Baldwin (2006) use information from verbal relations to find the relation between two nouns. The idea becomes rather intuitive, when we consider the name Levi (1978) gave to her semantic classes: "deletable predicates".[3] When two nouns form a compound, occurrences where both nouns are arguments of the same transitive predicate are retrieved from a corpus. These predicates are used to make a guess about the relation between the two nouns. This approach can thus be seen as investigating what predicate would be most likely to have been "deleted".

For each semantic class, Kim and Baldwin (2006) define a set of verbs that express the meaning of the class. These verbs are the *seed verbs* of the relation. The relation LOCATION , for instance, has the verbs *locate, place, site, settle* and *base* as seed verbs. Templates based on these seed verbs are created. The two following templates, point to the POSSESSOR relation:

1 Modifier *owns* Head

2 Head *belongs to* Modifier

Each semantic relation, except for TIME and EQUATIVE, has a set of seed verbs. We create a corpus that contains pairs of nouns and the verb relating them. The data is retrieved from the BNC, the Brown Corpus, and the Wall Street Journal part of the PennTreeBank parsed by RASP. Similarities between predicates in our corpus and seed verbs are obtained using both WordNet and Moby's Thesaurus. This information is used to relate the predicates to seed verbs.

Kim and Baldwin (2006) calculate the similarity between the predicate and the root verb. They replace all predicates by the most similar root verb in their corpus. We create a corpus of tuples containing two nouns and a seed verb in the same way.

Two relations of the classification, namely EQUATIVE and TIME, cannot be identified this way. For the TIME relation, Kim and Baldwin (2006) use CoreLex (Buitelaar (1998)). If the modifier is part of the *TIME* class of CoreLex, the compound will be classified as a TIME relation. Because the

[3] Kim and Baldwin (2006) only refer to Levi (1978) when discussing different classification systems. It is therefore unlikely that they were inspired by Levi's idea that compounds come to be by a predicate deletion process. The approach proposed by Kim and Baldwin (2006) does not suggest that compounds are indeed formed by deletion of predicates. We do not wish make such a claim either.

TIME class in CoreLex is formed by instances retrieved from WordNet, and WordNet is already integrated in our system, we use WordNet for this purpose. The TIME relation is assigned if the modifier is a hyponym of *time*.
For the EQUATIVE relation, coordination structures are used. If the two nouns occur in a coordination together, this indicates that they may be used as equatives in a compound. For all compounds, coordinations are retrieved next to verbal relations. The weight for the equative relations is computed as follows (Kim and Baldwin (2006)):[4]

$$EQUATIVE(NC_i) = -\log_2 \frac{freq(coord(M_i, H_i))}{freq(M_i) \cdot freq(H_i)} \tag{3.4}$$

We created a database containing pairs of nouns that occur in coordination in the BNC, accompanied by the frequency of this coordination. For the frequencies of the individual nouns, we use Kilgarriff's frequencies of words occurring in the BNC.[5]
Kim and Baldwin (2005) use the classifier TiMBL (Daelemans et al. (2004)). Because various factors can influence the distribution of seed verbs, this classifier which implements the k-nearest neighbour principle can help to improve results. Even though achieving higher results is important, we chose not to use a machine learning technique in our system. We are mostly interested in the direct relation between frequencies from our corpus and the semantic classes themselves.

3.6.2 Combining the two systems

In this experiment, compounds with a deverbal head, are disambiguated by both the nominalisation disambiguator (the first system) and the general (second) system. Compounds that do not have a deverbal head, are only

[4] The original equation (Kim and Baldwin (2006), p.495) contained "*TIME*" instead of "*EQUATIVE*". We assume this is an error, and the equation was meant to look as presented here.

[5] http://www.kilgarriff.co.uk/BNC$_l$ists/all.num.o5

treated by the second system.

The output of both systems corresponds to a set of frequencies. The first system provides frequencies for each grammatical relation, the second system provides frequencies of verbs which have semantics that points to a specific semantic relation. An intuitive way to combine the output of each system would be to compare these frequencies directly.

However, changes of finding a specific noun verb pair, are much higher than finding two specific nouns being arguments of the same verb. Despite the fact that we use a larger corpus for our second system, data sparseness forms a much bigger problem for this approach. Therefore, we also test performance of the system when the output of the second system is always preferred over the output of the nominalisation disambiguator.

In addition to these two comparisons, we also compare z-scores. For the nominalisation disambiguator, we use z-scores of SUBJ-OBJ and SUBJ-POBJ. In the second system, we compare z-scores of all attested relations. Because we include secondary synonyms from Moby Thesaurus when verbs are mapped to seed verbs, most verbs are mapped to more than one seed verb. As a result, all modifier-head pairs in our data that co-occur in our corpus get frequencies for more than one relation.

In general, more z-scores will be used in the second system than in the first. As explained in section 2.4, this introduces a slight bias towards results coming from the second system. In case the second system assigns equal (highest) z-scores to two relations, the most frequent occurring relation is assigned.

Chapter 4

Evaluation

4.1 Evaluation Data

We use three datasets to evaluate our system: Lapata (2002)'s manually selected data (henceforth dataset 1) , Nicholson (2005)'s data retrieved from open text (dataset 2), and our own dataset (dataset 3). Each of these datasets is described below.

Dataset 1 (Lapata (2002))

In the first set of experiments, we test our systems on a dataset created by Lapata. Each token in this set consists of two nouns and a gold-standard tag, indicating whether the compound contains a subject or an object relation.

Lapata retrieved 1,277 noun-noun sequences from the BNC, which had a deverbal head that was listed as a nominalisation in either CELEX or NOMLEX. The sample was manually inspected and prepositional relations (9.2%) and compounds exhibiting a semantic relation (28.0%) were removed from the dataset.

The data was annotated by two graduate students, using three sentences of context (the preceding sentence, the sentence containing the compound, and the following sentence). Annotators came to an agreement of 89.7% (Kappa coefficient 0.78 (Siegel and Castellan (1988))).

We tested all three systems on this dataset consisting of 800 tokens.[1]
We include Lapata's data to test the performance of our basic system. The main idea behind our experiments is to see how well a nominalisation disambiguation system performs on open data. It is therefore essential to see that the component identifying specific grammatical relations performs well. Note though, that in our experiments, we try to distinguish compounds containing a grammatical relation from compounds containing a semantic relation. Because Lapata only used compounds that she had identified as nominalisations, our experiments can only result in a drop of recall on this data set.

Dataset 2 (Nicholson (2005))

Nicholson is interested in the performance of his system on open text. The data he uses to test performance was developed accordingly. Annotators were asked to identify and classify compounds from 1,000 sentences retrieved randomly from the BNC. 464 compound nouns were identified in this text, 35 of which were excluded because they consisted of more than two nouns, and 119 because they contained a proper noun.
The compounds were annotated according to five classes: NA (a deverbal head, but no grammatical relation), NV (not a deverbal), SU (contains a subject relation), DO (contains a direct object relation), PO + preposition (contains a prepositional object relation). The distribution of the corpus is presented in table 4.1 (Nicholson (2005), p.17) below.

Nicholson's data includes all necessary information to test the performance of our systems, which are designed to identify the three possible grammatical relations, and compounds that do not contain a nominalisation. The most interesting class in Nicholson (2005)'s data is the NA class. Our system extends the experiment carried out by Nicholson with components that try to distinguish NA compounds from those containing a grammatical re-

[1] Lapata uses 796 tokens in her investigation. Her manually selected data should consist of 802 tokens (62.8% of 1,277). It is not clear why Lapata does not use all data in her own experiment, nor which data in the set (of 800) she sent to us was not present in her own experiment.

Table 4.1: Distribution within Nicholson's data

Class	Example	Frequency	Percentage
SUBJ	*eyewitness report*	22	(6.4%)
OBJ	*eye irritation*	63	(18.2%)
POBJ	*side show*	63	(12.8%)
NV	*scout hut*	58	(16.8%)
NA	*memory size*	158	(45.8%)

lation. Therefore this group will provide the clearest indication whether our identification techniques work.

Dataset 3

Next to testing how well our systems perform on the identification of nominalisations, we wish to test performance of our systems on a corpus annotated according to the twenty classes presented in table 1.3. For this purpose we created a dataset of over 900 tokens.

Just like Nicholson, we are interested in finding the performance of our system on open text. We retrieved sequences of two nouns from the BNC, using the BNC part-of-speech tags, and randomly selected 1,500 tokens from this set.

Our sample of 1,500 contained 1,347 unique pairs. Because the BNC contains tagging errors, and we purposely did not use a chunker or parser to see whether the sequence of nouns syntactically functions as one noun, not all examples in our set were compound nouns. We inspected the data set and selected the first 950 examples that seemed good candidates, and 50 bad examples to be used as filters. The filter examples consisted of tagging errors (14), bracketing errors (21), metaphorical relations (9), foreign words (3)[2], genitives (where the *'s* is missing) (2), and one spelling error. They were randomly distributed through the dataset.

Two native English speakers annotated the entire data set, and a third native speaker annotated 400 tokens. Compound nouns were presented together

[2] As one of the annotators pointed out, none of the foreign words were nouns.

with three sentences of context (preceding sentence, sentence containing compound and following sentence). Annotators were instructed to use the context to identify the most appropriate relation, and to classify each compound according to a classification containing the 20 classes presented in table 1.3, METAPHOR and NO SELECTION, thus a total of 22 options. Annotators were told that the data was filtered for metaphors and errors, and instructed to try to assign one of the 20 classes, when possible. The classes METAPHOR and NO SELECTION should only be used when contextual factors rule all other classes out.

The two annotators working on the complete data set assigned the same class to 54.2% of the data. In order to get a clearer indication of the agreement between annotators, we use Cohen's Kappa coefficient (Cohen (1960)), which corrects for coincidental agreement. Kappa (K) is calculated with the following formula:

$$K = \frac{P(\alpha) - P(E)}{1 - P(E)} \tag{4.1}$$

Where $P(\alpha)$ stands for the observed agreement between the annotators, and $P(E)$ represents the probability of the agreement being by chance. The Kappa-scores of agreement between the annotators is presented in table 4.2. The agreement between annotators A1 and A2 was $K = .503$ over the entire data set, which is a good result considering the difficulty of the task. Both annotators who annotated the full set (A1 and A2) showed good performance on detecting filter examples; annotator (A1) classified forty-eight as NO SELECTION or METAPHOR, annotator (A2) found forty-four. Forty-two of the fifty filter examples were identified by both annotators. All eight examples

Table 4.2: Inter-annotator Agreement

Annotators	Total Annotated	Total Agreed	Random Expectation	K
A1, A2	1,000	542	78.9	0.503
A1, A3	400	183	30.0	0.414
A2, A3	400	194	28.3	0.446
A1, A2	400	207	31.5	0.476

that had been missed by one of the annotators were part of higher-arity compounds, where bracketing would preferably apply in a different order.

Annotator (A1) and (A2) met, and discussed differently classified examples to form a gold-standard. The standardised data set contained 24 metaphors, and 59 examples for which no appropriate class could be assigned. This resulted in a set of 917 examples labelled according to the 20-fold classification. Table 4.3 represents the distribution of different classes, table 4.4 the distribution of nominalisation (NOM), and other compounds (NO NOM).

Table 4.3: Distribution of 20 Classes

Class	Tokens	Percentage	Class	Tokens	Percentage
SUBJECT	34	3.70%	LOCATED	19	2.07%
OBJECT	102	11.12%	LOCATION	78	8.51%
POBJECT	46	5.02%	MATERIAL	51	5.56%
BENEFICIARY	9	0.98%	POSSESSOR	61	6.65%
CAUSE	13	1.42%	PRODUCT	15	1.64%
CONTAINER	27	2.94%	PURPOSE	124	13.52%
CONTENT	41	4.47%	RESULT	10	1.09%
DESTINATION	2	0.22%	SOURCE	18	1.96%
EQUATIVE	34	3.71%	TIME	28	3.05%
INSTRUMENT	22	2.40%	TOPIC	183	19.96%

Table 4.4: Distribution of Nominalisations

Class	Tokens	Percentage
SUBJECT	34	3.70%
OBJECT	102	11.12%
POBJECT	46	5.02%
NO NOM	735	80.15%
NOM	182	19.85%

Note that the percentage of nominalisations identified by our annotators forms only 19.85% of the data. In Nicholson's dataset 37.4% of the compounds were interpreted using a grammatical relation. Both of the datasets are created from randomly selected compounds. With datasets containing over 900 and over 400 tokens respectively, it is unlikely that this difference lies in the nature of the datasets. Rather, the explanation should be sought in the different tasks assigned to the annotators.

Nicholson's annotators were asked to find nominalisations in a corpus. Our annotators, on the other hand, had to select the most appropriate relation from a set containing grammatical relations as well as semantic relations. In Nicholson's study, any compound that could possibly be interpreted as a nominalisation was identified as such by the annotators. Context was not taken into account, nor did the annotators check whether a grammatical relation formed a good interpretation of the compound. Nicholson's dataset contains nominalisations such as *school headmaster* (OBJECT, "to headmaster a school"), *database partner* (PREP OBJECT, "to partner in databases") and *mountain chain* (OBJECT "mountains are chained"). One could argue that the grammatical relations are possible interpretations in these cases, but semantic relations may provide better options.

In our investigation, annotators were asked to find the most appropriate relation from a larger set of possibilities. Compounds that are possible nominalisations, but preferably interpreted by another relation, are excluded from the nominalisation set. In addition, the annotators were instructed to take context into account. If a possible grammatical relation did not provide the best interpretation in the given context, our annotators assigned a different relation to the compound. Because only those compounds where grammatical relations provide the most likely interpretation are considered nominalisations, we believe that our data set provides a more accurate account of the frequency of nominalisations.

Note, however, that in our classification nominalisations were given a small preference over other relations as well. As we have pointed out in chapter 1, some of the semantic classes reveal similarities with thematic roles assigned by verbs. Because of this, some compounds can either be classified as one of the three grammatical relations, or as one of the semantic relations. Annotators were instructed to assign the class that seemed most appropriate. The first annotator noticed the overlap when annotating the first example (namely, *drug test*), and asked for guidelines. Therefore, we added the instruction that only in cases where both a grammatical and a semantic relation would result in a fully acceptable interpretation of the compound, the grammatical relation should be preferred over the semantic relation. We used the examples in table 4.5 (which do not occur in our data) to illustrate these instructions.

Table 4.5: Compounds with Grammatical or Semantic Relation

Example	Possible Relations	Preference
student protest	SUBJECT or SOURCE	SUBJECT seems more appropriate and should be assigned.
book review	OBJECT or TOPIC	Both seem appropriate, (maybe) assign OBJECT.
tomato can	CONTENT or OBJECT	CONTENT is more appropriate and should be assigned.

4.2 The Basic System

Before running the experiments described in chapter 3, we tested the basic system underlying our experiments, in order to make sure it shows equal performance to the classifiers made by Lapata and Nicholson. We repeated the experiments carried out by both authors. Table 4.6 presents the accuracy of our system, together with Lapata and Nicholson's results, respectively.

As table 4.6 reveals, our classification system performs comparable to both results achieved by Lapata (2002)[3] and by Nicholson (2005). Our system's accuracy is slightly higher than Lapata's, and lower than Nicholson's, though none of these differences are significant.
Our system achieves high recall of 94.6% on the nominalisation detection

[3] Lapata's best result is 86.1%, though this was only achieved when morphological information and direct context was taken into account. We compared our results to her experiment using only corpus statistics and WordNet smoothing.

Table 4.6: Accuracy of Basic System

Data & Classification	Our System	Lapata	Nicholson	Baseline
Dataset 1 Two-fold	72.6%	68.8%	72.1%	62.9%
Dataset 2 Two-fold	52.3%	–	52.7%	48.8%
Dataset 2 Three-fold	55.8%	–	56.1%	48.8%
Dataset 3 Two-fold	58.2%	–	–	56.0%
Dataset 3 Threefold	58.2%	–	–	56.0%
Dataset 2 Recall	123 (94.6%)	–	90 (69.8%)	–
Dataset 3 Recall	181 (99.5%)	–	–	–

task, compared to 69.8% in Nicholson (2005). This improvement can partially be explained by the step wise method applied to select data from Celex, increasing the set of reliable noun-root verb pairs from this resource with almost 4,000 tokens. Note though, that Nicholson tries to detect nominalisations in sentences. Part of his loss in recall can be explained because the nominalisation was not identified as a compound noun by his chunker and part-of-speech tagger.

In the previous chapter, we mentioned that Lapata (2002) performs smoothing for each relation that does not occur in the corpus. Nicholson (2005), on the other hand, only performs smoothing when relations are equally likely. We hypothesised that this difference could provide the main explanation for Nicholson's superior performance for two reasons. First, when Lapata's method is used, there are cases where approximated statistics are compared to statistics that are directly derived from the corpus. Second, intuitively, when one relation is found in the corpus where the other is absent, this could very well be evidence for the attested relation being the right interpretation

for the compound.
We tested both hypotheses on Lapata's data set. First, we performed overall smoothing when either one of the relations was not attested. In this experiment, our system achieved an accuracy of 70.1%, which is less high than the performance in general (72.6%), but the difference is not significant. When we repeated Lapata's method where relations were smoothed individually, accuracy dropped to 66.4%, a performance that is significantly worse than when overall smoothing was applied. Our hypothesis that comparing approximated frequencies to direct frequencies leads to errors is confirmed by this result.
The results presented in this section show that our basic system achieves similar performance to both Lapata (2002) and Nicholson (2005). The results from our experiments will thus provide a reliable account of the performance of the extensions we propose to the systems.

4.3 Experiment One, Evaluation

In our first experiment we test whether occurrences of the modifying noun — or a noun similar to the modifying noun— as an argument of the head noun's root verb can be used as evidence for the compound to contain a nominalisation. First, we only treat those compounds for which we find pairs of the modifying noun and root verb in our corpus as possible nominalisations. In a following experiment, we also include synonyms of the modifier, and its hyponyms. In the last part of the experiment, we include nouns within a specific WordNet distance from the modifier, next to synonyms and hyponyms. In this case, nouns that are within a WordNet-distance defined by a threshold θ of the modifier, are included in the smoothing process. If there are no occurrences of the root verb with a noun which dissimilarity to the modifier is smaller than θ, the compound will be classified as not a nominalisation.

4.3.1 Experiment One, Nominalisation set

The system was first tested on datasets consisting only of nominalisations (dataset 1, and subsets of dataset 2 and 3, respectively). Our filtering system can only result in a drop of recall on these sets. We test our system on the nominalisation sets in order to see whether the filter causes a significant decrease in accuracy. The table below presents the difference in accuracy between performance of our basic system, and performance of the system using a filter. Tables containing all results on each dataset can be found in the appendix.

The results for dataset 1 correspond to the filter that includes the modifier, its synonyms and its hyponyms (Mod Syn Hypo). Recall drops significantly when this filter is used.

Table 4.7: Differences in recall on nominalisation data

Dataset	Basic System	Filter	χ^2 (prob level)
1 (800, Mod Syn Hypo)	72.3%	58.8%	32.3 ($p < 0.01$)
2 (129, Mod)	54.3%	44.2%	2.62 ($p < 0.5$)
3 (182, Mod)	57.7%	48.4%	3.18 ($p < 0.1$)

For the smaller sets, even the strongest filter that only treats compounds whose modifier is an attested argument of the root verb as compounds, the decrease in accuracy is not significant. However, accuracy drops with 10.1% and 9.3% on these datasets. The small size of these sets (129 tokens, and 182 tokens, respectively) makes these differences insignificant.

4.3.2 Experiment One, Open set

The tables below present performance on nominalisation detection of our first experiment on dataset 2 and dataset 3. The first column represents the data that was considered evidence for a grammatical relation. In *Mod*, only instances where the modifier itself occurred in the corpus as an argument of the root verb were considered nominalisations. *Mod, Syn* includes the modifier and its synonyms, and in *Mod, Syn, Hypo*, hyponyms are considered evidence for a nominalisation as well. The θ values represent the thresholds of WordNet dissimilarity values. As θ increases, more data is included as evidence for a grammatical relation.

The column *Gram Rel* represent precision, recall and F-score ($\beta = 1$) on nominalisations, *Not Nom* presents these value for compounds that are not nominalisations. The column *Accuracy* presents the accuracy on nominalisation identification.

The baseline values correspond to the output of our basic system, without a filter applied. The results reveal that the method is successful in improving precision on nominalisation detection. In general, precision increases when less data is included as evidence. When only the modifier, its synonyms and hyponyms are considered evidence, results are significant on both datasets, with best results being an accuracy of 76% on dataset 2 (baseline = 64%)

Table 4.8: Experiment One, Dataset 2 (335), Nom detection

Included Data	Gram Rel			Not Nom			Accuracy
	P	*R*	*F*	*P*	*R*	*F*	
Mod	65.8	67.5	66.7	82.3	81.2	81.8	76.4%
Mod, Syn	65.6	68.4	66.9	82.6	80.7	81.7	76.4%
Mod, Syn, Hypo	63.4	70.9	66.9	83.3	78.0	80.6	75.5%
$\theta = 0.05$	62.7	76.1	68.7	85.5	75.7	80.3	75.8%
$\theta = 0.10$	62.0	79.5	69.7	87.0	73.9	79.9	75.8%
$\theta = 0.15$	58.9	84.6	69.5	89.2	68.3	77.4	74.0%
$\theta = 0.20$	57.6	84.6	68.5	89.0	66.5	76.1	72.8%
$\theta = 0.30$	52.8	87.2	65.8	89.4	58.3	70.6	68.4%
$\theta = 0.40$	52.6	88.0	65.8	89.9	57.3	70.0	68.1%
$\theta = 0.50$	52.8	89.7	66.5	91.2	56.9	70.1	68.4%
$\theta = 0.60$	52.2	89.7	66.0	91.0	56.0	69.3	67.8%
$\theta = 0.70$	52.5	90.6	66.5	91.7	56.0	69.5	68.1%
$\theta = 0.80$	52.5	90.6	66.5	91.7	56.0	69.5	68.1%
$\theta = 0.90$	52.2	90.6	66.3	91.7	55.5	69.1	67.8%
$\theta = 1.00$	52.2	90.6	66.3	91.7	55.5	69.1	67.8%
Baseline	49.1	94.0	64.5	93.7	47.7	63.2	63.9%

Table 4.9: Experiment One, Dataset 3 (917), Nom detection

Included Data	Gram Rel			Not Nom			Accuracy
	P	R	F	P	R	F	
Mod	37.8	79.1	51.2	92.9	67.8	78.4	70.0%
Mod, Syn	36.1	79.7	49.7	92.8	65.0	76.5	67.9%
Mod, Syn, Hypo	35.7	82.4	49.8	93.6	63.3	75.5	67.1%
$\theta = 0.00$	34.8	86.8	49.7	94.8	59.7	73.3	65.1%
$\theta = 0.05$	34.9	87.4	50.0	95.0	59.8	73.4	65.2%
$\theta = 0.10$	33.1	89.6	48.3	95.5	55.1	69.9	61.9%
$\theta = 0.15$	31.9	92.3	47.5	96.4	51.3	67.0	59.4%
$\theta = 0.20$	30.8	92.9	46.2	96.5	48.3	64.4	57.1%
$\theta = 0.30$	28.6	95.1	44.0	97.1	41.4	58.0	52.0%
$\theta = 0.40$	27.8	95.6	43.1	97.3	38.6	55.3	50.0%
$\theta = 0.50$	27.6	96.2	42.9	97.5	37.6	54.2	49.2%
$\theta = 0.60$	27.3	96.7	42.6	97.8	36.2	52.8	48.2%
$\theta = 0.70$	27.3	97.3	42.7	98.1	40.0	52.6	48.1%
$\theta = 0.80$	27.2	97.8	42.6	98.5	35.2	51.9	47.7%
$\theta = 0.90$	27.1	97.8	42.4	98.5	34.8	51.5	47.3%
$\theta = 1.00$	27.1	97.8	42.4	98.5	34.8	51.5	47.3%
Baseline	25.4	99.5	40.5	99.5	27.8	43.4	42.0%

and 70% on dataset 3 (baseline = 42%). On dataset 3, accuracy already increases significantly when only a small filter ($\theta = 1.00$) is used ($\chi^2 = 5.30$, $p < 0.05$). The filter $\theta = 1.00$ only excludes examples where either the root verb does not occur in the BNC, or the modifier is not listed in WordNet.
The gain in precision goes at the cost of recall. Two observations can be made on the overall effect of our filter. On the one hand, we observe a significant drop in recall before precision increases significantly. On the other hand, F-scores increase steadily when a stronger filter is used.
Next to results on nominalisation identification, we investigate the influence of our filter on precision and recall of each grammatical relation. The tables below represent these results for dataset 2 and 3 respectively.

As expected, results on the individual relations behave similarly to the results on nominalisation detection. As the filter excludes more tokens, precision in-

Table 4.10: Experiment One, Dataset 2 (335), Gram rel

Included Data	OBJECT			SUBJECT			POBJECT		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Mod	55.6	57.9	56.4	25.0	66.7	36.4	100	10.3	18.6
Mod, Syn	55.0	57.9	56.4	24.1	66.7	35.4	100	10.3	18.6
Mod, Syn, Hypo	50.0	59.7	54.4	23.7	66.7	35.0	100	10.3	18.6
$\theta = 0.00$	46.2	63.2	53.3	23.7	66.7	35.0	100	10.3	18.6
$\theta = 0.05$	49.3	64.9	56.1	22.2	66.7	33.3	100	10.3	18.6
$\theta = 0.10$	47.6	68.4	56.1	21.9	66.7	32.9	100	10.3	18.6
$\theta = 0.15$	44.1	71.9	54.7	19.7	66.7	30.4	100	10.3	18.6
$\theta = 0.20$	42.3	71.9	53.2	19.7	66.7	30.4	100	10.3	18.6
$\theta = 0.30$	37.0	70.2	48.4	18.5	71.4	29.4	100	10.3	18.6
$\theta = 0.40$	36.7	71.9	48.5	20.0	76.2	31.7	100	10.3	18.6
$\theta = 0.50$	36.4	70.2	47.9	18.8	76.2	30.2	100	10.3	18.6
$\theta = 0.60$	37.7	71.9	49.4	18.2	76.2	29.4	100	10.3	18.6
$\theta = 0.70$	37.5	73.7	49.7	18.6	76.2	29.9	100	10.3	18.6
$\theta = 0.80$	37.2	73.7	49.4	17.6	71.4	28.3	100	10.3	18.6
$\theta = 0.90$	37.2	73.7	49.4	17.4	71.4	28.0	100	10.3	18.6
$\theta = 1.00$	37.2	73.7	49.4	17.4	71.4	28.0	100	10.3	18.6
Baseline	33.6	78.9	47.1	17.4	71.4	28.0	100	10.3	18.6

creases at the cost of recall. Objects benefit most from the filter system, though F-scores increase for both relations.

The results of the first experiment indicate that occurrences of a compound's modifier as an argument of the verb provide a reliable indication of the compound being a nominalisation. The overall best results leading to an improvement of 28% in accuracy on the nominalisation identification task.

4.4 Experiment Two, Evaluation

In the second experiment, we test whether the selectional restrictions of the root verb allow for the modifier to be one of the verbs arguments. Just like

Table 4.11: Experiment One, Dataset 3 (917), Gram Rel

Included Data	OBJECT			SUBJECT			POBJECT		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Mod	36.9	66.7	47.6	9.8	50.0	16.4	12.5	6.5	8.6
Mod, Syn	34.8	67.6	46.0	9.4	50.5	15.9	12.5	6.5	8.6
Mod, Syn, Hypo	33.7	68.6	45.2	9.6	52.9	16.2	12.5	6.5	8.6
$\theta = 0.00$	31.0	73.5	43.6	9.6	52.9	16.2	12.5	6.5	8.6
$\theta = 0.05$	31.3	73.5	43.9	9.4	52.9	16.0	12.5	6.5	8.6
$\theta = 0.10$	29.8	76.5	42.9	8.7	52.9	14.9	12.5	6.5	8.6
$\theta = 0.15$	29.0	79.4	42.5	8.5	55.9	14.8	12.5	6.5	8.6
$\theta = 0.20$	27.6	79.4	40.9	8.2	55.9	14.3	12.5	6.5	8.6
$\theta = 0.30$	24.8	78.4	37.7	7.4	55.9	13.0	12.5	6.5	8.6
$\theta = 0.40$	24.2	80.4	37.2	7.3	55.9	12.8	12.5	6.5	8.6
$\theta = 0.50$	23.6	80.4	36.5	7.2	55.9	12.8	12.5	6.5	8.6
$\theta = 0.60$	23.0	80.4	35.7	7.2	55.9	12.8	12.5	6.5	8.6
$\theta = 0.70$	23.0	80.4	35.8	7.1	55.9	12.6	12.5	6.5	8.6
$\theta = 0.80$	22.8	80.4	35.5	7.0	55.9	12.5	12.5	6.5	8.6
$\theta = 0.90$	22.7	80.4	35.3	7.0	55.9	12.5	12.5	6.5	8.6
$\theta = 1.00$	22.7	80.4	35.3	7.0	55.9	12.5	12.5	6.5	8.6
Baseline	19.9	81.4	32.0	7.0	55.9	12.5	12.5	6.5	8.6

the first system, we examine performance on the three datasets containing only nominalisations, and on the full datasets 2 and 3. In addition, the system provides an output indicating which modifiers did not fit selectional restrictions of the head noun's root verb. We used this output to check whether the general methodology worked, and to see at what stages errors may have been introduced.

4.4.1 The Selectional Restriction Filter

Table 4.4.1 presents some example entities of argument-verb relations that were excluded by the system.

For the first three examples, the system took the right decision by excluding the modifier as a possible argument of the verb: shoes do not shop, spoons do not race and the autumn does not tree, or get treed.

Table 4.12: The Selectional Restriction Filter

Compound	Root-verb	Selectional Restriction	Correct
shoe shop	*shop*	$SUBJ$ = animate	yes
spoon race	*race*	$SUBJ\&OBJ$ = animate	yes
autumn tree	*tree*	$SUBJ\&OBJ$ = concrete	yes
government inquiry	*inquire*	$SUBJ$ = animate	no
ticket arrangement	*arrange*	OBJ = concrete	no
classroom assistance	*assist*	OBJ = animate	no

The last three examples reveal cases where the filter should not have applied for different reasons. In *government inquiry*, the restriction provided by WordNet is too limited: an organisation such as government can also inquire. The problem in *ticket arrangement* can be found in WordNet: our algorithm mapped the selectional restriction *concrete* to *physical entity*; *ticket* should have been recognised as a hyponym of this entry.

In the last example, the system is not at fault. VerbNet correctly indicated that only living things are assisted, and the result from WordNet that a classroom is not animate is also correct. However, *Classroom* is used as a metonymy for the students in the classroom, and can therefore be interpreted as the object of *assist*. We cannot expect that our selectional restriction filter handles these metaphorical examples correctly.

We manually checked the output of the filter run on dataset 3 and evaluated the judgements of the system. The system excluded 116 argument relations, 107 of which seemed appropriate (92%). Note though, that this does not imply that our system correctly identified 107 compounds with deverbal head that are not nominalisations. On the one hand, the modifier can be (wrongly) accepted as another argument of the verb. On the other hand, a candidate accepted as a nominalisation may have another preferable interpretation. The next section presents the results on the nominalisation identification task when the selectional restriction filter is applied.

4.4.2 Results

Contrary to the first experiment, the method checking selectional restrictions does not lead to a significant drop in recall on nominalisation data. Results of the experiments carried out on the nominalisation sets can be found in the appendix.

The tables below present the results of nominalisation detection on dataset 2 and 3, respectively. We test performance of the system on both the full dataset (Full) , and a smaller dataset (VN), in which all deverbal head nouns have at least one root verb present in VerbNet.

Table 4.13: Experiment Two, Dataset 2 (335, 278), Nom detection

Data	Gram Rel			Not Nom			Accuracy
Full (335)	*P*	*R*	*F*	*P*	*R*	*F*	
Basic	49.1	94.0	64.5	93.7	47.7	63.2	63.9%
VN-Sys	51.5	90.6	65.6	91.5	54.1	68.0	66.9%
VN (278)							
Basic	54.5	92.9	68.7	93.7	57.8	71.5	70.1%
VN-Sys	58.4	88.8	70.4	91.5	65.6	76.4	73.7%

Table 4.14: Experiment Two, Dataset 3 (917, 701), Nom detection

Data	Gram Rel			Not Nom			Accuracy
Full (917)	*P*	*R*	*F*	*P*	*R*	*F*	
Basic	25.4	99.5	40.5	99.5	27.8	43.4	41.9%
VN-Sys	25.8	96.7	40.8	97.5	31.3	47.4	44.3%
VN (701)							
Basic	30.4	99.3	46.6	99.5	37.2	54.1	50.6%
VN-Sys	31.4	96.1	47.3	97.5	41.9	58.6	53.6%

The results reveal an increase in precision at the cost of a loss in recall. As in the first experiment, F-score goes up, as well as accuracy, although none of the differences are significant. This may be due to the relatively small size of the dataset.

Tables 4.15 and 4.16 present performance on individual grammatical relations. Again, we present results on the complete dataset, as well on the reduced dataset including only deverbal heads listed in VerbNet. Results on the four-fold classification (subject, object, pobject, not a nominalisation) are presented in VN-All. VN-Id presents results on the identified nominalisation set.

Note that in both datasets, precision on subject identification decreases when the filter is applied. This could be related to the fact that we use selectional restrictions both as a filter and to assign relations. If a modifier cannot be the object of a verb, but does fit the restrictions of the subject-role, a subject relation will be assigned.

Table 4.15: Experiment Two, Dataset 2, (335, 278), Gram rel

	SUBJECT			OBJECT			POBJECT			Acc
Full (335)	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
Basic	35.6	63.2	45.6	63.1	76.3	69.1	100	13.6	23.9	49.5%
VN-All	33.9	66.7	45.0	65.1	69.5	67.2	100	13.6	23.9	52.8%
VN-Id	33.9	74.5	46.6	65.1	71.9	68.3	100	16.8	28.8	64.1%
VN (278)										
Basic	34.9	62.8	44.9	63.5	77.0	69.6	100	14.3	25.0	56.5%
VN-All	32.2	65.7	43.2	65.3	69.0	67.1	100	14.3	25	59.4%
VN-Id	32.2	74.2	44.9	65.3	71.6	68.3	100	17.7	30.1	74.3%

Table 4.16: Experiment Two, Dataset 3 (917, 701), Gram rel

	SUBJECT			OBJECT			POBJECT			Acc
Full (917)	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
Basic	37.5	62.2	46.8	65.5	79.0	71.6	90.3	11.9	21.1	35.9%
VN-All	35.7	58.8	44.4	66.1	74.5	70.0	60.0	6.5	11.8	35.9%
VN-Id	35.7	60.6	44.9	66.1	76.8	71.0	60.0	6.8	12.2	46.4%
VN (701)										
Basic	42.6	64.5	51.3	70.0	81.4	75.2	75.0	8.6	15.4	42.3%
VN-All	36.7	61.5	46.0	67.6	73.4	70.4	66.6	7.4	14.9	42.3%
VN-Id	36.7	63.5	46.5	67.6	75.8	71.5	66.6	7.8	14.0	50.8%

The trade-off between precision and recall seems to lean more towards a more significant loss in recall than a gain in precision: F-scores go down, and accuracy does not increase at all when the filter is applied on dataset 3. In dataset 2, accuracy does increase but the difference is not significant.

4.4.3 Discussion of Experiment 2

Despite the fact that the methodology seems to work well, the application of a selectional restriction filter did not lead to a significant improvement of precision in any of our experiments. There are two possible explanations of why results from the second experiment are not significant. First of all, our filter only checks whether a specific modifier can be the argument of the root verb. The fact that the modifier is a possible argument, does not mean it is plausible argument. In this sense the filter is too weak.

On the other hand, an explanation can be found in the resources used. Some of the selectional restrictions are too specific. For instance, verbs such as *decide*, *choose* and *investigate* must have a subject that is animate, or an organisation. However, a village can also decide, choose or investigate. In this sense, a village referring to a group of people could be seen as an organisation, but an ontology such as WordNet does not necessarily provide this interpretation.

At the same time, VerbNet includes selectional restrictions that are not specific enough for our purposes. For instance, subjects of verbs such as *talk* and *judge* tend to be human, not just animate. This property of VerbNet can be explained by the fact that the resource defines thematic roles and selectional restrictions for groups of verbs. A verb like *talk* is part of the same group as verbs like *scream*, for which animate is an appropriate expression. A more fine grained resource for thematic roles could therefore improve the results of this experiment.

4.5 Experiment Three, Evaluation

In the third experiment, we use both the nominalisation disambiguator (system 1) and the general compound noun disambiguator described in section 3.6 (system 2). We evaluate three methods of combining the two systems: direct frequency comparison, comparison of z-scores and preference of the output from system 2 over the output from system 1. All three comparison methods were tested on dataset 2 and 3. In addition, we test system 1 on both datasets, and system 2 on dataset 3, the only dataset that contains all 20 classes of the second system.

Performance of the second system was poor: coverage was 14.1%, and only 10.1% of the covered data was classified correctly. There are several explanations for the low performance of this system. They will be discussed in section 4.5.1. Results improved significantly when the system was combined with the nominalisation disambiguator. Coverage went up to 77.1% and 12.1% of this larger dataset was classified correctly according to the 20-fold classification.

Despite its poor individual performance, the second system does lead to an increase in accuracy of the first system. Tables 4.17 and 4.18 present results on nominalisation detection of system 1 and the three combined methods.

Table 4.17: Experiment 3, Dataset 2, Nom detection

Comparison	Gram Rel			Not Nom			Accuracy
	P	*R*	*F*	*P*	*R*	*F*	
Frequency	55.9	85.5	67.6	89.1	63.8	74.3	71.3%
Z-scores	50.8	82.9	63.0	86.1	56.9	68.5	69.8%
Pref. sys2	52.7	66.7	58.9	79.1	67.9	73.1	67.5%
System 1	52.2	90.6	66.3	91.7	55.5	69.1	67.8%

As the tables reveal, accuracy on nominalisation identification goes up for both datasets regardless of the comparison method used. However, the high-

Table 4.18: Experiment 3, Dataset 3, Nom detection

Comparison	Gram Rel			Not Nom			Accuracy
	P	*R*	*F*	*P*	*R*	*F*	
Frequency	28.6	91.2	43.5	95.2	43.5	59.8	53.0%
Z-scores	25.6	84.6	39.3	91.1	39.0	54.7	54.2%
Pref. sys2	25.9	66.5	37.2	86.4	52.8	65.5	55.5%
System 1	27.1	97.8	42.3	98.5	34.8	51.5	47.3%

est increase on dataset 2, using Frequency comparison, is not a significant improvement ($\chi^2 = 1.27$, $p < 0.5$). For dataset 3, on the other hand, all combination methods lead to significantly better accuracy ($\chi^2 \geq 5.90$, $p < 0.05$).

Both datasets respond differently to the comparison methods. The frequency comparison method, which filters out the fewest compounds out, performs best on dataset 2. On dataset 3, best performance is achieved when the output of the second system is always preferred over the output of system 1. A similar observation can be made for results on the individual grammatical relations. They are presented in tables 4.19 and 4.20.

Table 4.19: Experiment 3, Dataset 2, Gram rel

Comparison	SUBJECT			OBJECT			POBJECT			Acc
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
Frequency	15.4	71.4	25.4	48.0	63.2	54.5	57.1	10.3	17.4	57.9%
Z-scores	12.1	61.9	20.3	46.9	66.7	55.1	100	7.69	14.3	57.3%
Pref. sys2	12.2	47.6	19.4	46.9	52.6	49.6	100	5.13	9.76	56.7%
System 1	13.6	76.2	23.0	47.4	64.9	54.8	57.1	10.3	17.4	53.1%

An explanation for the different reaction of each dataset to the filter methods may be found in the nature of the two datasets. As explained in section 4.1, dataset 2 was created by annotators looking for nominalisations in open text. Dataset 3, on the other hand, contains compounds annotated according to the best possible interpretation in a given context. As a result, the percentage of nominalisations in dataset 2 is almost twice as high as the per-

Table 4.20: Experiment 3, Dataset 3, Gram rel

Comparison	SUBJECT			OBJECT			POBJECT			Acc
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
Frequency	5.95	61.8	10.9	35.2	67.6	46.3	9.38	6.52	7.69	45.0%
Z-scores	4.11	44.1	7.52	33.3	71.6	45.5	16.7	6.52	9.38	46.9%
Pref. sys2	2.90	23.5	5.16	33.1	58.8	42.4	18.2	4.34	7.02	49.9%
Basic System	5.81	70.6	10.7	33.5	68.6	45.0	11.4	8.70	9.88	38.6%

centage in dataset 3. This would explain why results improve on dataset 3 when more compounds are excluded from the nominalisation set, whereas a relatively weak filter performs better on dataset 2.

4.5.1 Discussion of Experiment 3

In general, combining the two systems has shown to be a successful approach improving results on both individual systems. Nevertheless, poor performance of the second system forms a problem in this experiment.

First, we cope with low coverage. Kim and Baldwin (2006) report coverage of over 20% when only including nouns that co-occur at least five times in the corpus. We hoped to improve coverage by including synonyms of the components of the compound noun, and by accepting nouns (and synonyms) that co-occur only once. However, despite this extension, coverage was only 14% on our data. It is not clear where the cause of this difference can be found. We followed Kim and Baldwin's methodology precisely in the creation of our corpus, as well as in the mapping of predicates to seed verbs. The only difference between the two approaches is that we used a more recent (improved) version of RASP, but it is unlikely that the cause of the loss in coverage lies here.

Next, there is the difference in accuracy for relation identification. Kim and Baldwin report 53%, our system achieves around 10%. The low performance can partially be explained by our choice not to use machine learning, but is mostly due to the seed verbs that are used as features. The relations and their seed verbs are presented in table 4.21.

First, for some irregularly occurring relations such as POSSESSOR and SUBJECT the set of seed verbs contains highly frequent verbs, like *have* and *do*. The relations PURPOSE and LOCATION, which occur far more often than the two mentioned above, have only verbs that occur relatively rarely. Second, for some of the seed verbs, it was not clear to us how they represent the relation they are meant to identify. It is for instance not clear how conjunction would provide proof for two nouns to stand in an equative relation. Do materials really occur frequently as the subject of *manufacture*, and how does *behave* relate to an object relation, and *consume* to instruments?

Table 4.21: Relations and their Seed Verbs

Relation	Seed verbs
SUBJECT	*perform, execute, act, do, carry_out, accomplish, run*
OBJECT	*act, perform, behave, do, play, work_on, move, practise*
BENEFICIARY	*benefit, receive, profit, gain, acquire, obtain, earn*
CAUSE	*cause, generate, result, induce, lead, bring, stimulate, bring_on, conclude*
RESULT	*cause, generate, result, induce, lead, bring, stimulate, bring_on, conclude*
CONTAIN	*contain, include, encompass, hold, carry, reserve, obtain*
CONTENT	*contain, include, encompass, hold, carry, reserve, obtain*
DESTINATION	*head, bound*
INSTRUMENT	*drive, power, consume, apply, employ*
LOCATION	*locate, place, site, settle, base*
LOCATE	*locate, place, site, settle, base*
MATERIAL	*create, manufacture, produce, compose, construct*
PRODUCT	*create, manufacture, produce, compose, construct*
POSSESSOR	*belong_to, possess, own, possessor, acquire, grab, occupy*
PURPOSE	*mean, intend, imply, signify, aim, desire, expect, match, fit*
SOURCE	*entail, come, imply, mean, involve, evoke, bring_about, cause, lead_to, result_in*
TOPIC	*concern, refer, relate, come_to, bear_on, touch(_on), interest, regard, pertain, associate, link, connect, be_about, deal_with*
EQUATIVE	conjunction

We first noticed these difficulties while implementing the second system. However, it went beyond the scope of our research to design new features for Kim and Baldwin's methodology. Moreover, they achieved good results with this set of verbs, and there was no reason to assume that this would not be the case in our experiment.

After having noticed the poor performance of the system, we inspected the features retrieved from the corpus. It turned out that in most cases, the system had not found any features that would have led to the right interpretation. We therefore stuck to our decision not to use machine learning in this experiment. When features are not informative, machine learning cannot do much. If we had used machine learning, results would have improved. But it would have been an artificial result based on the general distribution of classes over the training data, instead of a good indication on performance of the methodology.

Overall, it is surprising that Kim and Baldwin achieved an accuracy of more than 50%. We believe that the different distribution of their dataset may provide an explanation. In Kim and Baldwin (2006), the majority class covered 42.4% of the data, compared to 20% in our dataset. This makes their classification task significantly simpler than ours: improving accuracy from 42% to 53% using both features and distribution of classes in training data is easier than getting good accuracy using features alone.

As mentioned above, despite these short-comings of our second system, results on the nominalisation identification task improved. It is, however, questionable how reliable these results are. At best, we can conclude that co-occurrences of modifier and head noun within the same sentence provide positive evidence that the relation between the two is semantic rather than grammatical.

Chapter 5

Conclusion

5.1 Summary

This thesis investigated the identification and interpretation of nominalisations. We presented three methods for distinguishing nominalisations from other compounds. Each method was built on top of a nominalisation disambiguator designed for a three-way classification task between underlying subject, object and prepositional complement relations within the nominalisation.

The main goal of our experiments was to design a system that would achieve good recall and precision on nominalisation detection when applied to sets of randomly retrieved compounds from unrestricted text. In addition, we want the system to achieve good performance on the disambiguation task of the identified nominalisations.

Using a database created from Nomlex, Celex and Catvar, our system achieved recall of 94.5% and 99.5% for nominalisations from open sets of 335 and 917 compounds respectively. The three methods for nominalisation identification were directed towards increasing the precision of the system.

The first method used attested occurrences of the modifying noun as an argument of the head noun's root verb as evidence for the compound to be a nominalisation. This method proved to be the most successful of the three increasing accuracy in nominalisation detection to 76% (baseline = 64%) and 70% (baseline = 42%) respectively, for the two open datasets.

The second experiment tested whether selectional restrictions of the root verb could be used to find out whether a compound contained a grammatical relation. The methodology successfully filtered out modifiers that would be "bad" arguments of the root verb, and results improved compared to the basic system, though the improvements were not significant.

In the third experiment, the nominalisation disambiguation system was combined with a second system taking semantic relations into account. Despite the poor performance of this second system, the combination of the two led to better results on the nominalisation disambiguation task. Accuracy on the largest dataset (917 tokens) reached 53%. This is a significant improvement compared to a baseline of 47%. As mentioned in the previous chapter, there is room for improvement in the second system used in this experiment.

The research presented in this thesis is an extension to work by Lapata (2002), Grover et al. (2003) and Nicholson (2005). It improves on the first two studies because it shows good performance on data that is not filtered manually, and is not restricted to a specific domain. Our work forms a significant extension to all three works mentioned above, because it addresses the question of how nominalisations can be identified. Notably the first method is shown to be successful at this task. Taking all four studies together, compound nominalisations can now be detected, identified and interpreted from open text.

5.2 Future Work

Now that all possible subtasks of nominalisation disambiguation have been addressed (detection, identification, classification), the first natural step for future research would be to combine all components of the work carried out by Lapata (2002), Grover et al. (2003) and Nicholson (2005) with our own work.

Lapata's results show that suffixes and direct context can improve the nominalisation disambiguation task significantly. These features could also provide additional indications for the nominalisation identification task. Combining our methods with Nicholson's research would provide a complete nominalisation detection and identification system from open text.

Looking at results from our experiments, the first two methods we presented show optimal performance, considering the resources used. Results from experiment one could possibly be improved by using more data, results from method two could improve when a more extensive and fine-grained resource for selectional restriction would be used.

There is more potential for improvement in the third experiment. In section 4.5.1, we discussed several problems of the general classifier. Despite these problems, we believe that the methodology is suited to be combined with a statistical nominalisation disambiguation system, because of the similar information used by both classifiers. A system using Lapata's methodology and the methodology proposed by Kim and Baldwin (2006) can provide an interpretation of compounds based on relations between verbs and arguments only. The general classifier could be improved in the following ways.

First, our data should be split in a training and a test set. The training data can be used to define more appropriate seed verbs. Using machine learning on these potentially better seed verbs could lead to significant improvement of results. Next to this step, it would be interesting to see how the system performs when lemmas of predicates from the corpus are used directly as features, without the intervention of predefined seed verbs.

In general, we believe that future research on statistical disambiguation of nominalisations lies in the combination of systems. With an eye on applications, a system that can only treat a subset of compounds (between 20-40%) is not complete. It is therefore useful to investigate what the nominalisation disambiguation system can contribute to other classifiers with a larger scope, including other methodologies than the one proposed by Kim and Baldwin.

This last step naturally leads to an extension towards systems aiming at interpreting all compound nouns. We have addressed the controversy about the number and nature of semantic classes for this task at several stages in this thesis. Before our research could be extended to a larger classification task including semantic classes, the question of which classes are necessary needs to be addressed.

Researchers have focused on finding classes that are attested by human annotators. So far, psychological research has not provided an answer stating which classes are essential for the interpretation of compound nouns. The proposal of 14 classes made by Gagné and Shoben (1997) has been shown insufficient by the experiment carried out by Costello et al. (2006), where subjects needed 139 classes to come to a satisfactory interpretation for 1,000 compounds.

It is unlikely that our knowledge on human interpretation of compound nouns will provide answers on the matter in the near future. Moreover, humans find the that labelling compounds according to their best possible interpretation is an unnatural task.[1] Therefore, we believe that the question of necessary classes should be addressed differently in future research. Instead of identifying semantic labels that humans judge best for a given compound, research should focus on labels that are necessary for a specific task. In other words, we believe that compound noun disambiguation should be more oriented towards applications.

One of the questions that should be addressed is what information is missing in compounds that is needed for information extraction or question answering. In machine translation, it would be relevant to know which noun pairs can be expressed by a compound in both languages. If the target language uses syntax or morphology within the compound, what information does this reflect? How can we find this information in a source language that does not express relations overtly?

To our knowledge, none of the studies on compound noun disambiguation address these issues. This is rather surprising, because before making the effort of designing a system for a complex task, the question of what kind of output would be useful is one of the most relevant to ask.

Presumably, it is relevant to know whether a noun functions as the subject, object or a prepositional complement of a verb or deverbal noun. But before

[1] One native speaker who was willing to annotate our data spent 20 minutes trying to find an appropriate interpretation for the first example, and then gave up on the task.

extending this three-fold classification to one containing 14, 20, 60 or 139 classes, the question of relevance should be addressed.

Bibliography

Abney, Steve. 1996. Partial Parsing via finite-state cascades. In John Carroll (ed.), *Workshop on Robust parsing*, ESSLLI, pages 8–15.

Baayen, R.H., Piepenbrock, R. and van Rijn, H. 1995. *The CELEX lexical database, (Release 2) [CD-ROM]*. Philadelphia, PA.: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Baker, Collin F., Fillmore, Charles J. and Lowe, John B. 1998. The Berkeley FrameNet Project. In Christian Boitet and Pete Whitelock (eds.), *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 86–90, San Francisco, California: Morgan Kaufmann Publishers.

Barker, Ken and Szpakowicz, Stan. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

Bauer, Laurie. 1979. On the need for pragmatics in the study of nominal compounding. *Journal of Pragmatics* 3(1), 45–50.

Bergsten, N. 1911. *A Study on Compound Substantives in English*. Uppsala: Almquist and Wiksell.

Brekle, Herbert E. 1970. *Generative Satzsemantik und transformationelle Syntax im System der englischen Nominalkomposition*. München: Fink.

Briscoe, Ted, Carroll, John and Watson, Rebecca. 2006. The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australia.

Brugmann, K. 1900. Über das Wesen der sogennanten Wortzusammensetzungen. Eine sprachpsychologische Studie. In *Berichte über die Verhandlungen der königlichen sächsichen Gesellschaft der Wissenschaften zu Leipzig*, volume 52 of *Philosophische-Historische Class*, pages 359–401.

Buckeridge, Alan M. and Sutcliffe, Richard F. E. 2002. Disambiguating noun compounds with Latent Semantic Indexing. In *Proceedings of the 2nd International Workshop on Computational Terminology*.

Buitelaar, Paul. 1998. *CoreLex: Systematic Polysemy and Underspecification*. Ph. D.thesis, Brandeis University, USA.

Burnage, Gavin. 1990. *CELEX- A Guide for Users*. Nijmegen: Centre for Lexical Information, University of Nijmegen.

Burnard, Lou. 2000. *User Reference Guide to the British National Corpus*. Oxford University Computing Services.

Chomsky, Noam. 1970. Remarks on nominalizations. In Roderick A. Jacobs and Peter S. Rosenbaum (eds.), *Readings in English transformational grammar*, pages 63–130, Waltham, Massachusetts: Ginn and Company.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 37–46.

Copestake, Ann and Lascarides, Alex. 1997. Integrating symbolic and statistical representations: The lexicon-pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*.

Costello, Fintan J., Veale, Tony and Dunne, Simon. 2006. Using WordNet to Automatically Deduce Relations between Words in Noun-Noun Compounds. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, The Association for Computer Linguistics.

Daelemans, Walter, Zavrel, Jakub, van der Sloot, Ko and van den Bosch, Antal. 2004. TiMBL: Tilburg Memory-Based Learner - version 3.0, Reference Guide.

Davidson, Donald. 1967. The Logical Form of Action Semantics. In N. Rescher (ed.), *The Logic of Decision and Action*, pages 81–95, Pittsburg: University of Pittsburg Press.

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language* 53(4), 810–842.

Fellbaum, Christiane (ed.). 1998. *Wordnet: an electronic lexical database*. Cambridge, MA: MIT press.

Fillmore, Charles. 1971. The Semantic Interpretation of Compound Nominals. In D. Steinberg and L. Jakobovits (eds.), *Semantics*, Cambridge University Press.

Finin, Tim. 1980. The semantic interpretation of nominal compounds. In *Proceedings of the 1st National Conference on Artificial Intelligence*.

Gagné, Christina L. and Shoben, Edward J. 1997. Influence of thematic relations on the comprehension of modifier-noun compounds. *Journal of Experimental Psychology: Learning, Memory and Cognition* 23(1), 71–87.

Grover, Claire, Lapata, Mirella and Lascarides, Alex. 2003. A comparison of parsing technologies for the biomedical domain. *Natural Language Engineering* 1(1), 1–38.

Habash, Nizar and Dorr, Bonnie. 2003. A Categorial Variation Database for English. In *Proceedings of the North American Association for Computational Linguistics*, pages 96–102, Edmonton, Canada.

Hindle, David and Rooth, Mats. 1993. Structural ambiguity and lexical relations. *Computational Linguistics* 19(1), 103–120.

Hobbs, Jerry R., Stickel, Mark, Appelt, Douglas and Martin, Paul. 1993. Interpretation as abduction. *Journal of Artificial Intelligence* 63(1–2), 69–142.

Isabelle, Pierre. 1984. Another look at nominal compounds. In *Proceedings of the 10th International Conference on Computational Linguistics*.

Jespersen, Otto. 1942. *A modern English grammar on historical principles*. Copenhagen: Ejnar Munksgaard.

Katz, Slava M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics Speech and Signal Processing* 33(3), 400–401.

Kim, Su Nam and Baldwin, Timothy. 2005. Automatic interpretation of noun compounds using WordNet similarity. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*.

Kim, Su Nam and Baldwin, Timothy. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the ACL-06 Main Conference Poster Session*.

Kingsbury, P. and Palmer, M. 2002. From Treebank to PropBank.

Kipper, Karin, Dang, Hoa Trang and Palmer, Martha. 2000. Class-based construction of a verb lexicon. In *AAAI-2000*, pages 691–696, Austin, Texas.

Koziol, Herbert. 1937. *Handbuch der englischen Wortbildungslehre*. Carl Winter's Universitätsbuchhandlung.

Lakoff, George. 1970. *Irregularity in Syntax*. New York: Holt, Rinehart and Winston.

Lapata, Maria. 2002. The Disambiguation of Nominalisations. *Computational Linguistics* 28(3), 357–388.

Lauer, Mark. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph. D.thesis, Macquarie University, Sydney, Australia.

Lees, Robert B. 1960. The grammar of English nominalizations. *International Journal of American Linguistics* 26.

Leonard, Rosemary. 1984. *The Interpretation of English Noun Sequences on the Computer*. Amsterdam: North-Holland.

Levi, Judith. 1978. *The Syntax and Semantics of Complex Nominals*. New York: Academic Press, INC.

Lowe, Henry and Barnett, Octo. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of American Medical Association (JAMA)* 271(4), 1103–1108.

Macleod, C., Grishman, R., Meyers, A., Barrett, L. and Reeves, R. 1998. NOMLEX: A lexicon of nominalizations. In *EURALEX 1998*.

McCawley, James D. 1975. Review of Noam Chomsky (1972). In *Studies on Semantics in generative grammar*, volume 3 of *Studies in English Linguistics*, pages 209–311, The Hague: Mouton.

McDonald, David. 1982. *Understanding Noun Compounds*. Ph. D.thesis, Carnegie Mellon University.

Moldovan, Dan, Badulesdu, Andrea, Tatu, Marta, Antohe, Daniel and Girju, Roxana. 2004. Semantic Classification of Non-nominalized Noun Phrases. In *Proceedings of HLT/NAACL 2004 - Computational Lexical Semantics workshop*, Boston, MA.

Mosteller, Frederick and Wallace, Davad L. 1964. *Inference and Disputed Authorship: the Federalist*. New York: Springer-Verlag.

Motsch, Wolfgang. 1970. Analyse von Komposita mit zwei nominalen Elementen. In Manfred Bierwisch and Karl Erich Heidolph (eds.), *Progress in linguistics*, volume 43 of *Janua Linguarum, Series Maior*, pages 208–223, The Hague: Moutin.

Nastase, Vivi and Szpakowicz, Stan. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*.

Newmeyer, Frederick J. 1971. The source of derived nominals in English. *Language* 47, 786–796.

Newmeyer, Frederick J. 2005. Some Remarks on Roeper's Remarks on Chomsky's "remarks". *Journal of Theoretical Linguistics* (Special issue).

Nicholson, Jeremy. 2005. *Statistical interpretation of compound nouns*. Masters Thesis, University of Melbourne, Melbourne, Australia.

Nicholson, Jeremy and Baldwin, Timothy. 2006. Interpretation of compound nominalisations using corpus and Web statistics. In *Proceedings of the ACL-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*.

Quinlan, Ross J. 1993. *C4.5 Programs for machine learning*. Series in Machine Learning, San Mateo, CA: Morgan Kaufman.

Roeper, Thomas. 2005. Chomsky's Remarks and the Transformationalist Hypothesis. In R. Lieber and P. Stekauer (eds.), *Handbook of English Word-Formation*, pages 125–146, Dordrecht: Kluwer.

Rosario, Barbara and Hearst, Marti. 2001. Classifying the semantic relations in noun compounds via a Domain-Specific Lexical Hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Ross, John Robert. 1974. Nouniness. In Osamu Fujimura (ed.), *Three dimensions of linguistic theory*, pages 137–257, Tokyo: TEC Corporation.

S. Patwardhan, S. Banerjee, T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*.

Selkirk, Elisabeth. 1982. *The Syntax of Words*. Cambridge, MA: MIT Press.

Siegel, Sidney and Castellan, N. John. 1988. *Non Parametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Spärck Jones, Karen. 1983. Compound noun interpretation problems. In Frank Fallside and William A. Woods (eds.), *Computer Speech Processing*, Englewood Cliffs, NJ: Prentice-Hall.

Talmy, Len. 1985. Force dynamics in language and thought. In *Parasession on Causatives and Agentivity*, University of Chicago. Chicago Linguistics Society (21st Regional Meeting).

Turney, Peter D. 2006. Expressing implicit semantic relations without supervision. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*.

Vanderwende, Lucy. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th International Conference on Computational Linguistics*.

Warren, Beatrice. 1978. *Semantic Patterns of Noun-Noun Compounds*. Göteborg: Acta Universitatis Gothoburgensis.

Wu, Zhibiao and Palmer, Martha. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics*, pages 133–138, Las Cruces NM.

Zholkovskij, A. and Mel'čuk, I. 1972. Sur la synthèse sémantique. *T.A. Informations* 11:2.

Appendix A

Sample Entries of Resources

A.1 Nomlex, Celex and Catvar

A.1.1 Nomlex

```
(NOM :ORTH ”abandonment”
        :PLURAL *NONE*
        :VERB ”abandon”
        :NOM-TYPE ((VERB-NOM))
        :VERB-SUBJ ((DET-POSS)
                         (PP :PVAL (”by”)))
        :SUBJ-ATTRIBUTE ((COMMUNICATOR))
        :VERB-SUBC ((NOM-NP :SUBJECT ((DET-POSS)
                                     (PP :PVAL (”by”)))
        :OBJECT ((DET-POSS)
                                     (PP :PVAL (”of”))))
(NOM-NP-PP :SUBJECT ((DET-POSS)
                                     (PP :PVAL (”by”)))
        :OBJECT ((DET-POSS)
                                     (PP :PVAL (”of”)))
        :PVAL (”for” ”to”))
(NOM-NP-TO-INF-OC :SUBJECT ((DET-POSS)
                         (PP :PVAL (”by”)))
                            :OBJECT ((DET-POSS)
                         (PP :PVAL (”of”)))
                            :NOM-SUBC ((TO-INF :OC T)))
(NOM-NP-AS-NP :SUBJECT ((DET-POSS)
                         (PP :PVAL (”by”)))
                            :OBJECT ((DET-POSS)
                         (PP :PVAL (”of”)))
                            :NOM-SUBC ((AS-NP-PHRASE :OC T))))
:DONE T
:REVISED ”Jan-Update@sapir 17:48 1/13/1999”)
```

A.1.2 Celex

```
11\abandon\35\Z\\1\N\N\N\N\Y\abandon\2\S\N\N\N\\N\N\S\((abandon)[V])[N]\N\N\N

12\abandon\975\M\\1\N\N\N\N\Y\abandon\2\S\N\N\N\\N\N\S\(abandon)[V]\N\N\N

13\abandoned\25\Z\\1\N\N\N\N\Y\abandoned\2\F\N\N\N\\N\N\F\((abandoned)[V])[A]\N\N\N
```

A.1.3 Catvar

```
abdicate_V#abdicator_N#abdicable_AJ#abdication_N
abdomen_N#abdominal_AJ#abdominous_AJ#abdominousness_N
abdominocentesis_N
abdominovesical_AJ
abduce_V#abducent_N#abducent_AJ
abducens_N
abduct_V#abductor_N#abducted_AJ#abduction_N#abducting_AJ
```

A.2 RASP output

A.2.1 Active sentence

```
(1_MC |:_:| |The_AT| |Fulton_NP1| |County_NNL1| |Grand_NP1| |Jury_NN1|
|say+ed_VVD| |Friday_NPD1| |an_AT1| |investigation_NN1| |of_IO| |Atlanta_NP1|
|'s+_$| |recent_JJ| |primary_JJ| |election_NN1| |produce+ed_VVN| |“_”| |no_AT|
|evidence_NN1| |”_”| |that_CST| |any_DD| |irregularity+s_NN2| |take+ed_VVD|
|place_NNL1| ._.) 1 ; (-66.970)

(|ncsubj| |say+ed_VVD| 2_MC _)
(|ncmod| _ |say+ed_VVD| |further_RRR|)
(|iobj| |say+ed_VVD| |in_II|)
(|dobj| |in_II| |presentment+s_NN2|)
(|ncmod| _ |presentment+s_NN2| |term-end_NN1|)
(|ccomp| |that_CST| |presentment+s_NN2| |deserve+s_VVZ|)
(|ncsubj| |deserve+s_VVZ| |Committee_NNJ1| _)
(|ncmod| _ |deserve+s_VVZ| |for_IF|)
(|dobj| |for_IF| |manner_NN1|)
(|det| |manner_NN1| |the_AT|)
(|cmod| _ |manner_NN1| |conduct+ed_VVN|)
(|arg_mod| _ |conduct+ed_VVN| |in_II|)
(|ncsubj| |conduct+ed_VVN| |election_NN1| _)
(|aux| |conduct+ed_VVN| |be+ed_VBDZ|)
(|passive| |conduct+ed_VVN|)
(|det| |election_NN1| |the_AT|)
(|dobj| |in_II| |which_DDQ|)
(|ta| |quote| |ellip| |deserve+s_VVZ|)
(|iobj| |deserve+s_VVZ| |of_IO|)
(|dobj| |deserve+s_VVZ| |and_CC|)
(|dobj| |of_IO| |City_NNL1|)
(|det| |City_NNL1| |the_AT|)
(|iobj| |City_NNL1| |of_IO|)
(|dobj| |of_IO| |Atlanta_NP1|)
(|det| |and_CC| |the_AT|)
(|conj| |and_CC| |praise_NN1|)
(|conj| |and_CC| |thanks_NN2|)
(|ta| |bal| |Committee_NNJ1| |have+ed_VHD|)
(|ncsubj| |have+ed_VHD| |which_DDQ| _)
(|dobj| |have+ed_VHD| |charge_NN1|)
```

```
(|ncmod| _ |charge_NN1| |over-all_JB|)
(|iobj| |charge_NN1| |of_IO|)
(|dobj| |of_IO| |election_NN1|)
(|det| |election_NN1| |the_AT|)
(|det| |Committee_NNJ1| |the_AT|)
(|ncmod| _ |Committee_NNJ1| |City_NNL1|)
(|ncmod| _ |Committee_NNJ1| |Executive_JB|)
(|ta| |colon| 2_MC |jury_NN1|)
(|det| |jury_NN1| |The_AT|)
```

A.2.2 Passive sentence

(2_MC |:_:| |The_AT| |jury_NN1| |further_RRR| |say+ed_VVD| |in_II| |term-end_NN1| |presentment+s_NN2| |that_CST| |the_AT| |City_NNL1| |Executive_JB| |Committee_NNJ1| |,_,| |which_DDQ| |have+ed_VHD| |over-all_JB| |charge_NN1| |of_IO| |the_AT| |election_NN1| |,_,| |“_”| |deserve+s_VVZ| |the_AT| |praise_NN1| |and_CC| |thanks_NN2| |of_IO| |the_AT| |City_NNL1| |of_IO| |Atlanta_NP1| |”_”| |for_IF| |the_AT| |manner_NN1| |in_II| |which_DDQ| |the_AT| |election_NN1| |be+ed_VBDZ| |conduct+ed_VVN| ._.) 1 ; (-78.276)

```
(|xmod| _ 3_MC |consider+ing_VVG|)
(|ta| |quote| |ellip| |consider+ing_VVG|)
(|dobj| |consider+ing_VVG| |and_CC|)
(|conj| |and_CC| |interest_NN1|)
(|conj| |and_CC| |number_NN1|)
(|conj| |and_CC| |size_NN1|)
(|det| |size_NN1| |the_AT|)
(|iobj| |size_NN1| |of_IO|)
(|dobj| |of_IO| |city_NNL1|)
(|det| |city_NNL1| |this_DD1|)
(|det| |number_NN1| |the_AT|)
(|iobj| |number_NN1| |of_IO|)
(|dobj| |of_IO| |voter+s_NN2|)
(|det| |interest_NN1| |the_AT|)
(|ncmod| _ |interest_NN1| |in_II|)
(|dobj| |in_II| |election_NN1|)
(|det| |election_NN1| |the_AT|)
(|ncmod| _ |interest_NN1| |widespread_JJ|)
(|ta| |voc| |say+ed_VVD| 3_MC)
```

```
(|ncsubj| |say+ed_VVD| |jury_NN1| _)
(|det| |jury_NN1| |the_AT|)
(|ta| |colon| 3_MC |charge+ed_VVN|)
(|ncsubj| |charge+ed_VVN| |jury_NN1| _)
(|aux| |charge+ed_VVN| |have+ed_VHD|)
(|aux| |charge+ed_VVN| |be+en_VBN|)
(|passive| |charge+ed_VVN|)
(|ccomp| _ |charge+ed_VVN| |by_II|)
(|ccomp| _ |by_II| |Pye_NP1|)
(|ncsubj| |investigate_VV0| |Pye_NP1| _)
(|xcomp| _ |investigate_VV0| |primary_JJ|)
(|ncsubj| |primary_JJ| |report+s_NN2| _) (|dobj| |investigate_VV0| |report+s_NN2|)
(|ccomp| _ |primary_JJ| |win+ed_VVN|)
(|ncsubj| |win+ed_VVN| |which_DDQ| _)
(|aux| |win+ed_VVN| |be+ed_VBDZ|)
(|passive| |win+ed_VVN|)
(|iobj| |win+ed_VVN| |by_II|)
(|dobj| |by_II| |Mayor-nominate_NN1|)
(|ccomp| _ |Mayor-nominate_NN1| ..._VVD)
(|ta| |quote| ..._VVD |receive+ed_VVN|)
(|ta| |quote| |ellip| |receive+ed_VVN|)
(|ncmod| _ |receive+ed_VVN| |Only_RR|)
(|ncsubj| |receive+ed_VVN| |handful_NN1| _)
(|aux| |receive+ed_VVN| |be+ed_VBDZ|)
(|passive| |receive+ed_VVN|)
(|det| |handful_NN1| |a_AT1|)
(|ncmod| _ |handful_NN1| |relative_JJ|)
(|iobj| |handful_NN1| |of_IO|)
(|dobj| |of_IO| |report+s_NN2|)
(|det| |report+s_NN2| |such_DA|)
(|ncsubj| ..._VVD |Allen_NP1| _)
(|ncmod| _ |Allen_NP1| |Jr_NNSA1|)
(|ncmod| _ |Allen_NP1| |Ivan_NP1|)
(|ncmod| _ |report+s_NN2| |in_II|)
(|dobj| |in_II| |hard-fight+ed_VVN|)
(|det| |hard-fight+ed_VVN| |the_AT|)
(|iobj| |report+s_NN2| |of_IO|)
(|dobj| |of_IO| |irregularity+s_NN2|)
(|ncmod| _ |irregularity+s_NN2| |possible_JJ|)
(|ta| |quote| |ellip| |irregularity+s_NN2|)
(|ncmod| _ |Pye_NP1| |Fulton_NP1|)
```

```
(|ncmod| _ |Pye_NP1| |Superior_JJ|)
(|ncmod| _ |Pye_NP1| |Judge_NNS1|)
(|ncmod| _ |Pye_NP1| |Durwood_NP1|)
(|ncmod| _ |Judge_NNS1| |Court_NNJ1|)
(|det| |jury_NN1| |The_AT|)
(|ncmod| _ |jury_NN1| |September-October_JJ|)
(|ncmod| _ |jury_NN1| |term_NN1|)
```

A.3 VerbNet

A.3.1 Mapping VerbNet to WordNet

Table A.1: VerbNet's selectional restrictions to WordNet

Sel. restriction	WordNet entry	Sel. restriction	WordNet entry
animate	*living thing*	force	*physical phenomenon*
spatial	*location*	int_control	*physical entity*
loc	*location*	dest_dir	*physical entity*
concrete	*physical entity*	src	*vehicle*
abstract	*abstract entity*	scalar	*number*
comestible	*food*	pointy	*instrument*
non rigid	*physical entity*	elongated	*physical entity*
path	*way*	dest_conf	*container*

A.3.2 VerbNet Entry

```
<?xml version="1.0" encoding="UTF-8"?/ >
<!DOCTYPE VNCLASS SYSTEM "vn$_c$lass-3.dtd"/ >
<VNCLASS ID="admit-65" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="vn$_s$chema-3.xsd"/ >
    <MEMBERS/ >
        <MEMBER name="admit" wn=""// >
        <MEMBER name="allow" wn=""// >
        <MEMBER name="include" wn=""// >
        <MEMBER name="permit" wn=""// >
        <MEMBER name="welcome" wn=""// >
    </MEMBERS/ >
    <THEMROLES/ >
        <THEMROLE type=”Agent”/ >
               <SELRESTRS logic=”or”/ >
                    <SELRESTR Value=”+” type=”animate”// >
                    <SELRESTR Value=”+” type=”organization”// >
               </SELRESTRS/ >
        </THEMROLE/ >
        <THEMROLE type=”Theme”/ >
               <SELRESTRS logic=”or”/ >
                    <SELRESTR Value=”+” type=”animate”// >
                    <SELRESTR Value=”+” type=”organization”// >
               </SELRESTRS/ >
        </THEMROLE/ >
        <THEMROLE type=”Location”/ >
        <SELRESTRS// >
    </THEMROLE/ >
</THEMROLES/ >
<FRAMES/ >
    <FRAME/ >
        <DESCRIPTION descriptionNumber=”8.1” primary=”NP”
                    secondary=”” xtag=”0.2”// >
        <EXAMPLES/ >
               <EXAMPLE/ >
She admitted us.</EXAMPLE/ >
        </EXAMPLES/ >
    <SYNTAX/ >
        <NP value=”Agent”/ >
               <SYNRESTRS// >
        </NP/ >
    <VERB// >
    <NP value=”Theme”/ >
        <SYNRESTRS// >
    </NP/ >
</SYNTAX/ >
<SEMANTICS/ >
    <PRED value=”admit”/ >
        <ARGS/ >
               <ARG type=”Event” value=”during(E)”// >
               <ARG type=”ThemRole” value=”Agent”// >
               <ARG type=”ThemRole” value=”Theme”// >
               <ARG type=”ThemRole” value=”?Location”// >
        </ARGS/ >
    </PRED/ >
    </SEMANTICS/ >
</FRAME/ >
<FRAME/ >
    <DESCRIPTION descriptionNumber=”8.1” primary=”NP-ADVP-PRED” secondary=”” xtag=”0.2”// >
```

Appendix B

Results on Nominalisation Data

B.1 Experiment One

Table B.1: Experiment One, Dataset 1 (800 tokens)

Included Evidence	Accuracy	Nom	Acc(Nom)
Mod	54.63%	589	74.19%
Mod, Syn	57.38%	624	73.88%
Mod, Syn, Hypo	58.75%	634	74.13%
$\theta = 0.05$	60.88%	663	73.45%
$\theta = 0.10$	65.12%	710	73.38%
$\theta = 0.15$	67.63%	741	73.01%
$\theta = 0.20$	68.88%	754	73.08%
$\theta = 0.30$	70.50%	778	72.49%
$\theta = 0.40$	70.63%	781	72.43%
$\theta = 0.50$	71.13%	785	72.48%
$\theta = 0.60$	71.25%	787	72.43%
$\theta = 0.70$	71.38%	789	72.37%
$\theta = 0.80$	71.25%	791	72.06%
$\theta = 0.90$	71.12%	791	71.93%
$\theta = 1.00$	71.12%	791	71.93%
MSH & all	72.25%	800	72.25%

B.2 Experiment Two

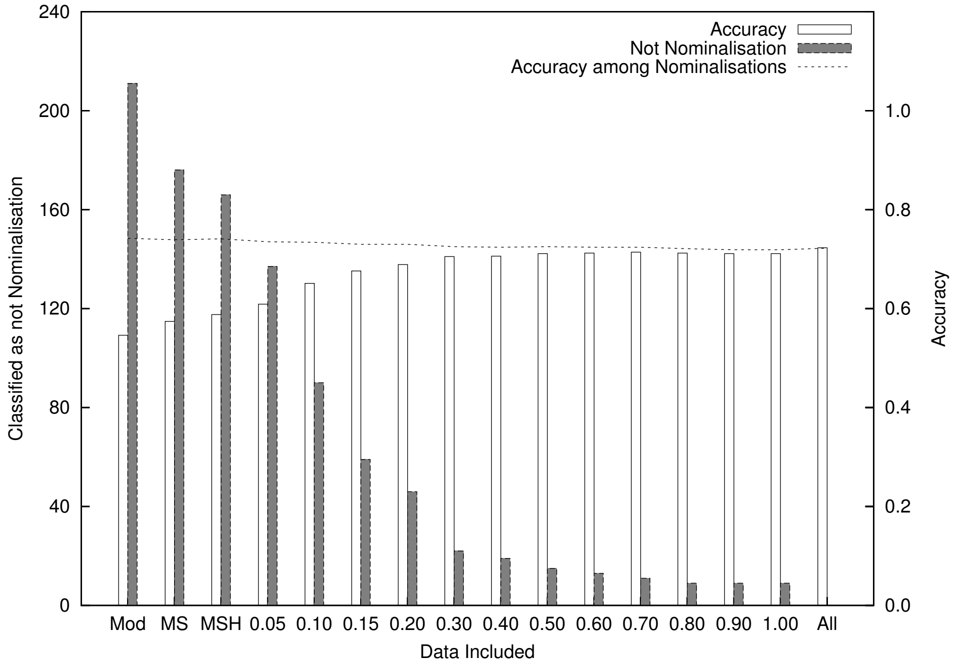

Figure B.1: Results Experiment One, Dataset 1

Table B.2: Experiment One, Dataset 2, Nominalisations (129)

Included Evidence	Accuracy	Nom	Acc(Nom)
Mod	44.2%	88	64.7%
Mod, Syn	44.2%	89	64.0%
Mod, Syn, Hypo	45.0%	92	63.0%
$\theta = 0.00$	46.5%	97	61.9%
$\theta = 0.05$	47.3%	98	62.2%
$\theta = 0.10$	48.8%	102	61.8%
$\theta = 0.15$	50.4%	109	59.6%
$\theta = 0.20$	50.4%	109	59.6%
$\theta = 0.30$	50.4%	112	58.0%
$\theta = 0.40$	52.7%	114	59.6%
$\theta = 0.50$	51.2%	116	56.9%
$\theta = 0.60$	51.9%	116	57.8%
$\theta = 0.70$	52.7%	117	58.1%
$\theta = 0.80$	51.9%	117	57.3%
$\theta = 0.90$	51.9%	117	57.3%
$\theta = 1.00$	51.9%	117	57.3%
MSH & all	54.3%	122	57.4%

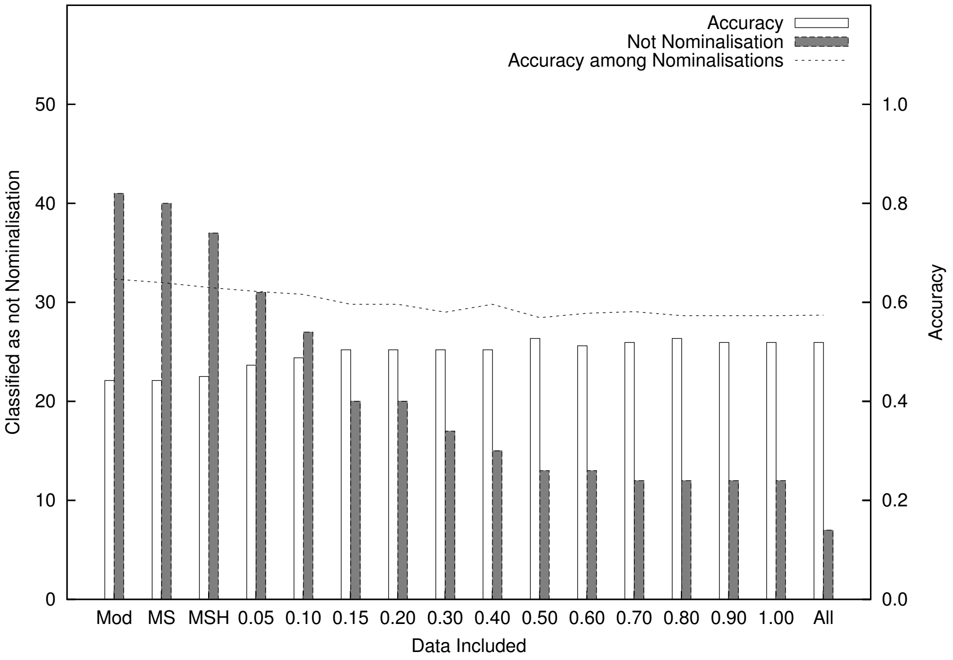

Figure B.2: Results Experiment One, Dataset 2, Nominalisations

Table B.3: Experiment One, Dataset 3, Nominalisations (182)

Included Evidence	Accuracy	Nom	Acc(Nom)
Mod	48.4%	144	61.1%
Mod, Syn	48.9%	145	61.4%
Mod, Syn, Hypo	50.0%	150	60.7%
$\theta = 0.00$	52.7%	158	60.8%
$\theta = 0.05$	52.7%	159	60.4%
$\theta = 0.10$	54.4%	163	60.7%
$\theta = 0.15$	56.6%	168	61.3%
$\theta = 0.20$	56.6%	169	60.9%
$\theta = 0.30$	56.0%	173	59.0%
$\theta = 0.40$	57.1%	174	59.8%
$\theta = 0.50$	57.1%	175	59.4%
$\theta = 0.60$	57.1%	176	59.1%
$\theta = 0.70$	57.1%	177	58.8%
$\theta = 0.80$	57.1%	178	58.4%
$\theta = 0.90$	57.1%	178	58.4%
$\theta = 1.00$	57.1%	178	58.4%
MSH & all	57.7%	181	58.0%

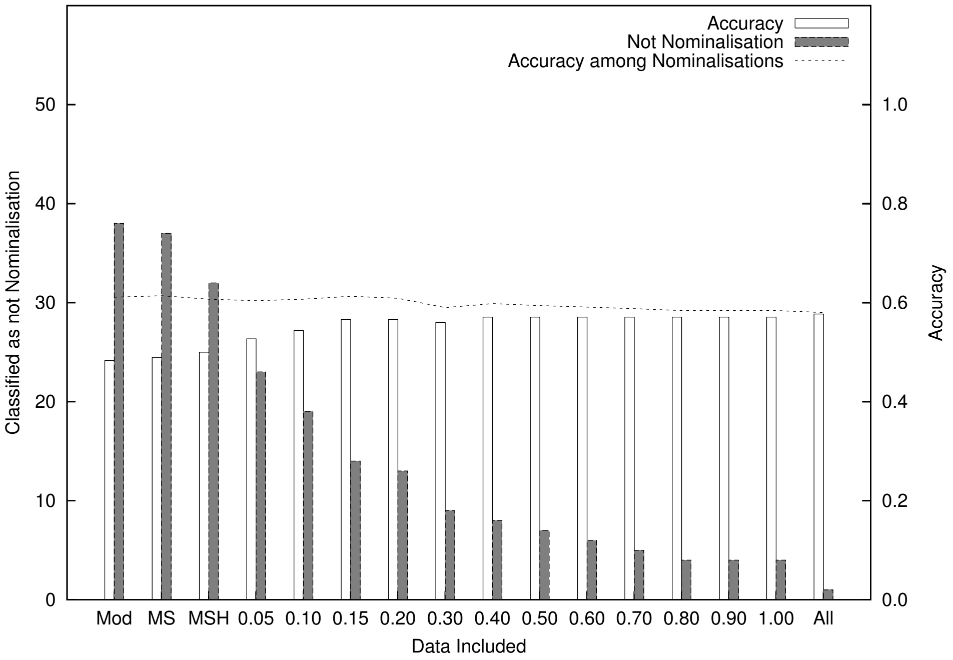

Figure B.3: Results Experiment One, Dataset 3, Nominalisations

Table B.4: Experiment Two, Dataset 1 (800, 684)

	SUBJECT			OBJECT			Acc
Full (800)	*P*	*R*	*F*	*P*	*R*	*F*	
Basic	63.9	57.0	60.3	74.0	79.2	76.5	70.5%
VN-All	59.5	63.7	61.5	75.4	72.0	73.7	68.8%
VN-Id	62.7	67.3	64.6	78.3	74.6	76.4	71.8%
VN (684)							
Basic	66.2	54.5	59.8	74.0	82.3	77.9	71.5%
VN-All	59.9	61.3	60.6	75.0	73.9	74.5	69.0%
VN-Id	61.3	62.9	62.1	76.3	75.0	75.6	70.3%

Table B.5: Experiment Two, Dataset 2, Nominalisations (129, 106)

	SUBJECT			OBJECT			POBJECT			Acc
Full (129)	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
Basic	35.9	63.6	45.9	62.8	75.4	68.5	100	11.9	21.3	52.7%
VN-All	36.6	68.2	47.6	64.8	70.8	67.7	100	11.9	21.3	51.2%
VN-Id	36.6	75.0	49.2	64.8	73.0	68.7	100	14.7	25.6	56.4%
VN (106)										
Basic	33.3	58.3	42.4	63.0	76.7	69.2	100	15.2	26.4	53.8%
VN-All	31.9	63.9	42.6	65.4	69.2	67.2	100	15.2	26.4	49.1%
VN-Id	31.9	71.9	44.2	65.4	71.6	68.3	100	19.0	32.0	55.3%

Table B.6: Experiment Two, Dataset 3 Nominalisations (182, 152)

	SUBJECT			OBJECT			POBJECT			Acc
Full (182)	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	
Basic	36.2	61.5	45.6	64.4	78.4	70.7	92.6	12.5	22.0	58.2%
VN-All	33.2	63.5	43.6	65.5	70.7	68.0	92.6	12.5	22.0	54.4%
VN-Id	33.2	69.5	44.9	65.5	73.2	69.2	92.6	14.9	25.6	56.3%
VN (152)										
Basic	37.5	62.2	46.8	65.5	79.0	71.6	90.3	11.9	21.0	61.2%
VN-All	34.1	63.7	44.4	66.4	71.0	68.6	90.3	11.9	21.1	55.9%
VN-Id	34.1	68.8	45.6	66.4	73.5	69.8	90.3	13.9	24.1	58.2%