UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

DISAMBIGUATING SEMANTIC ROLES IN SWEDISH COMPOUNDS

with Swedish FrameNet and SALDO

Karin Hedberg

Master’s Thesis:	30 credits
Programme:	Master’s Programme in Language Technology
Level:	Advanced level
Semester and year:	Autumn, 2016
Supervisor:	Richard Johansson
Examiner:	Lars Borin
Report number:	-
Keywords:	compounding, Swedish, semantic roles, NLP, SRL, FrameNet, SALDO

Abstract

The compounding of words in Swedish is productive, recursive, and frequent in both text and speech. Compounds can be ambiguous on many levels, and the processing of them involves segmentation, lemma disambiguation, word sense disambiguation, and semantic analysis. In this thesis, we focus on the latter.

We concretise the semantic analysis as semantic role disambiguation, meaning the automatic analysis of the relationship between the two parts of a compound (prefix and suffix) given a set of semantic roles selected by the suffix. The system architecture revolves around lexical resources such as the Swedish FrameNet (SweFN) and SALDO. In two experimental rounds, we train on (1) chunked and semantic role-analysed sentences, and (2) compounds marked up using the frames and semantic roles of SweFN. For instance, *laxröra* 'salmon casserole' is analysed as Constituent_parts+LU (LU=lexical unit) in the Food frame.

The training data of tagged sentences used in predicting compound semantic roles is deemed too sparse, and produces only a small improvement over a most-frequent-class baseline. In our final experiments, we use a narrowed down set of frames and compounds as both train and test data. We reach a best classification accuracy of 62% against a 33% baseline on 100 unseen compounds.

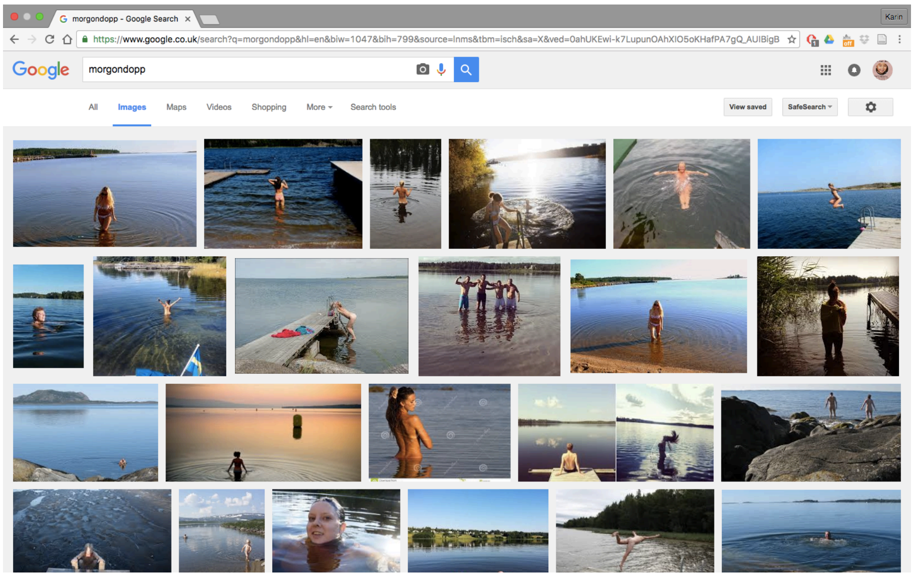
Preface

In addition to my fantastic supervisor, Richard Johansson, I would like to express my gratitude to Lisa Loenheim and Karin Friberg Heppin, both of whom have contributed to this thesis through interesting discussions. The same goes for Anna Ehrlemark, who also helped me with the annotations – Thank you! Finally, I thank Joel Wästberg for standing by my side during and beyond the work on this thesis.

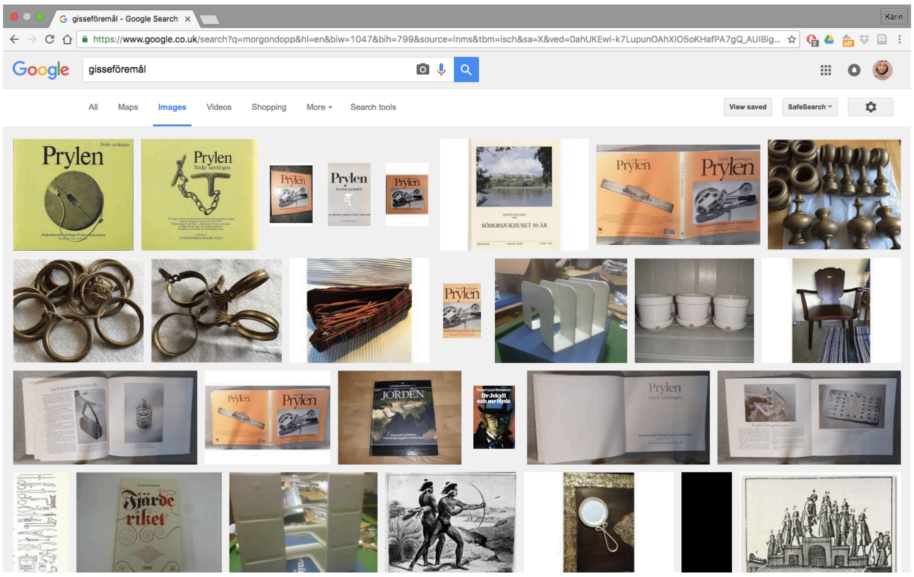
The original software and data sets described in this thesis may be obtained via email: `kkmhedberg@gmail.com`.

Please print in colour.

My favourite compounds are:


morgondopp - Google Search
Karin
https://www.google.co.uk/search?q=morgondopp&hl=en&biw=1047&bih=799&source=lnms&tbm=isch&sa=X&ved=0ahUKEwi-k7LupunOAhXIO5oKHafPA7gQ_AUIBigB
Google
morgondopp
All
Images
Maps
Videos
Shopping
More
Search tools
View saved
SafeSearch

morgondopp ‘morning swim’


gisseföremål - Google Search
Karin
https://www.google.co.uk/search?q=morgondopp&hl=en&biw=1047&bih=799&source=lnms&tbm=isch&sa=X&ved=0ahUKEwi-k7LupunOAhXIO5oKHafPA7gQ_AUIBig...
Google
gisseföremål
All
Maps
Images
Videos
Shopping
More
Search tools
View saved
SafeSearch
Prylen
Tredje samlingen
Prylen
SÖDERSJUKHUSET 50 ÅR
JORDEN
Fjärde riket

gisseföremål ‘artifact up for guessing’

Contents

1 Introduction

The compounding of words in Swedish is a productive, recursive, and frequent linguistic phenomenon. This means that many compounds are non-lexicalised, i.e. they may be produced on the fly – *hemmaladdaren* (home+charger) → 'the charger one keeps at home', and are interpreted by semantic knowledge about the parts, the compositionality principle and by context.

In the automatic processing of compounds, thus, we need to mimic the human interpretation process. Beyond the issues of segmentation – Swedish compounds are written together – and word sense disambiguation, semantic role disambiguation is required in order to make out how the constituents of a compound relate semantically to one another. Compared to semantic role labelling (SRL) at the sentence level, compounds have no internal structural or prepositional information. For instance: while *läderväska* 'leather bag', is a bag made of leather, *skolväska* 'school bag', is not interpreted as a bag **made of** schools, but as a bag whose **use** is to take to school. Therefore, it seems that the problem we need to solve is mainly one of semantic grouping.

The current work describes a Frame Semantics approach to semantic role disambiguation in Swedish compounds. In FrameNet terms (Baker et al., 1998; Ruppenhofer et al., 2010), this means that we consider each (right-headed) compound as fitting into the following structure: FE+LU, where FE stands for frame element, and LU stands for lexical unit, an item pertaining to a certain semantic frame. The task at hand is thus: given the frame that the suffix belongs to, predict the semantic role of the prefix. In the leather/school bag examples above, the frame Containers is given for the suffix, and the roles to be predicted for the prefixes are Material and Use, respectively.

In a first experiment, we consider all of the annotated sentences of the Swedish FrameNet (Borin et al., 2010a,b) extracting features using frame information, part of speech, different levels of semantic closeness by Brown clustering (Brown et al., 1992), the Synlex synonym lexicon (Kann & Rosell, 2005), and the SALDO (Swedish associative lexicon) graph (Borin et al., 2008), (Borin et al., 2013). We then train a Support Vector Machine on these features.

In order to prepare compounds to test our classifier on, we extract non-lexicalised compounds from a literary corpus and a web forum corpus from the Swedish Language Bank, Språkbanken, and set up an annotation interface through which we segment each compound and assign it a frame and a semantic role. Due to the sparsity of annotated sentences per frame in the Swedish FrameNet, we see only a slight improvement over baseline, which has frame as its only feature. This warrants our second experiment, for which we prepare a new data set of annotated compounds only, narrowed down to five frames. Using this set, we acquire a system that soundly outperforms baseline.

1.1 Motivation

In Swedish as well as in many other languages, compounding is a highly productive phenomenon. Compounds have been reported to make up roughly ten percent of content words in Swedish, German and Finnish running text (Hedlund, 2002). This means they must be carefully considered for good performance in any NLP application, for example natural language understanding, text generation, search, and machine translation (Stymne et al., 2013). While the correct segmentation of an orthographically joined compound provides more readily available tokens for e.g. machine translation, mere segmentation is prone to producing errors if there is no understanding of how the internal part of the compound relate semantically to one another. Thus, the internal semantic role disambiguation of Swedish compounds is the topic of the present thesis.

1.2 Contributions

The main contributions of this thesis are:

- Software:
  - Classification model: A classifier that assigns semantic roles to the prefix of compositional, non-lexicalised Swedish compounds in five semantic areas, at 62% average accuracy.
  - Annotation interface: An annotation tool for marking up Swedish compounds within the Swedish FrameNet framework. The annotation interface is available from the author upon request.
- Compound analyses: A set of 2,000+ annotated Swedish compounds, with sentence context, in the Swedish FrameNet FE+LU format, ready to go into the project.
- Statistics: Statistical insights about the distribution of semantic areas (frames) in compounds in two corpora are presented.
- Insights: Although there are some promising indications that Swedish FrameNet-annotated sentences may be successfully applied in learning the semantic roles of compounds, we see the need of more annotated sentences in order to give such an experiment a fair evaluation. Among our more enjoyable insights is the fact that our model for compound semantic role disambiguation trained on compounds almost doubles the accuracy of our baseline, with a relatively small training set.

1.3 Research questions

When tackling the issue of semantic role disambiguation in Swedish compounds, we need a framework and an infrastructure. We find an interesting approach in using (1) Frame Semantics annotated sentences and (2) the compound markup scheme of the Swedish FrameNet. This poses two research questions:

- Can we apply Frame Semantic theory in semantic role disambiguation in Swedish compounds?
- Are FrameNet annotations for sentences transferrable, i.e. useful in the semantic role classification of compounds?

1.4 Delimitations

The main problem we address in this thesis is the one of semantic role disambiguation in Swedish compounds. Related problems, such as segmenting and word sense disambiguation, are discussed. However, any attempts at improving existing technology in these areas are beyond the scope of this thesis.

1.5 Terminology and naming conventions

This section is for the reader's reference in case of any uncertainty regarding terminology and concepts.

In the text, the terms **prefix** and **FE** (frame element) are used interchangeably about the left constituent of a compound.

The term FE is sometimes also used to refer to semantic role in FrameNet contexts. We try to avoid confusion in this matter. The right constituent of a compound is referred to as the **suffix** or **LU** (lexical unit, in FrameNet terms).

Semantic role refers to the relation of the prefix with regards to the suffix of a compound. Since the semantic roles are also the class labels in our experiments, **class**, **label**, and (semantic) **role** all refer to the same concept in the experimental setting.

Swedish FrameNet and **SweFN** are used interchangeably.

BFN is short for the **Berkeley FrameNet**, which is the first and English-language adaptation of Frame Semantics.

We codify FrameNet and SweFN examples by font and case. Semantic frames are written in small caps: BRAGGING, and semantic roles are capitalised: Addressee.

SRL – semantic role labelling, is an established term used for the computational processing of running text. Semantic role **disambiguation** refers to the general idea, and may apply to both running text and compounds.

2 Background

In this section, we first present an overview of Swedish compounding. We then review linguistic theoretic work on (Swedish) compounds and automatic semantic role labelling (SRL). To this we add also the semantic framework in which we situate our disambiguation system: frame semantics.

To the best of our knowledge, the present thesis is original in its combination of problem formulation (SRL within compounds), language (Swedish), and approach (frame semantics). However, while previous work on the automatic semantic role disambiguation of Swedish compounds is scarce, semantic role disambiguation and the structure of compounds have been extensively explored as separate topics, as well as compounds in other languages.

2.1 Definition of compounds

First something about our definition of compounds. What we mean by the term compound in the context of this thesis is: **a linguistic concept expressed by two lexical items in conjunction, orthographically represented without an intervening space**. However, compounding may be recursive, i.e. consisting of more than two lexical units. Although either of the lexical items that make up a compound may be a compound in and of itself, we consider it as consisting of exactly two parts at surface level: a prefix and a suffix. Considering for instance the recursive compound *diskmedelsbubblor* (washing up liquid+bubbles) ‘bubbles caused by washing up liquid’, we only deal with the primary segmentation point despite the fact that the prefix, *diskmedel* may be further decomposed into diska+medel (washing up+detergent) (see Figure 1).

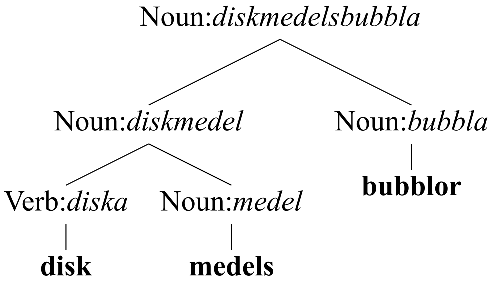


Figure 1: Recursive compounding.

2.2 Joining and segmenting

Swedish compounds are orthographically signified by being written together. In the process of joining the lexemes making up a compound, word-final vowels in the prefix may be omitted and the *s*-interfix may be inserted, especially after a long prefix or to mark the primary segmentation point in recursive compounding. This interfix may also less frequently be a vowel. A hyphen is used in conjunction with acronyms and where the prefix is a clause. Each joining process type is exemplified in Table 1.

The automatic segmentation of Swedish compounds is a well-researched subject as it forms the base for any further analysis of compounds, such as the present thesis topic. It is treated in Friberg (2007) by way of memory-based learning of possible character clusters, and in Sjöbergh & Kann (2004) by statistical hybrid methods involving word lists, POS information, and character n-grams. In our experiments, we use automatically pre-segmented data.

Prefix	Suffix	Process	Compound	Translation
hus ‘house’	tak ‘roof’	∅	*hustak*	‘roof of a house’
skriva ‘write’	kunnig ‘able’	a → ∅	*skrivkunnig*	‘able to write’
fotboll ‘football’	lag ‘team’	∅ → **s**	*fotbollslag*	‘football team’
barn ‘child’	tro ‘faith’	∅ → **a**	*barnatro*	‘childhood faith’
hälsa ‘health’	vård ‘care’	a → **o**	*hälsovård*	‘health care’
flicka ‘girl’	barn ‘child’	a → **e**	*flickebarn*	‘little girl’ (literary)
gata ‘street’	upplopp ‘riot’	a → **u**	*gatuupplopp*	‘riot’
TV ‘TV’	apparat ‘device’	∅ → **-**	*TV-apparat*	‘television set’
styr och ställ ‘steer and put’	cykel ‘bike’	∅ → **-**	*styr och ställ-cykel*	‘Gothenburg rental bike’

Table 1: Types of joining processes in Swedish compounds.

Often, the segmentation point in a compound is ambiguous: *glasskål* may be correctly segmented into the equally likely *glas+skål* (glass+bowl) or *glass+skål* (ice cream+bowl), and the less likely *glass+kål* (ice cream+cabbage). In a similar way, *fotbollslag* ‘football team’ has two structurally possible readings: the intended *fotboll+s+lag* (football+[interfix]+team) and the unlikely *fotboll+slag* (football+beat). In these examples, lemma disambiguation is required before further analysis is possible.

2.3 Lemma and word sense

Between the steps of segmentation and semantic relation disambiguation of Swedish compounds is the issue of determining which lemma (base form of a conjugation pattern) that each segment belongs to.

Consider the following example from Östling (2010): *mossflora* ‘flora of mosses’. Although the segmentation point is unambiguous (moss+flora), the prefix must be disambiguated into the correct lemma: *mossa* ‘moss (plant)’ or *mosse* ‘bog, wetland’. Furthermore, the suffix *flora* requires word sense disambiguation. The SALDO lexicon, which we will describe in greater detail in a later section, has three senses for the lemma in question: *flora*^{1} ‘plant life’, *flora*^{2} ‘catalogue of plant life’, and *flora*^{3} ‘collection’ (Borin et al., 2008). In other words, the processing of one compound may contain several instances of disambiguation. In the *mossflora* case, lemma disambiguation is required for the prefix, word sense disambiguation (WSD) in the suffix, and semantic role disambiguation regarding the entire compound.

Although a model for Swedish WSD that significantly outperforms a first sense baseline is yet to be seen, there is still hope. Johansson et al. (2016) present annotation work for Swedish word senses at the sentence level. The authors incorporate a structure for sense-marking not only simplex words, but also the internal parts of compounds, which may be of use to future improvements of the system described in this thesis.

2.4 Subtypes of compounds

While the majority of Swedish compounds belong to the noun+noun type, most combinations of parts of speech are possible, including but not limited to noun+verb *vitlöksmarinera* ‘garlic-marinate’, verb+noun *spränggranat* (explode+grenade) ‘explosive shell’, and noun+adjective *nervsjuk* (nerve+sick) ‘neurotic’. Even proper nouns may occur as modifiers in Swedish compounds, e.g. *Palmemordet* (Palme+murder) ‘the murder of (Olof) Palme’ (Koptjevskaja-Tamm, 2009).

Before we proceed, it is necessary to make a few distinctions between different types of compounds. For the purposes of this thesis, two discriminative concepts are particularly important: Compositionality and Lexicality.

In the context of compounds, the notion of **Compositionality**, commonly attributed to Frege (1884), refers to the degree to which the parts of a compound mean the same as when the parts stand alone. **Lexicalisation** refers to the process that words undergo once they are common enough to be interpreted as a specific concept. This may of course vary between speakers. For example, *sommarställe* 'summer place' may literally be interpreted as some location one goes to in the summer, however for most speakers (and lexicons), it has become established into the specific meaning of a (usually owned) house for stays primarily during the summer. Indeed, there is no ground truth as to when a compound goes from being non-lexicalised to being lexicalised, as a compound may be considered as a specific concept by some speakers, while not by others. Our definition of lexicality, for simplicity, is whether the full compound has its own entry in the SALDO lexicon.

In Table 2, we show how the two concepts of lexicality and compositionality interact. Naturally, the combination of non-compositional and non-lexicalised compound does not exist, as such a word would be unintelligible.

	Compositional	**Non-compositional**
Lexicalised	*sommarställe* 'summer place'	*jordgubbe* (earth+chap) 'strawberry'
Non-lexicalised	*torsdagsträning* 'Thursday workout'	∅

Table 2: The intersection between compositionality and lexicality.

Libben (1998) makes a case for the differences in the linguistic processing of compounds of different levels of **transparency**. While we will not go into discussing Libben's psycholinguistic evidence for English, his use of the terms transparency and opacity are applicable to Swedish, too. Libben et al. (2003) define **transparent** compounds as

> [words in which] the meaning of the entire string can be derived from the combination of the meanings of its constituents

which is parallel to the notion of compositionality. **Opaque** compounds are those in which none of the constituents represents its lexical meaning, e.g. *humbug*, and **partially transparent** compounds are those in which either the first or the second constituent are transparent, and the other opaque, such as 'chopstick' and 'shoehorn'. In a Swedish context, partially transparent or semi-compositional compounds are words like *krokodiltårar* 'crocodile tears' and *tranbär* 'cranberry', while fully opaque compounds may be exemplified by *nyckelpiga* (key+maid) 'ladybug'.

Regarding noun-final compounds, Teleman (1972) makes type distinctions between:

(a) Determinative compounds, which has significant overlap with the notion of compositionality and in which the meaning of the compound is a subset of the meaning of the semantic head, which is generally considered to be the suffix. Example: *torkhandduk* (dry+towel) 'towel for drying dishes'.

(b) Bahuvrihi compounds (also known as exocentric or possessive), in which the referent of the compound is found outside of it, such as *dumskalle* (stupid+head) 'idiot', which is not a subtype of heads, but descriptive of the person who possesses the head.

(c) Copulative compounds, in which both prefix and suffix are of equal weight (and both inflected): *prinsen-regenten* (the prince+the regent) 'the person that is both prince and regent'.

(d) Imperative compounds – a small group of strongly lexicalised compounds: *förgätmigej* ‘forget-me-not’.

For the purposes of this thesis, we focus on compositional, non-lexicalised compounds, i.e. those analogue with ‘Thursday workout’ in Table 2. The reason for this decision is that the meaning of non-lexicalised, infrequent compounds are hardly available through a search query, and therefore especially interesting from an NLP perspective.

2.5 Compounds in other languages

Compounding is productive not only in Swedish, but in many languages. The other Scandinavian languages, for instance, as well as German and Finnish, all share the feature of orthographic joining. Several eye-movement studies have been carried out with regards to (long) compounds, see e.g. Pollatsek et al. (2011) for an interesting study on the effect of lexicality versus novelty in the processing of compounds in Finnish. In this section we will look closer at two works involving English compounds. There is a tendency in English to join non-compositional and lexicalised compounds: *greenhouse*, *blackmail*, *killjoy* and to separate compositional compounds, (Cf. *summer house*, *boy band*). Swedish does not make this orthographic distinction, making the automatic processing of compounds different between the languages.

Although English compounds are different to the Swedish in their orthographic representation, there are conceptual similarities. In their corpus study of semantic patterns in compounding, Maguire et al. (2010) report the distribution of noun type combinations with regards to semantic groups in the British National Corpus. Out of 25 semantic classes, the three most common combinations were artifact+artifact ‘bicycle shed’, person+person ‘peasant soldier’, and artifact+act ‘guitar tuning’. Among their findings was also the interesting notion that compounds made up of two semantically related lexemes, such as plant+plant ‘elm tree’ and substance+substance ‘lithium metal’ are more likely to occur than expected with regards to the general frequency of that semantic group in the corpus.

Many attempts have been made to formalise the relation patterns between compound constituents. In her work on the syntax and semantics of complex nominals, (Levi, 1978, p. 280) presents a record of relations in English compounds: e.g. the bilateral Cause$_1$: ‘tear gas’ (the second causes the first), and Cause$_2$: ‘birth pains’ (the first causes the second).

In the next section, we provide an overview of attempts at doing the same for Swedish.

2.6 Semantic relations within Swedish compounds

Many attempts have been made at mapping the general types of semantic relations within Swedish compounds. The process underlying compounding may be described as the partial deletion of a clause. Notably, the aforementioned Teleman (1972) identifies 21 clause-type semantic relation types among determinative noun-final compounds, as well as 10 noun-initial compound relation types such as Time, Reason, Manner, and Location. In (Teleman et al., 2010, p. 44-5), a more readable list of noun+noun compound relation types is presented, including Material, Contents, Owner, Source, Place, Part, Whole, etc.

Järborg (2003) defines 14 groups (with subtypes) of semantic relations in noun+noun and noun+adjective compounds as types of constructions. In Järborg (2003), *klassrumsundervisning* ‘classroom education’ is formalised as belonging to the group ‘**Y** carried out **LOC X**’, where **X** refers to the prefix, **Y** is the suffix, and **LOC** is a placeholder for a preposition: ‘**education** carried out **in** (a) **classroom**‘. While reminiscent of the formalisations in Teleman (1972) and Teleman et al. (2010), Järborg's definitions are more readable, yet rather cumbersome to overview.

Headedness is an interesting discussion when it comes to the meaning of compounds. While Swedish compounds are generally considered syntactically right-headed, the semantic headedness does not always follow the syntax. For instance: is *vattendroppe* 'water drop' more saliently a **drop** that happens to consist of water (rather than some other substance, like milk), or is it some amount of **water** that has the characteristic of being the shape of a drop? Following Ruppenhofer et al. (2010), we take the stance of strict right-headedness in our analysis of compounds as it fits our FrameNet method, which will be described in later sections.

2.7 Frame semantics

Frame semantics is a theory of meaning in language proposed by Fillmore (1976). Frame semantics is based on the notion of **frame**, analogue with schema and scenario, as a cognitive architecture for mapping the meaning of, and relationships between, linguistic concepts. As an example of what frame entails, let us take a look at the widely used type example of a semantic frame: a **commerce scenario**. It is argued that we as linguistic beings conceptualise the participants i.e. the necessary or optional concepts involved in such a scenario, in relation to each other. In other words, we store linguistic units or constructions such as **buyer**, **seller**, **goods**, and **money** in relation to the commerce scenario frame as well as to one another.

Frame semantics theory has been put into practice in the Berkeley FrameNet project (BFN) as a lexical-semantic digital resource for language technology research (Baker et al., 1998; Fillmore et al., 2002; Ruppenhofer et al., 2010), which has inspired many other FrameNet projects, including those of Japanese, Brazilian Portuguese, and Swedish.

In FrameNet formalisation a target word, also called a lexical unit (LU), evokes a semantic frame. The frame has a finite set of frame elements (FEs) which relate to the predicate LU in various ways. The FEs come as two subtypes: core and peripheral. Let us look at the formalisation of the aforementioned commerce scenario. The definition of the COMMERCE_SCENARIO frame[1] in BFN begins as follows:

> Commerce is a situation in which a Buyer and a Seller have agreed upon an exchange of Money and Goods (possibly after a negotiation), and then perform the exchange, optionally carrying it out with various kinds of direct payment or financing or the giving of change. The Seller indicates their willingness to give the Goods in their possession to a Buyer who would give them some amount of Money. [...]

In the definition, FEs are represented by capitalisation: Buyer, Seller, Money, and Goods. These are the core FEs. The peripheral FEs in this frame are Manner, Means, Purpose, Rate, and Unit, and these are also capitalised (for the purposes of our experiments, no difference is made between core and peripheral FEs). Each FE has a definition within the frame. See the definition of Seller:

> The **Seller** has the Goods and wants the Money.

along with a tagged example, in which the LU is in all upper-case letters:

[1] `https://framenet2.icsi.berkeley.edu/fnReports/data/frame/Commerce_scenario.xml` (last accessed 8 September 2016)

[My local grocery store]$_{\text{Seller}}$ raised PRICES on meat

The frames range from abstract (e.g. Idiosyncrasy with LUs such as *unique* and *peculiar*) to concrete (Death: *die*, *starve*, *perish*). Though most frames revolve around scenarios, states, or events, there are also frames that treat physical objects, such as Accoutrements: *hat*, *ring*, *anklet*. We shall return to applied Frame semantics as used in this thesis, in future sections.

2.8 Semantic role labelling

Semantic roles, also called thematic roles, is a device used in semantic functional analysis of relationships between linguistic constituents, usually a relationship between a predicate and its arguments. The concept has been revisited by many researchers (Longacre, 1983; Larson, 1984).

Current literature propose different size sets of semantic roles, however they usually revolve around roles like Agent ('*Hannah* drove the car'); Patient ('the dog ate *the meat*'); Recipient ('hand *me* the phone, will you?'); Experiencer ('*He* suffered'); Source ('I took it *from Longacre*'); Goal ('let's go *to school*'); Path ('she walked *along the canal'*); Instrument ('They broke the window *with a hammer*') etc.

In Frame semantics, which we discussed in Section 2.7, the view is that the situation defines the set of semantic roles. In FrameNets, thus, each frame or scenario has its separate set of semantic roles, nevertheless with some overlap between frames.

Semantic role analysis differs from grammatical or functional analysis such that the semantic role of an argument does not change with voice:

Syntactic function: The cat caught [the mouse]$_{\text{Object}}$
[The mouse]$_{\text{Subject}}$ was caught by the cat

Semantic role: The cat caught [the mouse]$_{\text{Victim}}$
[The mouse]$_{\text{Victim}}$ was caught by the cat

Semantic role *labelling* (SRL), a well-researched NLP topic, is the process of automatically assigning semantic roles to (chunks of) words, given a (parsed) sentence and a predicate:

[Daisy]$_{\text{Eater}}$ **ate** [mum's flowers]$_{\text{Eaten}}$

In building an automatic SRL system, researchers are simultaneously confronted with (1) which (size) set of semantic roles to use, and (2) how to treat and evaluate the chunking of sentence elements. The latter involves deciding how strictly or leniently to treat the boundaries around the frame element – when predicting *mum's flower* in the recent example, which is larger than a single token, the evaluation measure has to handle whether to consider a subpart of the chunk, e.g. *flower* as completely incorrect or partially correct.

Gildea & Jurafsky (2002) chose the FrameNet infrastructure for the first problem, and a precision and recall measure for the second. The FrameNet corpus at the time consisted of 50,000 sentences. They hand-crafted a set of 18 abstract roles (Agent, Patient, Path, Result etc.) and achieved a 82% classification accuracy on presegmented constituents, and 65%/61% precision/recall on the system that first chunked the sentences, and then classified the constituents.

A FrameNet-based SRL system was developed by Johansson & Nugues (2006) prior to the conception of a Swedish FrameNet. Instead, for training, they automatically translated English FrameNet-annotated sentences into Swedish using parallel corpora. They reported 75% labelling accuracy for manually pre-chunked frame elements (baseline 41%), and 67%/47% precision and recall on the complete task.

Later, an experiment similar to the one described in Gildea & Jurafsky (2002) was carried out with the Swedish FrameNet by Johansson et al. (2012), but with a much smaller data set (∼3,000 sentences) and with dependency parsing rather than phrase structure parsing. They evaluated the chunking or bracketing task separately from the SRL task on presegmented frame elements, and acquired 71%/65% precision/recall for the bracketing task, and 64% accuracy for the labelling task (baseline 30%). Among the findings in Johansson et al. (2012) is the fact that treating class labels as frame-specific, even if they share the same name e.g. Time in SELF_MOTION and COMMERCE_BUY, impacts the accuracy significantly and negatively.

With regard to features in sentence-level SRL systems, there are a few recurring ones in the literature. Common features include: the **Predicate** itself, the syntactic **Path** from the constituent under scrutiny to the predicate, the **Phrase Type** of the constituent to be labelled e.g. PP or NP, the **Position** of the word or words to be classified in relation to the predicate (before or after), the **Voice** of the predicate (if a verb), the **Head Word** of the constituent in question, and the **Sub-categorisation** of the predicate verb (Xue & Palmer, 2004). Note how these features are mainly syntactic – this is expected as SRL in the traditional sense aims to create semantic relations between constituents and a predicate in a sentence. Compounds, on the other hand, may be viewed as inverted paraphrases stripped off most of their syntactic and structural information:

~~A~~ **scarf** ~~made of~~ **silk** → *silk scarf*

This means that the standard feature set for SRL is non-applicable to semantic role disambiguation in compounds. When it comes to such a task, thus, it seems we have to approach things differently.

2.9 Automatic-semantic compound analysis

Friberg Heppin & Petruck (2014) propose an encoding scheme for the labelling of compounds in the Swedish FrameNet, in which the suffix of a compound is treated as a frame-evoking lexical unit, and to tag its modifier with the appropriate frame element. In other words, the compound is seen as a sort of microenvironment, equivalent to a phrase or a clause. For instance: the suffix of the compound *råttgift* 'rat poison' – *gift* 'poison evokes the frame TOXIC_SUBSTANCE. The prefix is then analysed as representing one of the frame elements pertaining to this frame – the core FEs Toxic_substance, Toxin_source, and Victim, and the peripheral FEs Body_part, Circumstances, Degree, Duration, Reason, Type. The compound is thus formalised as:

råttgift (TOXIC_SUBSTANCE) – Victim+LU.

As future directions for research, Friberg Heppin & Petruck (2014) mention the topic of the present thesis, i.e. automatic semantic role labelling/disambiguation of Swedish compounds using their proposed encoding scheme. With the FrameNet compound structure serving as a framework, we are able to build such a system.

To the best of our knowledge, there is no previous work on semantic role disambiguation in Swedish compounds. A, few comparable works have been carried out for English, however none with FrameNet.

Rosario & Hearst (2001) use 18 semantic relation classes in their experiment with English noun compounds from the medical domain. They use hierarchical lexical resources for the specified semantic group that is their focus of interest, i.e. the medical domain, and acquire accuracies around 60%. They also experiment with softer evaluation measures, and report up to 78% accuracy in listing the correct relation among the top three prediction hypotheses. Rosario & Hearst (2001) compare their results to two other works for English: Vanderwende (1994) with 52% accuracy in a relation disambiguation task with 13 classes, and Lapata (2000) with 82% in a binary classification task.

After this background chapter, we move on to describing our scientific method and system design.

3 Method

The problem we address in this thesis is how to automate the human intuition about what a novel compound means. Consider for instance a situation in which a speaker of Swedish has never heard or read the word *boskapsauktion* 'livestock auction', but she or he is familiar with its components, *boskap* 'livestock' and *auktion* 'auction', from contexts other than the compound. The Swedish speaker will have little to no problem processing and interpreting the compound correctly as a commercial event in which livestock act as the **goods** to be sold and bought. A machine, on the other hand, needs help in learning what kind of roles that livestock may occupy in a given situation. Could a livestock auction be considered an auction for livestock to buy and sell their goods? Although this interpretation is entirely possible grammatically, it is semantically odd and so implausible that it hardly even crosses the mind of the human speaker of Swedish.

We have already discussed SRL on the sentence level. Although the problem we face is related to traditional SRL, it differs such that there are no syntactic, positional, or prepositional indicators in compounds – they are the same no matter in what context they appear. In a sentence such as 'He sold the livestock at an auction', the syntactic object relation between 'livestock' and 'sell' is a strong indicator of 'livestock' occupying the Goods role. When disambiguating compounds, we have no access to this type of information. Indeed, the opposite side of the coin is that we do not have the chunking issue faced in sentence-level SRL systems. Nevertheless, it does mean we must teach the machine something about the meaning of the words that make up the compound in order for it to correctly predict the semantic relation between them. Therefore, we make use of a range of semantic resources: the SALDO lexicon (Borin et al., 2008, 2013) with its semantic descriptors, Brown clusters (Brown et al., 1992), and the Synlex synonym lexicon (Kann & Rosell, 2005).

In this chapter, we describe the procedure of building and evaluating a classification system for disambiguating semantic roles in Swedish compounds. In order to build such a system, we need data to train and test our classifier on, and in those cases where the data is not desirably structured or labelled, we do so ourselves. We also describe the application of lexical resources and a machine learning algorithm.

3.1 System design

The list below describes the system implementation step by step. More details about each component follows in later sections.

1. Defining the task: The classification task is defined as follows: Given the semantic frame evoked by the suffix of a compound, predict the semantic role of its prefix. In practice: given the compound *laxrosa* 'salmon pink' and the frame Color, predict the role Comparand out of the set [Attribute, Color, Entity, Cause, Color_qualifier, Comparand, Degree, Descriptor, Sub-region, Type].

2. Defining sub-experiments: We use two different types of training data. Henceforth, the **first experiment** refers to the experiment in which annotated example sentences from SweFN make up the training data. The **second experiment** refers to the experiment(s) in which we use annotated compounds for both training and test.

3. Structuring SweFN: For the first experiment, we take all of the marked-up sentences of SweFN (see section 3.4.1) and process them such that the following sentence from the Clothing frame:

   [Jag]$_{\text{Wearer}}$ hade en [glansig]$_{\text{Descriptor}}$ [klänning]$_{\text{LU}}$ [från YSL]$_{\text{Creator}}$ på mig
   [I]$_{\text{Wearer}}$ had a [shiny]$_{\text{Descriptor}}$ [dress]$_{\text{LU}}$ [from YSL]$_{\text{Creator}}$ on me
   'I wore a shiny dress from YSL'

generates three training instances:

Frame	LU (lemma)	LU POS	FE lemma	FE POS	Role
Clothing	klänning	NN	jag	PN	Wearer
Clothing	klänning	NN	glansig	AV	Descriptor
Clothing	klänning	NN	YSL	PM	Creator

In order to strip the train data of the most common unnecessary words, we make use of the POS-information provided by the sentences in the SweFN file, which are also dependency parsed. For maximal recall, we add each non-function word in a multiword chunk as a new training instance. E.g. in the third and last training instance above, the tagged frame element consists of more than a single word, [from YSL]. Only the content word of the bracketed chunk is made into a train instance. As we can see, YSL+klänning (YSL+dress) makes for a more plausible candidate for a compound than från+klänning (from+dress).

4. Abstraction: As each frame has its own set of roles, inheritance links for the Berkeley FrameNet are used to obtain abstract representations of frames and roles. For example, the Clothing frame inherits from Artifact, which in turn inherits from Entity, and the role Style inherits from the role Type (Matsubayashi et al., 2009). In the featurisation of the train/test instances, both the specific and the most abstract frames are used as features. The abstract role, if present, is used as the classification label.

5. Feature selection: In the search for relevant patterns between compounds of the same semantic role type, we experiment with features based on lemma, part of speech (POS), frame, abstract frame, SALDO entries and primary descriptors, synonyms from the Synlex synonym lexicon, Brown semantic clusters for lemmas and their Synlex synonyms, and in the second experiment, sentence context features.

6. Training a classifier: A machine learning algorithm is applied to learn from the feature vectors the patterns of what types of prefixes belong to what semantic roles. We train several different models using this algorithm.

7. Prepare test data: In preparing the test instances for the first experiment and the entire data set for the second experiment, we start from existing compounds rather than from sentences. By segmenting and marking up compounds with frame and semantic role, we produce a data set of the same format as the sentence based train data. A detailed account of the annotation procedure follows in a later section.
   Example of a test instance from the frame People_by_vocation – *kökspiga* 'kitchen maid':

Frame	LU (lemma)	LU POS	FE (lemma)	FE POS	Role
People_by_voc.	piga	NN	kök	NN	Place_of_employm.

8. Evaluation: In the first experiment, the model is evaluated using an average accuracy measure: the number of correct predictions divided by the total number of predictions. In the second experiment, the SweFN annotated sentences are no longer used, but instead we use our hand-prepared data set of around 1,000 instances containing five frames only. Because of the modest size of this data set, we shuffle and divide all but 100 instances of the data ten times into 90%+10% folds, assuring that within each fold, no unseen classification label is put in the 10%. This will henceforth be referred to as the **development set**. We report the 10-fold cross-validation average accuracy for each experiment in the development set, and the remaining 100 compounds are used as the final **test set**.

3.2 Data

3.2.1 Train data

The train data for the **first experiment** consists of the entirety of the 8,500+ SweFN example sentences annotated for frame, lexical unit(s), and frame elements. The material is also dependency parsed, lemmatised, and tagged with part of speech. A few sentences also embed compounds analysed in the FE+LU format, however the number is negligible. The method described in Section 3.1, step 3, renders a body of training instances just shy of 20,000. Although this is not small in absolute terms, the average number of sentences per frame becomes slight as there are 1003 frames, i.e. on average 8 sentences or 20 training instances per frame. A handful of frames are represented by zero or one training instance. Considering that there is only a certain degree of overlap in semantic roles (FEs) between frames, there is also an abundance of class labels which entails in no way optimal experimental circumstances.

The train data for the **second experiment** is of the same type as the test data in the first experiment, i.e. annotated compounds rather than test instances generated automatically from sentences. In the next section, we provide details about the corpora used and our selection requirements for compounds.

3.2.2 Test data

For test data in the first experiment, we prepare a set of labelled non-lexicalised, compositional compounds. What follows in this section holds for both train and test in the second experiment.

We use two corpora from Språkbanken, the Swedish Language Bank[2]: (1) a collection of 23 novels published by Norstedts publishing company in 1999 (hereafter 'the literary corpus') and (2) the fathers' section of the family themed web forum *Familjeliv* from 2004-2014, hereafter 'the web forum corpus' or 'the web corpus'. The structuring of the two corpora bears strong resemblance to the SweFN examples: Each sentence is dependency parsed and POS-tagged, and each word is lemmatised and suggestions are given for possible SALDO entries (see 3.4.2). For our purposes, it is particularly helpful that the corpora also provide suggestions for the segmentation of compounds. Both corpora are shuffled beforehand.

The compounds are selected by the criteria of (1) the corpus structure showing suggestions for segmentation, and (2) non-lexicality as defined by an empty search result in the SALDO lexicon. This method returns compounds such as the non-lexicalised *duvskit* 'pigeon droppings', but not the lexicalised *ljusblå* 'light blue', since it has an entry in SALDO. This selection method is successful, returning more compounds than time permits us to analyse. The selection method also results in a small amount of noise caused by false segmentation, e.g. the surname *Olaisen* is interpreted as the first name *Ola* + *isen* 'the ice'. These are manually cleaned during annotation.

3.3 Annotation of compounds

We develop an annotation interface for preparing the compounds extracted from the literary and web forum corpora. While a description follows, screenshots of the annotation interface may be found in Appendix A (page 36).

The annotator (a native Swedish speaker well conversant with FrameNet and SALDO, primarily the author of this thesis) is presented with a compound along with one or more suggestions for segmentation, and the sentence it appears in for context. The annotator is asked to confirm the compositionality of the

[2] https://spraakbanken.gu.se/eng

compound and to identify the correct lemmas if ambiguous, i.e. lemma disambiguation. The annotator is then asked to assign a frame to the suffix. In order to select the best frame, the annotator may need some help to choose between the 1003 frames. This is done in the following way: Each frame in SweFN has a finite set of lexical units, which are in the format of SALDO entries. The chosen suffix lemma may translate into several possible SALDO entries, so the annotation program looks up all of them, as well as their primary descriptors, in SweFN. If the SALDO entry or its primary descriptor is found as a lexical unit of a frame, that frame is presented as an option. However – in the case that the desired word sense does not evoke a frame, the annotator may override the suggestions and select whichever of all frames she or he sees fit. When a suitable frame has been selected, the annotator is presented with the set of semantic roles pertaining to the chosen frame, and selects the appropriate role.

In the **first annotation round**, 1147 compounds are annotated – 918 from the literary corpus and 229 from the web forum corpus.
We note that 80% of the compounds from the literary corpus and 76% from the web corpus are nn+nn (noun+noun) compounds.[3] The internal order of the next four most common compound types with regards to part of speech differs slightly between the two corpora, as seen in Table 3.

Literary corpus (size: 918)			
nn+nn	80%	*frukostkorven*	‘breakfast sausage’
nn+av	3.7%	*hastighetsbegränsade*	‘speed-restricted’
vb+nn	2.9%	*bindvävnad*	‘connective tissue’
nn+vb	2.6%	*bråtebelamrad*	‘debris-cluttered’
av+nn	2.2%	*extrabyxor*	‘spare trousers’

Web corpus (size: 229)			
nn+nn	76%	*spöval*	‘choice of fishing rod’
vb+nn	5.7%	*grillplatsen*	‘BBQ-ing spot’
av+nn	4.8%	*gopappor*	‘sympathetic dads’
nn+vb	3.5%	*handtvätta*	‘hand wash’
nn+av	2.6%	*båtintresserade*	‘boat-interested’

Table 3: Distribution of compounds with regards to POS (nn=noun, vb=verb, av=adjective).

The semantic distribution between compounds is also skewed. With a total of 1003 frames, an even distribution would entail that each frame was represented approximately once, or in 0.1% of our annotated set of 1147 compounds. This is clearly not the case, since 85% of our compounds are noun-headed while most frames tend to be event-oriented (verb LUs) rather than entity-oriented (noun LUs). In fact, only 323 of 1003 possible frames are represented in the first annotation round.
Furthermore, the distribution between the frames that do occur is heavily skewed – 45% of the 323 frames only occur once, and 17% twice. Conversely, 9% or 30 of the frames make up 45% of the instances.

As it becomes obvious that the frame distribution of our compounds is vastly different from that of the SweFN example sentences, we narrow our focus for the preparation of compounds for the second experiment.

In the **second annotation round**, we consider only the five most common frames of the literary corpus from the first round: CONTAINERS (3.6% of occurrences), CLOTHING (2.4%), FOOD (2.4%), FURNITURE (2.4%), and PEOPLE_BY_VOCATION (2.3%). A further 200 compounds for each of these frames are annotated from

[3] Full tagset: https://spraakbanken.gu.se/eng/research/saldo/tagset

the literary corpus. The decision to proceed with the literary corpus only is based on the judgement that it is more representative of general language use than the web forum corpus, whose top five frames are VEHICLE, AGGREGATE, HUNTING, CONTAINERS, and PEOPLE.

3.3.1 Inter-annotator agreement measure

Cohen's Kappa (Cohen, 1960) is a means of measuring pairwise agreement in a given judgement task. We use it to test the reliability of our hand-labelled data, i.e. our judgement of which of a set of roles a compound modifier belongs to. In order to be able to perform this pairwise evaluation, two annotators mark up the same body of compounds. We then compute the degree to which they agree in their choice of class labels for each annotation instance.

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{1}$$

Let us explain Cohen's Kappa (κ) formula as seen in Equation 1 using a toy example. Imagine that frame X has two FEs: y and z, and occurs four times in the data. Annotator 1 assigns the roles [y, y, z, y], while Annotator 2 assigns roles [y, y, z, z,]. We thus observe agreement in three out of four cases, i.e. $P(a) = 0.75$. However, Cohen's κ statistic not only calculates the average agreement, but it compensates for the expected agreement or chance agreement i.e. the probability that the annotators may agree purely by chance. This can be thought of as a combination of the number of classes available, and the two annotators' individual preference for a certain label. Calculating the chance agreement involves adding the marginal frequency of each class and dividing by the total number of instances, i.e. for our example:

$$P(e) = \frac{(1+2)(0+2)/4 + (1+0)(1+1)/4}{4} = 0.5 \tag{2}$$

The measure of chance agreement is particularly important when measuring the reliability of agreement in a judgement task with few categories, but it confirms reliability also in our case, where the number of class labels per frame range from one or two to over 20 classes.

The result for our toy example is shown in Equation 3:

$$\kappa = \frac{0.75 - 0.5}{1 - 0.5} = 0.5 \tag{3}$$

We will report the κ value for agreement between two annotators from two annotation rounds in Chapter 4.

3.4 Resources

3.4.1 SweFN

The Swedish FrameNet (SweFN) (Borin et al., 2010a,b; Ahlberg et al., 2014) is a lexical and semantic resource developed by researchers and hosted at Språkbanken at the University of Gothenburg. It is based on Frame Semantics theory (see section 2.7) with the core notion that the meaning of words are learned and conceptualised within their context, i.e. for a MEMORIZATION event to take place, there needs to be a Cognizer who memorises, as well as a Pattern for the Cognizer to memorise. The word written in small caps refers to the **frame** name, and a frame is evoked by a predicate, an item out of a set of lexical units, **LU**s. In the case of MEMORIZATION, the set of LUs include verbs like *memorera* 'memorise' and nouns like *memorerande* 'memorising'. The upper case-initial words above (Cognizer, Pattern) are examples of frame elements (**FE**s), more specifically **core FE**s. **Peripheral FE**s are usually semantic roles representing optional information, e.g. in the MEMORIZATION frame, the peripheral FEs are Completeness, Time, Place, Means, Manner, and Purpose.

SweFN bases its frames on those of Berkeley FrameNet (BFN) but adds, deletes, or amends frames and frame elements where necessary to fit the Swedish language. As of mid-2016, the resource consists of 1003 frames and 8,500 example sentences with annotations. It also has information about common or observed compound types in around 200 of the frames, e.g. in the ARCHITECTURAL_PART frame:

- **Material**+LU *kakelugn* 'tiled stove'
- **Whole**+LU *takfönster* (roof+window) 'skylight'
- **Orientation**+LU *yttervägg* 'outer wall'
- **Description**+LU *sågtak* 'saw-tooth roof'

SweFN is integrated with other lexical and language technology resources at Språkbanken. For an overview on the building and integration of the Swedish FrameNet, see Dannélls et al. (2014). Most relevant for the purposes of this thesis is its linkage to the sense lexicon SALDO. Each frame in SweFN has a designated set of lexical units (**LU**s) belonging to it in the form of SALDO entries. Since SALDO is a word-sense disambiguated lexicon, all lexical units are exclusive to one frame.
The main difference between BFN and SweFN is that while the Berkeley project has focused on providing a substantial number of tagged example sentences, SweFN has prioritised the adding of LUs, i.e. linking it up with SALDO, which is unfortunate for our initial experiments in which we use the example sentences as training data.

3.4.2 SALDO

SALDO (Swedish Associative Thesaurus version 2), (Borin et al., 2008, 2013) is, like SweFN, a lexical and semantic resource developed by researchers and hosted at Språkbanken at the University of Gothenburg.

The SALDO lexicon has one entry per word sense, and it is organised as a network of word senses with semantic links between the entries. The base of the lexicon is a top node, PRIM, which is the parent of 43 core senses or primitives (Borin et al., 2013). The lexicon also contains lemma information and inflection tables for each lemma. Contrary to other semantic lexica such as the English and Swedish WordNets (Fellbaum, 1998; Viberg et al., 2003), SALDO is organised by cognitively motivated associations between word senses rather than by taxonomy. While the entries of WordNet are defined strictly in relation to their synonyms, hyper-/hyponyms, meronyms and antonyms, SALDO entries are defined by at least one semantic **descriptor**. The **primary descriptor** has to fit the criteria of being: (1) a semantic neighbor of the entry to be described, and (2) more central than it. By semantic neighbor, Borin et al. (2013) intend a formal semantic relationship '[...] such as synonymy, hyponymy, meronymy, argument-predicate relationship and so on', and by centrality, being more frequent and/or less stylistically marked than the entry it describes. Table 4 exemplifies part of the SALDO structure — note how it treats different word senses as separate entries.

Entry	Descriptor(s)
läder ‘leather’	*hud* ‘hide’ + *djur* ‘animal’
äpple ‘apple’	*frukt* ‘fruit’
knivig ‘tricky’	*svår* ‘difficult’
gissa ‘guess’	*ana* ‘forebode’ + *försöka* ‘try’
marmor ‘marble’	*kalksten* ‘limestone’
[…]	
bar ‘bare’	*naken* ‘naked’
bar..2 ‘bar’	*lokal* ‘locale’ + *äta* ‘eat’
bar..3 ‘bar’	*mått* ‘measure’ + *lufttryck* ‘air pressure’

Table 4: A selection of SALDO entries with their semantic descriptors.

3.4.3 Brown clusters

The Brown clustering algorithm (Brown et al., 1992) is a language model-type method of semantic grouping. It takes as input a large corpus of running text and a subset of the vocabulary occurring in the text. This vocabulary represents the items to be clustered. The basic idea is that starting as single ‘islands’, the items (or words, for simplicity) merge into a tree-structure of semantic relatedness. To find semantically related words, the algorithm takes into account the history of each word, as defined by an n-gram of preceding words. In other words: words are judged as semantically close if they frequently occur in the same environment: *for lunch, I had a* __ (salad/pizza/pear).

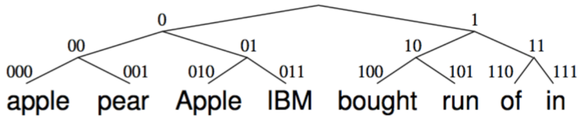


Figure 2: Brown clustering.

Given a large corpus of written text to train on, the algorithm outputs the input vocabulary tagged with cluster IDs for each word. The cluster IDs are in the form of bit strings, such that i.e. the clusters 010 and 011 are generalisable into the cluster 01. The longer the bit string, thus, the smaller the cluster, with the most specific ‘cluster’ being a single word, see Figure 2 from Koo et al. (2008). This way, different magnitudes of semantic relatedness is captured.

For our experiments, we use the clustered output of training corpora of around 1 billion words from Språkbanken. The vocabulary size is 1 million items.

3.4.4 Synlex

Synlex is a multi-user collaborative synonym lexicon developed by Kann & Rosell (2005). It obtains its synonym pairs by letting users suggest and rate synonym pairs before being able to search the lexicon. It is controlled for sabotage through the random generation of ratable pairs, i.e. a user is highly unlikely to encounter a pair to rate that she or he suggested themselves. Synlex is searchable[4] and downloadable.[5]

[4] `http://folkets-lexikon.csc.kth.se/cgi-bin/synlex` (last accessed 29 August 2016)

[5] `http://folkets-lexikon.csc.kth.se/synlex.html` (last accessed 29 August 2016)

The downloadable version of the lexicon includes only synonym pairs which have a rating of 3 and above on a 0-5 scale, and a minimum of reviews is set for the synonyms to appear in search. We use the late 2013 version of Synlex.

3.4.5 Machine learning algorithm

There are a number of machine learning algorithms to consider for a multi-class classification problem like the one at hand. The Support Vector Machine (SVM) (Cortes & Vapnik, 1995) is chosen because of a good track record in Natural Language Processing applications. Without going into the mathematics of the SVM, the algorithm takes the feature vectors of the training data and updates the weight of each feature based on seen instances. Compared to other machine learning algorithms, it does this in a way that maximises the separation between classes. Consider the toy example of a two-dimensional, two-class classification problem, illustrated in Figure 3[6]. It shows the difference between an algorithm that settles as soon as it finds a hyperplane that separates the classes, and the SVM, which finds the maximum margin between the classes. In this two-dimensional example, the hyperplanes are in the form of lines, the left visualisation showing several possible ways of separating the data, while the right illustrates the SVM. We use an SVM module from Scikit-learn (Pedregosa et al., 2011): Linear Support Vector Classifier (LinearSVC)[7]. Our word-based vectors are translated into number vectors with DictVectorizer[8].

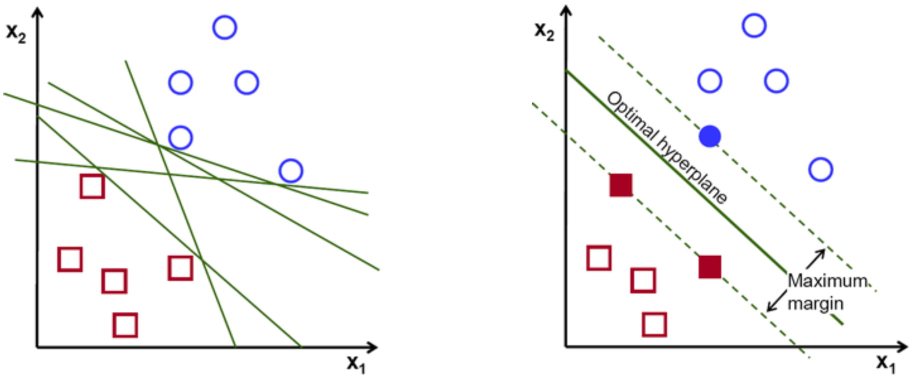


Figure 3: Support Vector Machine. Left: several possible hyperplanes separating two classes. Right: the hyperplane with the maximal margin.

3.5 Experiments

In this section, we first give a detailed account of each feature type used in the experiments. Then, we define our baseline, and finally we describe in short our two main experiment sections.

3.5.1 Feature design

In this section, we describe in detail the features we explore in our experiments. The intuition behind the selection of these features is that the classifier primarily needs semantic information in order to find patterns in the data. Therefore, we experiment mainly with different methods of semantic grouping:

[6] Figures from OpenCV: `http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html`

[7] `http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html` (last accessed 28 August 2016)

[8] `http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.DictVectorizer.html#sklearn.feature_extraction.DictVectorizer` (last accessed 28 August 2016)

clustering, the semantic descriptors in SALDO, and synonyms. Of course, we also use word features in various forms. We use the feature vector of the compound *företagsväska* ‘company bag’ as an example:

Word: Word value features for prefix and suffix in three versions. Full lemgram: `LU_full='väska..nn.1'`, lemma with POS: `LU_with_POS='väska..nn'`, and bare lemma: `LU_short='väska'`. Boolean value features are encoded with their position, e.g. `FE_företag=True` and `LU_väska=True`.

POS: Part-of-speech features are considered for both prefix and suffix: `FE_POS='NN'` `LU_POS='NN'`.

SALDO: The first SALDO entry + primary descriptor in a lookup of the prefix and suffix lemmas. In other words: no manual or automatic word sense disambiguation is carried out. `LU_saldo='väska..1'`, `LU_prim='förvara..1'`.

Frame: Frame feature: `Frame='Containers'`.

SuperFrame: Abstract frame feature: `SuperFrame='Entity'`.

FrameNeighbours: All other LUs pertaining to the same frame as the suffix. E.g. `LU_box..1=True`.

Compound: Where annotated compounds make up both train and test data, the whole compound is considered a feature: `compound='företagsväska'`.

Context: Where a context sentence is present, the string values of two positions on each side of the compound are considered each one feature: `compound-2='en'`, `compound-1='jätteful'` etc.

Synlex: We search the Synlex lexicon for synonyms of both prefix and suffix and encode them like the boolean word features so that the information is shared across feature vectors, e.g. if *påse* is a synonym of the suffix (LU), then `LU_påse=True` is added to the vector. For maximum recall, we tweak the lexicon by taking each of the synonyms of the target word, searching for *their* synonyms, and saving those that appear at least twice among the synonyms’ synonyms. For example: *företag* ‘company, business’ has the synonyms *affärsverksamhet*, *bolag*, *firma*, and *näringsverksamhet*. Two of them, *affärsverksamhet* and *firma*, also have *rörelse* among their synonyms. We consider this as an indicator of strong relatedness to the target word, and add it to the rest of the synonyms.

Clusters: Semantic cluster IDs for prefix, suffix, and any Synlex synonyms are retrieved from a document[9] of nearly one million words produced using the Brown clustering algorithm. For instance: `FE_cluster_001110001=True` (this may come from either the prefix lemma or from one of its synonyms).

3.5.2 Baseline

As a baseline, we train a classifier with Frame as the only feature. This leads to the classifier consistently predicting the most frequent class label given the frame. Our test results will also be discussed in relation to the κ value for inter-annotator agreement.

[9] Courtesy of the supervisor of this thesis, Richard Johansson.

3.5.3 First experiment

In the first experiment, the SweFN annotated sentences are processed into train instances, which are subsequently used to train an SVM. As test instances, we use 500 compounds from the literary corpus, annotated as described in Section 3.3. For features in this experiment we use: Words, POS, SALDO (entry and primary descriptor), Frame, SuperFrame, FrameNeighbours, Synlex and Clusters. We report average accuracy.

3.5.4 Second experiment

In the second experiment, we use compounds both as train and test data. As described in Section 3.3, we narrow it down to the five most common frames in the literary corpus: Containers, Clothing, Furniture, Food, and People_by_vocation. In this experiment, all features from the list in 3.5.1 except FrameNeighbours are explored. We report 10-fold cross-validation average accuracy on the development set (960 compounds), and average accuracy on the test set (100 compounds).

4 Results

In this chapter, we report the average accuracies of a number of trained models. Recall that the **first experiment** refers to the model trained on the SweFN sentence data and tested on annotated compounds with no pruning of frames, while the **second experiment** refers to the one in which we narrow down to five frames and annotate a new development set plus test set with compounds only.

4.1 First experiment

In the first experiment, the SweFN annotated sentences are processed into train instances which are subsequently used to train an SVM. As test instances, we use 500 compounds annotated as in section 3.3. For features in this experiment we use Frame, SuperFrame, FrameNeighbour, Word features for prefix and suffix, the best guess SALDO entry and primary descriptor for prefix and suffix, Synlex synonyms for prefix and suffix, and Cluster for all Word and Synlex features, again for both prefix and suffix. We call this model SweFN-ex, and the result is shown in Table 5.

	Average accuracy
Baseline	0.216
SweFN-ex	0.290

Table 5: First experiment: Results.

As we see, the result is but a small improvement over the baseline model. We attribute this mainly to the sparsity of data: out of ∼1,000 frames in the train set, only ∼200 are represented in the test set of 500 instances, and one frame in the test set is unseen in the train set. With the abundance of frames, each having their own set of semantic roles (although there is some overlap), it is rather surprising that no more than four class labels (Empathy_target, Conflict, Enclosed_region, Service_provider) in the test set are previously unseen. Nevertheless, this makes them impossible to predict.

4.2 Second experiment

In our second experiment, the development set consists of 960 annotated compounds from five frames: Containers, Clothing, Food, Furniture, and People_by_vocation. In addition to the features from the first experiment, we now also consider the Context and Compound features (see Section 3.5.1). We do, however, drop the FrameNeighbours features as they merely add weight to the Frame feature. The FrameNeighbours features were necessary in the first experiment in order to avoid confusion about what set of labels should be considered for each frame, but in our narrowed down second experiment they only add computational cost without improvement.

We split this subsection into one accuracies and feature analysis part and one distribution and error analysis part.

4.2.1 Accuracies and feature analysis

In Table 6, we see a big improvement over baseline, with 0.614 average accuracy for our model.

We carry out an ablation test i.e. omitting one feature at the time to see what individual contribution each feature has on the accuracy. The last column in Table 7 represents the reversed effect of the omission of that feature i.e. if it shows a positive value the feature is good, and conversely, a negative value indicates a harmful feature. The figures in bold face represent the extremes.

	Average accuracy	**Standard deviation**
Baseline	0.379	±3.1%
All features	0.614	±2.8%

Table 6: Development set: 10-fold cross-validation results. All features vs Baseline.

	Omitted feature(s)	**Average accuracy**	**Standard deviation**	**Effect**
	(All features	0.614)		
1	Context	0.608	±3.0%	0.6%
2	Frame	0.602	±2.6%	1.2%
3	SuperFrame	0.61	±2.8%	0.4%
4	Frame+SuperFrame	0.593	±2.7%	2.1%
5	Cluster (all)	0.605	±3.1%	0.9%
6	Cluster (prefix)	0.597	±3.4%	1.7%
7	Cluster (suffix)	0.612	±2.6%	0.2%
8	Synlex (all)	0.616	±2.9%	**-0.2%**
9	Synlex (prefix)	0.595	±2.9%	1.9%
10	Synlex (suffix)	0.613	±3.5%	0.1%
11	SALDO (prefix)	0.609	±3.1%	0.5%
12	SALDO (suffix)	0.611	±2.9%	0.3%
13	SALDO prim.descr. (prefix)	0.585	±3.6%	2.9%
14	SALDO prim.descr. (suffix)	0.616	±2.2%	**-0.2%**
15	Word+SALDO (prefix)	0.521	±3.5%	**9.3%**
16	Word+SALDO (suffix)	0.602	±2.5%	1.2%
17	POS (prefix)	0.61	±3.2%	0.4%
18	POS (suffix)	0.612	±3.0%	0.2%

Table 7: Development set: Ablation test with 10-fold cross-validation.

The group of features that clearly contributes the most to the performance of the classifier (9.3% reversed effect) is the lexical features for the prefix (line 15), which is unsurprising as the prefix is the word that ultimately evokes the semantic role.

The second most contributing feature according to our ablation test is the SALDO primary descriptor for the prefix (line 13). It seems intuitive that some prefixes pertaining to the same semantic role would be children of the same descriptor, e.g. for both prefixes in *tennisströja* 'tennis shirt' and *golftröja* 'golf shirt' (Use), *sport* is the primary descriptor.

Another observation is that the effect of deleting Frame and SuperFrame (line 4) is greater than the sum of their individual effects (lines 2, 3). This is due to the fact that the two features often share the same value, i.e. frames that do not appear in the BFN inheritance hierarchy file simply have the same value for Frame and SuperFrame.

We note that two features contribute negatively to the accuracy: the Synlex synonyms (line 8) and the primary descriptor of the suffix (line 14). The fact that Synlex impacts negatively is surprising, since we expect that adding synonyms should widen the vocabulary of the classifier. What is even more surprising is that the Synlex features for prefix and suffix *respectively* contributed positively (line 9, 10). In other words, there seems to be a measure of confusion among the Synlex features despite the fact that they are codified with their position i.e. prefix or suffix. This may have to do with their interaction with the Cluster features. A reason for the harm that the Synlex features make may be the lack of word sense disambiguation. In other words, there may be false synonyms mixed in with the true. Indeed, it is true that we equally do no WSD when it comes to polysemous words in selecting the SALDO feature. Instead, we choose the first of the suggested SALDO entries, which are sorted in order of saliency according to the lexicographers. The first SALDO sense, however, is reported as the correct choice in 77% of the cases among polysemous words in running text from a novels corpus in Nieto Piña & Johansson (2016). This may explain why the SALDO feature for prefix does not confuse the classifier the way Synlex does. The SALDO primary descriptor feature of the suffix, however, had a negative impact. Again, this is best described with possible mismatches due to WSD or confusion caused by the SALDO graph.

The fact that the POS feature had little effect is unsurprising with regard to the distribution of compound types – a vast majority of the compounds are noun+noun.

Overall, the features representing lexical meaning for the prefix had the most positive effects (lines 6, 9, 13). This is expected since it is the prefix that predicts the semantic role, while the suffix determines the frame, which is already given.

As regards the rest of the ablation tests, it is interesting that few of the deletions cause any dramatic drop or increase in accuracy. This could be attributed to certain levels of shared information between features, making them hard to isolate.

Not present in Table 7 are the results of experiments involving different levels of granularity in the Brown clusters (see Section 3.5.1), which had no effect. We attribute this to the fact that the clusters were already quite coarse: 1024 clusters distributed over 1 million words, which took away the generalisation effect.

We proceed to the final test set, and train two models: one with all features used in the development set, and one with an optimised set stripped of the negatively impacting features in the development experiments, i.e. LU primary descriptor and Synlex. In addition to testing the full feature set against the optimised, we evaluate the effect of shared class labels between frames. Following Johansson et al. (2012) we rename the role labels, e.g. Material is split into Containers_Material, Clothing_Material, and Furniture_Material, and retrain a classifier on the optimised feature set.

	Accuracy
Baseline	0.33
All features	0.61
Optimised feature set	**0.62**
Frame-specific classes	0.52

Table 8: Test set: Results for models trained on the development set and tested on 100 previously unseen compounds.

In table 8, the results of the three classifiers are presented along with the result of the baseline classifier. As can be seen, the model optimised on the development set perform the best also on the test set. The frame-specific classes model performs much lower, ten percentage points, than that of the corresponding model *with* label generalisation. This follows the linguistic intuition of overlap between e.g. furniture materials and container materials (wood, steel, plastic, etc.) – it is expected that the frames learn lexical patterns from each other.

The baseline accuracy in the final test round is 0.33, i.e. almost five percentage points lower than the average baseline accuracy in the development round, while the accuracies of our best classifiers match or slightly outperform the best development set model. This may be because of the increase of 100 training instances in the final test round, or within the margin of error.

4.2.2 Distribution and error analyses

Before analysing the errors that our classifiers make, let us take a look at the distribution of classes in Table 9. Evidently, the classes are unevenly distributed. For sparsity reasons, however, no attempts are made at balancing the data.

If we treat each semantic role as frame specific, there are 41 populated classes, i.e. at least one instance per class. Our approach, however, is to generalise over frames, so that the Descriptor role of Containers is considered equivalent to the Descriptor roles of Clothing, Furniture, People_by_vocation, and Food. Given this merge, there are 27 populated classes in the development set. The test set has been selected randomly, with the reservation that there be no unseen classes in the test set.

Containers	Furniture	Clothing	Food	People_by_voc.
Contents (107)	Function (43)	Material (70)	Type (79)	Type (121)
Material (41)	Material (43)	Use (66)	Constit._parts (71)	Place_of_emp. (38)
Use (26)	Type (43)	Style/Type (39)	Descriptor (38)	Employer (9)
Type (19)	Place (26)	Wearer (22)	Food/Entity (14)	Rank (6)
Relative_loc. (12)	Relative_loc. (8)	Descriptor (12)		Descriptor (5)
Descriptor (6)	Descriptor (3)	Subregion (4)		Contract_basis (4)
Owner (3)	Name (2)	Creator (2)		Origin (4)
Construction (2)	Time_of_cr. (1)	Garment/Entity (1)		Person/Entity (4)
Container (1)		Body_location (0)		Persistent_char. (2)
Part(0)				Context_of_acq. (2)
				Ethnicity (2)
				Age (0)
				Compensation (0)
Total: 217	+ 169	+ 216	+ 202	+ 197 = **1001**

Table 9: Development set: Distribution of FEs per frame in the development set.

The most common misclassification types as illustrated in Table 10 reflect to some extent the difficulty with which the roles are discerned from one another, and we will look at some examples in a moment. We identify this as a documentation issue, since no definitions of the kind that are available for BFN, are available for SweFN. We therefore rely on the definitions in BFN while annotating, to no avail in the SweFN-specific Furniture frame.

As expected, the most common misclassifications generally correspond to the most frequent FEs. The exception to this is Containers – Owner, which had three instances in the whole development set. What probably happen here is that the classifier has no best guess, and therefore goes for the most frequent class for that frame: Contents. Another exception is the Material role, which we will return to in a moment.

Among the most common classification errors are the Descriptor and Type roles in Food. In the Food frame, the examples of Type and Descriptor in the Berkeley FrameNet documentation are rather similar. For example: the modifier ‘low-fat’ in relation to ‘milk’ is tagged as Descriptor, while ‘cooking’ in ‘cooking apple’ is tagged as Type, which seems quite fine-grained and may lead to confusion during annotation.

Frame	Predicted	Correct	%	Example
PEOPLE_BY_VOCATION	Type	Place_of_employment	6.2%	*bodexpedit* 'store clerk'
FOOD	Type	Descriptor	4.3%	*dip-grönsakerna* 'dip vegetables'
FOOD	Descriptor	Type	4.0%	*Galaäpplen* 'Gala apples'
FURNITURE	Type	Function	4.0%	*läsfåtöljen* 'reading chair'
FOOD	Constit._parts	Type	3.8%	*dilamm* 'milk-fed lamb'
CLOTHING	Use	Type (Style)	3.2%	*poloskjortan* 'polo shirt'
CONTAINERS	Contents	Use	3.2%	*frysväskan* (freeze+bag) 'coolbox'
PEOPLE_BY_VOCATION	Type	Employer	2.9%	*advokatsekreterare* (lawyer+secretary) 'legal secretary'
FURNITURE	Function	Type	2.7%	*orgelpallen* 'organ stool'
CONTAINERS	Contents	Owner	2.4%	*personalgrytan* 'staff pot'

Table 10: Development set: Most common misclassifications. The percent value stands for the proportion of total mistakes. The pre-abstracted FE is in parenthesis.

We note that the Material role, shared by and popular in three frames, is absent from our common misclassifications table. This may be because of the close-knit character of this class, with prefixes such as 'leather', 'wool', 'fabric', 'wood', 'tin' etc.

Style and Use in the CLOTHING frame is a difficult distinction when it comes to real-life examples – is a *poloskjorta* 'polo shirt' primarily signified by its Use or by its Style? To an extent, the conceptualisation of what is Use versus Style goes hand in hand with lexicalisation. As a CLOTHING compound goes from Use to Style, it gets closer to lexicalisation. The same discussion goes for Use versus Wearer. In the development set, we have marked e.g. *seglarbyxor* 'sailor trousers' as Style as there is no indication in the context sentence of the wearer being a sailor. Again, this may be an indication that *seglarbyxor* should indeed be considered lexicalised and be given an entry in SALDO.

Parallel to the CLOTHING problems of Use/Style and Wearer/Style is the issue of FURNITURE – Function/Type. Since the FURNITURE frame has no definitions for frame or FE available, the roles are interpreted in parallel with CLOTHING and CONTAINERS, as it shares most of its FEs with them. It is interesting that misclassifications between FURNITURE Function and FURNITURE Type are common in both directions. Here too, we see a potential lexicalisation pattern between Function and Type: A piece of furniture, e.g. *orgelpall* 'organ stool' has a particular Function, and therefore a specific design (wide and adjustable) to fit its function. Thus, the shape or design, i.e. the Type, may be more salient in the minds of speakers than its original Use.

4.3 Inter-annotator agreement

Using Cohen's κ metric described in Section 3.3.1, we compute the inter-annotator agreement between two annotators twice.

First, we consider 100 compounds from the first annotation round. As the number of possible categories i.e. semantic roles ranges from one to 20+ per frame, we calculate a separate chance agreement per frame. There is a slight difference between the task of the annotators and the task of the system: while the system has the frame as a given, the annotation task consists of assigning both a relevant frame and the appropriate role from the set of roles pertaining to the chosen frame. In order to capture equivalence with our classification system, therefore, we consider only those compounds for which the two annotators agree with regard to frame, segmentation, and lemma disambiguation. The inter-annotator agreement score of $\kappa = 0.71$ is considered good, i.e. the categories are relatively intuitive to humans.

We compute another κ on the final test set of the second experiment round, i.e. with only five frames. This set, like the first, is 100 instances (average number of classes: 9.6), and we reach a lower average score than previously: $\kappa = 0.65$, which is considered fair-good. The κ values vary between frames, see Table 11.

The absolute agreement between the two annotators was 66 out of 100. The most common disagreements involved Contents/Use in CONTAINERS: *ginglaset* 'gin glass' and Constituent_parts/Type in FOOD: *fågelköttet* 'bird meat'. Both of these disagreements occurred four times, however in the former, each annotator consistently assigned the same label, while in the latter, the disagreement went both ways.

With the hypothesis that these for humans tricky cases would also be tricky cases for our system, we expect to see a similar pattern in the most common system misclassifications. Indeed, the two mentioned disagreements are reflected among the most common misclassifications in the development set for our automatic system (Table 10), in places 5 and 7.

Other recurring disagreements were Descriptor/Type in FOOD – *gästabudsäpple* 'banquet apple', Type/Use in CONTAINERS – *skolväskor* 'school bags', and Descriptor/Style in CLOTHING – *platåskor* 'platform shoes'. These occur three times each, however only the first is reflected in Table 10. Disagreement regarding Constituent_parts/Food in FOOD – *sköldpaddsbiffar* 'turtle patties' occurred twice. All other disagreement types were unique.

CLOTHING	$n = 21$	$\kappa = 0.71$
CONTAINERS	27	0.40
FOOD	14	0.60
FURNITURE	17	**0.77**
PEOPLE_BY_VOCATION	21	0.66

Table 11: Test set: Inter-annotator agreement per frame.

To summarise this chapter, we have seen the results from several experiments. First, we saw the results of the SweFN-ex model, which only by a few percent outperformed the baseline. We then saw 10-fold cross-validation results for our five-frame development data set, which was around 0.61 against a 0.38 baseline, and analyses of the contributions of the different features. Finally, we saw the results of our optimised classifier tested on a set of 100 compounds: 0.62.

Our best accuracy score of 0.62 should be seen in light of the fact that there is a measure of conflict between human readers about which role to assign to the compound prefix. Recall our inter-annotator agreement of $\kappa = 0.65$ for the five-frame task. The observed agreement before compensating for chance agreement was 0.66. With this in mind, the performance of our system is close to equal to that of the agreement between two human Swedish speakers.

5 Discussion

In this chapter, we discuss the results presented in Chapter 4 and reason about how they came to be. We also compare our results to related work, and look into future improvements of our system.

Our best semantic role disambiguation model performed 62% accuracy on 100 novel compounds. The full test set is appended to this paper for the reader's reference (Appendix B, page 37, in Swedish). We compare the result to that obtained in Rosario & Hearst (2001), who obtained around 60% for a 18-class classification task. Compared to Rosario & Hearst (2001), our set of classes may be seen as rather large at 27 semantic roles, or as modest with an average of 8 classes per frame in the second experiment(s). As we see, the classification accuracy in Rosario & Hearst (2001) is about the only common feature between their experiments and ours. While we used SweFN and SALDO, they used other lexical resources for English, and while we looked at five semantic areas, they restricted their experiments to one domain (medicine).

Our best classification accuracy in relation to baseline was also close to those of running-text SRL in Johansson & Nugues (2006); Johansson et al. (2012). This is interesting given the differences between the tasks – where SRL can incorporate structural features in addition to semantic, compound semantic role disambiguation relies to a higher extent on semantic and lexical features. However, our limited number of frames compared to the breadth of frames treated in Johansson & Nugues (2006) and Johansson et al. (2012) certainly accounts for some of the performance, making the comparison weak. To our knowledge, there are no system equivalent enough to ours for a fair comparison. Therefore, we move on to a qualitative discussion about Swedish compounds.

In the error analysis, we discussed briefly the blurry line between Style and Wearer/Use in Clothing. Recall our argument that e.g. 'sailor trousers' is conventionalised from Wearer into Style. In the Containers frame, the Use, Type, and Contents roles exhibit a similar phenomenon. For instance, *whiskykaraff* 'whisky decanter', although perhaps conventionally interpreted as Contents, was marked as Use in an attempt to capture the temporary use of such a utensil.

Another example is *snapsglas* 'snaps/shot glass': *Sedan hällde Olaisen upp i snapsglas* 'then Olaisen poured the snaps glasses'. Depending on the aspect considered, *snaps* may be interpreted either as 'a glass with snaps in it' – *Oi, give me my snaps glass!* (Contents) or 'a small glass of a certain shape' – *Do we have any snaps glasses left or did they all break during last midsummer?* (Type).
One could argue here that we are complicating things by transferring the meaning of what is ultimately a liquid, *snaps*. However, what may be thought of as the primary meaning of 'snaps' in 'snaps glass', i.e. Contents, is in fact harder to exemplify than the Type aspect. Compare for instance *would you a glass of snaps?* and *would you like a snaps glass?* The latter seems odd if the asker wants to know if the other person wants the Contents, i.e. a drink. One way of looking at it is that a compound may have a primary role (here: Contents), and a transferred role (Type). For the purposes of automatic classification, in the future more focus may be put on the sentence context in order to disambiguate pairs like the ones we have discussed.

This phenomenon was discussed with linguist Lisa Loenheim in analogy with examples from her as yet unpublished research about the semantic interpretation of compounds among native and non-native Swedish speakers. A 'false friends' pair was discussed: *fruktkorg* 'fruit basket', which for native speakers had connotations of being **filled with fruits**, i.e. a Contents relation, versus *svampkorg* 'mushroom basket', which many native speakers conceptualised as a container for **picking mushrooms**, i.e. a Use relation (personal communication, 9 March 2016). In other words, superficially equivalent compounds may bear different connotations, which in turn impact the interpretation of semantic roles.

A similar issue of conceptual ambiguity relates to the base form of each compound constituent. The lemma interpretation of the compound parts may in fact vary between speakers without causing communicative confusion. For example: the compound *skithus* (lit. shit+house) ‘privy’ may be thought of as a place for the event, *skita* ‘to excrete’, or as a place for the end result, *skit* ‘faeces’. Although the noun and the verb are ultimately derived from the same root, there is still a slight conceptual difference between the two readings. This detail decides the part of speech of the prefix, which in turn informs training and classification in automatic analysis. For the mentioned example, it could mean the difference between the Function and Descriptor roles in the Buildings frame.

As we saw in Section 4.3, our inter-annotator agreement scores were far from perfect in both annotation rounds, although they are not considered unreliable. This is in spite of both annotators being (1) native Swedish speakers, (2) linguists, and (3) well conversant with both FrameNet and SALDO. In order to reach high and reliable results, the annotation procedure must be meticulous. The annotation procedure was incremental (see Appendix A, page 36) and preceded by discussions between the annotators as well as a small test-round. Written guidelines were provided for the second annotator. Despite this, the agreement score remained lower than desired for a clear-cut classification task. This emphasises (a) the importance of annotation training, and (b) the fine-grained nature of the task.

We noted a tendency for the annotators to conceptually rank the FEs in terms of specificity. While it can be said for most prefixes that they represent some Type of the suffix (usually a hyponym), one wants to see first if there is another FE that might fit more specifically. In future, this could be applied to the system as well as to the annotation guidelines. E.g. as we saw in Table 10 on page 26, *advokatsekreterare* ‘legal secretary’ (lit. ‘lawyer secretary’) is labelled as Employer+LU while in truth the predicted Type+LU should not be seen as a fatal mistake. As a future improvement, we might consider a ranking system using a softer decision boundary and thus a fairer evaluation (Rosario & Hearst, 2001). This would also capture some of the inter-annotator disagreement in the case that the data set is prepared by more than one person.

We note that in the SweFN compound examples, exceptions are made from the semantic right-headedness principle for certain lexicalised compounds, e.g. Color – LU+Descriptor: *färgrik* (colour+rich) ‘colourful’, Being_dry – LU+Subregion: *torrskodd* ‘dry-shod’. In our non-lexicalised compounds, there were cases in which the prefix was judged more semantically salient than the suffix, e.g. *iskub* (‘ice cube’), *sköldpaddsbiffar* (‘turtle patties’). We dealt with these either by trying to fit them into the frame evoked by the suffix, or by skipping them.

We note that the omission of the Frame features did not impact the accuracy scores too negatively, only 2.1% down. This is promising for generalisation between frames and for the future inclusion of more and different data.

The biggest challenge regarding our choice of framework, FrameNet, is its somewhat cumbersome nature. There are over 1,000 frames, each with its own definitions and set of roles, which was reflected in the set of labels. In our second experiment round, even though the data was streamlined to five frames, there were 27 populated classes as well as a handful of FEs for which we found no compounds. The size of the frame inventory also plays a part in annotation training time.

We generalised our data in two ways. First, using the BFN inheritance hierarchy, we were able to add a SuperFrame feature, i.e. the ‘grandest’ parent of each frame. This way, the SuperFrame feature for Clothing was Entity (Clothing < Artifact < Entity). Second, we considered semantic roles generalisable over frames, such that Type, Use, Descriptor, Entity and Material were classes populated by compounds from more than one frame. We note that the generalisation of frames and roles works in favour of our experiments. We saw this in Section 4.2.1, where accuracy dropped by 10% when considering each role as frame-specific. Indeed, there is the possibility that prefixes occupy different roles in different frames, however we did not see any confusion of this kind among our five frames. We identify possible conflicts between frames such as the prefix *barn* 'child'. In *barnmisshandel* 'child abuse', *barn* is Victim, while in *barnskor* 'children's shoes', *barn* is Wearer. As long as the roles are not shared between frames, however, this should not pose a problem given sufficient data.

We have a few notes about the Food frame. This was the most difficult one with regard to annotation. It seems that many of the compounds would fit better in a 'meal' category, i.e. prepared food rather than produce. This would entail a set of FEs more well-suited for meals and dishes than the current, which are Type, Descriptor, Constituent_parts, and Food. Following our suggestion, *limefrukt* 'lime fruit' would stay in Food while *uppvisningsbakelser* 'display cake' would go into the proposed, new frame. Constituent_parts would be better described as Ingredient in such a meal frame. Additional FEs could include cook – *Mannerströmfisk* 'fish à la Mannerström', time – *nattmacka* 'midnight snack', and occasion – *examenstårta* 'graduation cake'.

To answer the research questions posed in Chapter 1, we were indeed able to use a Frame Semantics based resource – SweFN – in the automating of semantic role disambiguation in Swedish compounds. What we had hoped for, however, was for the annotated *sentences* in the SweFN database to act as useful training instances. We learned that the distribution between frames was too scattered for this to succeed. Also, the distribution of compounds was skewed in favour of entity-oriented rather than event-oriented frames. Instead, we put substantial effort into preparing a data set by hand, which enjoyed more success than the sentence examples when it came to classification accuracy.

6 Conclusions

6.1 General findings

In this thesis, we have discussed compounding as a productive and frequent phenomenon in Swedish. We have treated different levels of ambiguity, relating to segmentation point, lemma, word sense, and finally semantic role, for which we built our system.

Our best performing system, which involved five semantic frames and a train data set of 960 compositional, non-lexicalised compounds distributed over 27 semantic roles, had an accuracy of 62% when tested on 100 unseen items. The baseline accuracy for the same data was 33%, which was equivalent to the average proportion of the majority class per frame. We have discussed the sub-optimal inter-annotator agreement score of $\kappa = 0.65$. We conclude from this that the task is difficult for human as well as for machine, but also that our system performs rather well given the size of the data set.

Before commencing this work, the hypothesis was that the annotated sentences in SweFN could be used as training data in analogy with the intuition that a paraphrase-like relationship holds between a phrase like *a dress from Prada* and a compound like *Prada dress*. In our first experiment, in which we explored this hypothesis, we saw that the number of example sentences per frame were not enough for giving the model a chance at learning patterns about each frame. For the most common frame among our compounds, CONTAINERS, the full body of examples in SweFN was four sentences, translating into six training instances. Nevertheless, the SweFN-ex model outperformed baseline by a few percent with 29% against 21.6%, showing the potential improvement of such a system given more training data per frame.

With regard to the first research questions posed in Chapter 1, thus, we conclude that Frame Semantics is indeed applicable to our classification problem given the compound markup scheme (Friberg Heppin & Petruck, 2014). While there is no reson to doubt the potential success of other frameworks of semantic compound analysis (Järborg, 2003; Teleman, 1972; Teleman et al., 2010), we find the integrated nature of SweFN advantageous for our purposes. Regarding the second research question, we conclude that running-text annotations have the potential of being more useful than our result suggests, with reservations for sparsity. We could also imagine a mixed training data set with both compounds and sentences being successful.

Finally, we conclude that (Swedish) compounds are a vast and heterogeneous group of linguistic units. We have dealt with a sub-type of them – compositional, non-lexicalised compounds, and we have done this in an attempt to provide automatic analyses of infrequent linguistic units that do not have their own representation in lexical knowledge bases such as SALDO. Given further improvements, it is hoped that our system can aid further NLP applications.

6.2 Future directions

Future directions of research and applications may be: (1) Further improvement of the system through the production of training instances for more frames, as well as evaluating the generalising effects between frames. In addition, it is possible to go deeper into the SALDO graph than the current primary descriptor. For instance, recursively adding primary descriptors under a common 'is_relative' feature to capture more relationships between prefixes than the current 'parent' relation between an entry and its primary descriptor. (2) Setting up a public web interface for automatic analysis of novel compounds. (3)

Integrating the system in Sparv, Språkbanken's annotation tool.[10] (4) Integrating with computer vision. With reference to our discussion about e.g. Style in CLOTHING and Type in CONTAINERS, an interesting integrated artificial intelligence approach would be to link an image search component to our system. See the preface for illustrations. For example, typing *vinglas* 'wine glass' into Google's image search[11] returns an image inventory more visually homogeneous than that of *favoritglas* 'favourite glass'. Using visual similarity measures between the top n image results, a high visual similarity score could indicate a Type relation.

[10] `https://spraakbanken.gu.se/sparv/` (last accessed 12 September 2016)

[11] `https://images.google.com/` (last accessed 12 September 2016)

References

Ahlberg, M., Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., Friberg Heppin, K., Johansson, R., Kokkinakis, D., Olsson, L.-J., & Uppström, J. (2014). Swedish FrameNet++ The Beginning of the End and the End of the Beginning. *Proceedings of the Fifth Swedish Language Technology Conference, Uppsala, 13-14 November 2014*.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98 (pp. 86–90). Stroudsburg, PA, USA: Association for Computational Linguistics.

Borin, L., Dannélls, D., Forsberg, M., Gronostaj, M. T., & Kokkinakis, D. (2010a). Swedish FrameNet++.

Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., & Kokkinakis, D. (2010b). The past meets the present in Swedish FrameNet++. In *14th EURALEX international congress*.

Borin, L., Forsberg, M., & Lönngren, L. (2008). Saldo 1.0 (svenskt associationslexikon version 2). *Språkbanken, University of Gothenburg*.

Borin, L., Forsberg, M., & Lönngren, L. (2013). SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4), 1191–1211.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 467–479.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297.

Dannélls, D., Heppin, K. F., & Ehrlemark, A. (2014). Using language technology resources and tools to construct Swedish FrameNet. In *Workshop on Lexical and Grammatical Resources for Language Processing* (pp. 8–17).

Fellbaum, C., Ed. (1998). *WordNet: an electronic lexical database*. Language, speech, and communication. Cambridge, Mass: MIT Press.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1), 20–32.

Fillmore, C. J., Baker, C. F., & Sato, H. (2002). The FrameNet Database and Software Tools. In *LREC*.

Frege, G. (1884). *Die Grundlagen der Arithmetik (translated in 1950 as Foundations of Arithmetic by J. L. Austin)*. NY: Bantam Books.

Friberg, K. (2007). Decomposing Swedish compounds using memory-based learning. In *Proceedings of the 16th Nordic Conference on Computational Linguistics (Nodalida'07)* (pp. 224–230).

Friberg Heppin, K. & Petruck, M. R. (2014). Encoding of Compounds in Swedish FrameNet. *EACL 2014*, (pp. 67–71).

Gildea, D. & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3), 245–288.

Hedlund, T. (2002). Compounds in dictionary-based cross-language information retrieval. *Information Research*, 7(2), 7–2.

Johansson, R., Adesam, Y., Bouma, G., & Hedberg, K. (2016). A Multi-domain Corpus of Swedish Word Sense Annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)* Portorož, Slovenia.

Johansson, R., Friberg Heppin, K., & Kokkinakis, D. (2012). Semantic Role Labeling with the Swedish FrameNet. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12); Istanbul, Turkey; May 23-25* (pp. 3697–3700).

Johansson, R. & Nugues, P. (2006). A FrameNet-based semantic role labeler for Swedish. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 436–443).: Association for Computational Linguistics.

Järborg, J. (2003). *Semantisk uppmärkning – Metoder, problem och resultat*. Technical report, Research Reports from the Department of Swedish, University of Gothenburg.

Kann, V. & Rosell, M. (2005). Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th NODALIDA conference, Joensuu 2005* (pp. 105–110).: Citeseer.

Koo, T., Carreras Pérez, X., & Collins, M. (2008). Simple semi-supervised dependency parsing. In *46th Annual Meeting of the Association for Computational Linguistics* (pp. 595–603).

Koptjevskaja-Tamm, M. (2009). Proper-name nominal compounds in Swedish between syntax and lexicon. *Rivista di linguistica, A special issue on Compounds between syntax and lexicon*, 21(1).

Lapata, M. (2000). The automatic interpretation of nominalizations. In *AAAI/IAAI* (pp. 716–721).

Larson, M. L. (1984). *Meaning - Bared Translation: A Guide to Cross - Language Equivalence*. University Press of America.

Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Libben, G. (1998). Semantic Transparency in the Processing of Compounds: Consequences for Representation, Processing, and Impairment. *Brain and Language*, 61(1), 30–44.

Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language*, 84(1), 50–64.

Longacre, R. E. (1983). *The Grammar of Discourse*. Plenum.

Maguire, P., Wisniewski, E. J., & Storms, G. (2010). A corpus study of semantic patterns in compounding. *Corpus Linguistics and Linguistic Theory*, 6(1).

Matsubayashi, Y., Okazaki, N., & Tsujii, J. (2009). A comparative study on generalization of semantic roles in FrameNet. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1* (pp. 19–27).: Association for Computational Linguistics.

Nieto Piña, L. & Johansson, R. (2016). Embedding Senses for Efficient Graph-based Word Sense Disambiguation. In *Proceedings of TextGraphs-10* San Diego, United States.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Pollatsek, A., Bertram, R., & Hyönä, J. (2011). Processing novel and lexicalised Finnish compound words. *Journal of Cognitive Psychology*, 23(7), 795–810.

Rosario, B. & Hearst, M. (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 (EMNLP-01)* (pp. 82–90).

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., & Scheffczyk, J. (2010). *FrameNet II: Extended theory and practice*.

Sjöbergh, J. & Kann, V. (2004). Finding the Correct Interpretation of Swedish Compounds, a Statistical Approach. In *LREC*.

Stymne, S., Cancedda, N., & Ahrenberg, L. (2013). Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4), 1067–1108.

Teleman, U. (1972). *Om svenska ord*. Gleerups.

Teleman, U., Hellberg, S., Andersson, E., Christensen, L., & Svenska akademien (2010). *Svenska akademiens grammatik*. Stockholm: Svenska akademien : Norstedt i distribution. OCLC: 704529673.

Vanderwende, L. (1994). Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th conference on Computational linguistics-Volume 2* (pp. 782–788).: Association for Computational Linguistics.

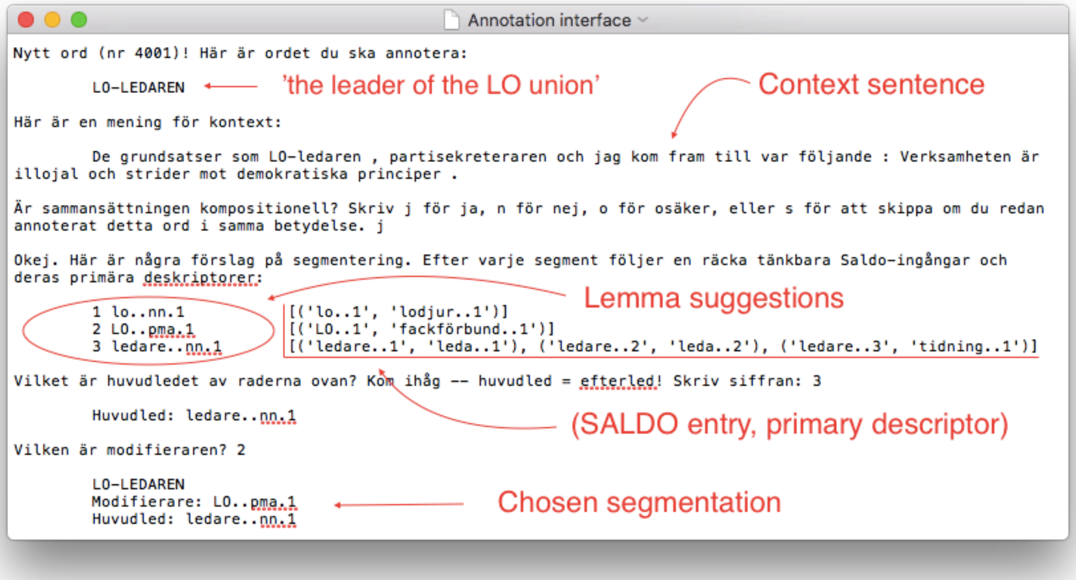
Viberg, A., Lindmark, K., Lindvall, A., & Mellenius, I. (2003). The swedish wordnet project. In *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002: Copenhagen, Denmark, August 13-17, 2002* (pp. 407–412).

Xue, N. & Palmer, M. (2004). Calibrating Features for Semantic Role Labeling. In *EMNLP* (pp. 88–94).

Östling, R. (2010). A Construction Grammar Method for Disambiguating Swedish Compounds. In *SLTC 2010 Workshop on Compounds and Multiword Expressions*.

Appendix A: Annotation interface

The annotation procedure described in section 3.3 is illustrated below by screen shots of the interactive annotation interface with comments in red.



```
Nytt ord (nr 4001)! Här är ordet du ska annotera:

        LO-LEDAREN

Här är en mening för kontext:

        De grundsatser som LO-ledaren , partisekreteraren och jag kom fram till var följande : Verksamheten är
illojal och strider mot demokratiska principer .

Är sammansättningen kompositionell? Skriv j för ja, n för nej, o för osäker, eller s för att skippa om du redan
annoterat detta ord i samma betydelse. j

Okej. Här är några förslag på segmentering. Efter varje segment följer en räcka tänkbara Saldo-ingångar och
deras primära deskriptorer:

        1 lo..nn.1              [('lo..1', 'lodjur..1')]
        2 LO..pma.1             [('LO..1', 'fackförbund..1')]
        3 ledare..nn.1          [('ledare..1', 'leda..1'), ('ledare..2', 'leda..2'), ('ledare..3', 'tidning..1')]

Vilket är huvudledet av raderna ovan? Kom ihåg -- huvudled = efterled! Skriv siffran: 3

        Huvudled: ledare..nn.1

Vilken är modifieraren? 2

        LO-LEDAREN
        Modifierare: LO..pma.1
        Huvudled: ledare..nn.1
```

'the leader of the LO union'

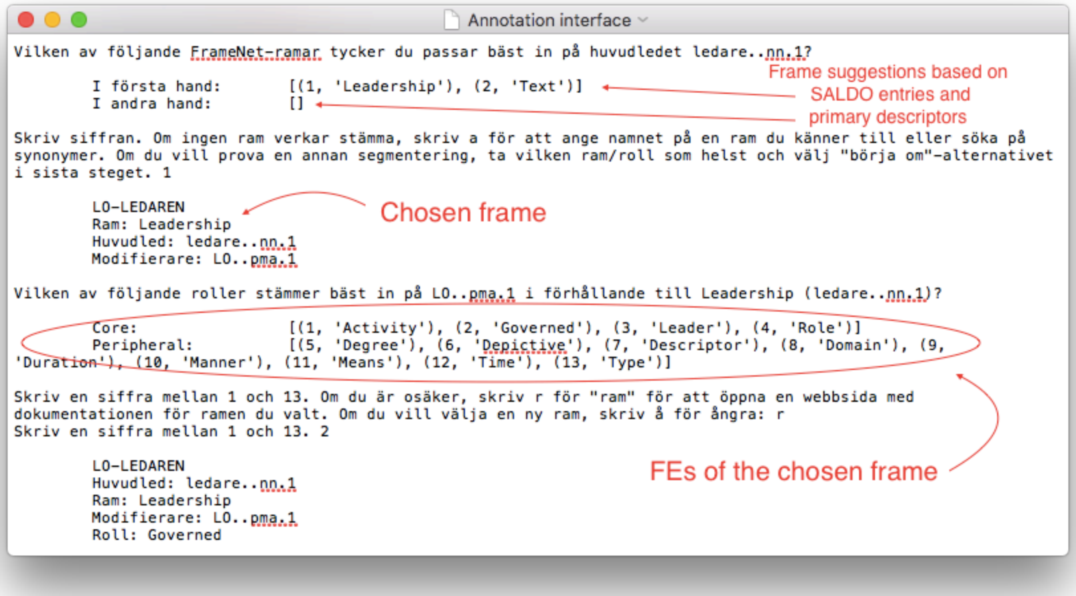
Context sentence

Lemma suggestions

(SALDO entry, primary descriptor)

Chosen segmentation

Figure 4: Annotation interface: first half.



```
Vilken av följande FrameNet-ramar tycker du passar bäst in på huvudledet ledare..nn.1?

        I första hand:          [(1, 'Leadership'), (2, 'Text')]
        I andra hand:           []

Skriv siffran. Om ingen ram verkar stämma, skriv a för att ange namnet på en ram du känner till eller söka på
synonymer. Om du vill prova en annan segmentering, ta vilken ram/roll som helst och välj "börja om"-alternativet
i sista steget. 1

        LO-LEDAREN
        Ram: Leadership
        Huvudled: ledare..nn.1
        Modifierare: LO..pma.1

Vilken av följande roller stämmer bäst in på LO..pma.1 i förhållande till Leadership (ledare..nn.1)?

        Core:                   [(1, 'Activity'), (2, 'Governed'), (3, 'Leader'), (4, 'Role')]
        Peripheral:             [(5, 'Degree'), (6, 'Depictive'), (7, 'Descriptor'), (8, 'Domain'), (9,
'Duration'), (10, 'Manner'), (11, 'Means'), (12, 'Time'), (13, 'Type')]

Skriv en siffra mellan 1 och 13. Om du är osäker, skriv r för "ram" för att öppna en webbsida med
dokumentationen för ramen du valt. Om du vill välja en ny ram, skriv å för ångra: r
Skriv en siffra mellan 1 och 13. 2

        LO-LEDAREN
        Huvudled: ledare..nn.1
        Ram: Leadership
        Modifierare: LO..pma.1
        Roll: Governed
```

Frame suggestions based on SALDO entries and primary descriptors

Chosen frame

FEs of the chosen frame

Figure 5: Annotation interface: second half.

Appendix B: Test compounds

Clothing	Descriptor	*finklänningen* *gangstersvid*
	Material	*bomullspullover* *denimskjorta* *gabardinkjolen* *linnejacka* *linnekavaj* *näthandskarna* *sidendräkt* *sidensjalarna* *ullbyxor* *ylletröja*
	Style	*armépälsen* *militärbyxor* *platåskor* *pyjamasbyxorna* *tropiksvid*
	Subregion	*skörtklädnaden*
	Use	*jaktväst* *snökåpor*
	Wearer	*cowboystövlar*
Containers	Construction	*gallerlådor*
	Contents	*brödlådor* *ginglaset* *kissburken* *murbrukstråg* *terpentinflaskor*
	Material	*bleckfat* *kanvasväska* *lergodsskålen* *läderhölster* *nylonkasse* *pappväskor* *sidenhandväskan* *silverskrinet* *träaskar* *trälådan*
	Owner	*kantinkärl* *tiggarskålar*
	Type	*läkarväska* *pistolkolven* *sherryfat* *torgväska*

Containers	Use	*bärväska* *påfyllningskärlen* *resekista* *skolväskor* *whiskykaraff*
Food	Constituent_parts	*lammgryta* *nudelsoppa* *skinkstek* *tomatsoppa*
	Descriptor	*gästabudsäpple* *tolvgröten* *uppvisningsbakelser*
	Food	*fiskkebab* *sköldpaddsbiffar*
	Type	*aprikospeppar* *fågelköttet* *kalvtunga* *limefrukt* *steksås*
Furniture	Descriptor	*favoritfåtölj* *kaffekoppssäng*
	Function	*ledarbordet* *läsbordet* *medicinskåp* *patientstolen* *presentbordet*
	Material	*korgsoffa* *lönnbordet* *skinnstol* *teakbordet* *träbänkarna*
	Place	*balkongstol* *matsalsstolar*
	Relative_location	*skåphyllan*
	Type	*högskåp* *stolpsäng*
People_by_vocation	Contract_basis	*säsongsarbetare*
	Descriptor	*kändiskocken* *älsklingsservitris*
	Employer	*TASS-korrespondenten* *arbetarpartipolitiker*
	Ethnicity	*navaho-krigares*
	Persistent_characteristic	*upplysningsförfattaren*

People_by_vocation	Place_of_employment	*ambassadtjänsteman* *lagerarbetare*
	Rank	*andrepiloten*
	Type	*Engelsklärarinnan* *Generalinspektören* *Hingstföraren* *fiskhandlarens* *gatuunderhållare* *juridikprofessorn* *maskinreparatör* *programplanerarna* *statspolisen* *trädgårdsläraren* *videoteknikern*