## Examensarbete

## Swedish-English Verb Frame Divergences in a Bilingual Head-driven Phrase Structure Grammar for Machine Translation

Sara Stymne

LIU-KOGVET-D--06/08--SE 2006-06-07

## Swedish-English Verb Frame Divergences in a Bilingual Head-driven Phrase Structure Grammar for Machine Translation

Sara Stymne

LIU-KOGVET-D--06/08--SE 2006-06-07

Handledare: Lars Ahrenberg

Examinator:Lars Ahrenberg

## **Abstract**

In this thesis I have investigated verb frame divergences in a bilingual Head-driven Phrase Structure Grammar for machine translation. The purpose was threefold: (1) to describe and classify verb frame divergences (VFDs) between Swedish and English, (2) to practically implement a bilingual grammar that covered many of the identified VFDs and (3) to find out what cases of VFDs could be solved and implemented using a common semantic representation, or interlingua, for Swedish and English.

The implemented grammar, BiTSE, is a Head-driven Phrase Structure Grammar based on the LinGO Grammar Matrix, a language independent grammar base. BiTSE is a bilingual grammar containing both Swedish and English. The semantic representation used is Minimal Recursion Semantics (MRS). It is language independent, so generating from it gives all equivalent sentences in both Swedish and English. Both the core of the languages and a subset of the identified VFDs are successfully implemented in BiTSE. For other VFDs tentative solutions are discussed.

MRS have previously been proposed as suitable for semantic transfer machine translation. I have shown that VFDs can naturally be handled by an interlingual design in many cases, minimizing the need of transfer.

The main contributions of this thesis are: an inventory of English and Swedish verb frames and verb frame divergences; the bilingual grammar BiTSE and showing that it is possible in many cases to use MRS as an interlingua in machine translation.

## Acknowledgments

First and foremost I would like to thank my supervisor, Professor Lars Ahrenberg at NLPLab at the Department of Computer Science at Linköping University. Thanks for being a great support throughout my work, from helping me find the subject of the thesis, through many interesting and giving discussions, to your many useful opinions on the essay itself.

I would also like to thank everyone in the DELPH-IN collaborations who helped me with the LKB and the Matrix, especially to Stephan Oepen for debugging my grammar after an LKB update.

Another thanks to everyone at the grammar engineering workshop at the 21st Scandinavian Conference of Linguistics in Trondheim for encouraging me on my work and for many interesting discussions.

Yet another thanks to Martin for your support and for always being there. And also for all the times you read through my drafts and for your valuable comments on them!

Finally, thanks to all my friends, family and colleagues who at one time or other helped me out or simply cheered me up during this time.

Sara Stymne Linköping, March 2006

## Table of Contents

1	Intr	oduction
	1.1	Objectives
		1.1.1 Delimitation
	1.2	Method
	1.3	Conventions
	1.4	Thesis outline
2	The	oretical background
	2.1	Basic grammatical concepts
	2.2	Translation divergences
	2.3	Machine translation
	2.4	Head-driven Phrase Structure Grammar
		2.4.1 Typed feature structures and attribute value matrices
		2.4.2 The grammatical theory
3	Res	ources and frameworks 2
	3.1	Linguistic Knowledge Builder
		3.1.1 Type definition language
	3.2	Minimal Recursion Semantics
	3.3	The LinGO Grammar Matrix
	3.4	The DELPH-IN MT architecture
		3.4.1 Semantic transfer
		3.4.2 Multilingual grammar design
4	Obj	ectives revisited 3
5	Cate	egorization of verb frames 39
-	5.1	Verb frame divergence categories
	J	5.1.1 Structural divergences
		5.1.2 Conflational divergences
		5.1.3 Head inversion divergences

		5.1.4	Categorial divergences	
		5.1.5	Syntactic divergences	 42
6	The		E grammar	43
	6.1	Basic of	design	 43
	6.2	The co	ore grammar	 43
		6.2.1	Constructions for language	 45
		6.2.2	Verb phrases	 45
		6.2.3	Noun phrases	 50
		6.2.4	Empty and contentive constituents	 52
		6.2.5	Modifiers	 53
		6.2.6	Polar questions	57
7	Solu	tions f	For divergence problems	59
	7.1		ural divergences	 59
		7.1.1	Reflexives	60
		7.1.2	Empty prepositional complements	62
		7.1.3	Particles	63
		7.1.4	Infinitive markers	65
	7.2		tional divergences	66
	•	7.2.1	Understood optional complements	68
	7.3		inversion divergences	69
		7.3.1	Adverb modifiers	70
		7.3.2	Aspectual particles	
	7.4		prial divergences	74
	7.5	_	etic divergences	
	1.0	7.5.1	Ditransitive alternations	
			Distansierve discrimations	
8		ussion		77
	8.1		ingual grammars	
	8.2		as interlingua	
	8.3	BiTSE	and the Matrix	 80
	8.4	Contri	butions	 80
	8.5	Future	e work	 81
	8.6	Conclu	asion	 81
Re	eferen	ices		83
Α	Verb	frame	es in English and Swedish	89
			ntences and the translations proposed by BiTSE	93

# List of Figures

2.1	Non-clausal arguments of English verb frames	8
2.2	Clausal arguments of English verb frames	8
2.3	MT at different depth of analysis	14
2.4	Typical HPSG lexical entry for the word "put"	16
2.5	Typical TFS for the simple phrase "Bo swims"	17
2.6	TFS for the head feature principle	18
2.7	Multiple inheritance hierarchy	19
3.1	Screenshots of LKB	22
3.2	Example of unification	23
3.3	Example TFSs in TDL and AVM notation	24
3.4	MRS for "every big horse sleeps" using AVM-notation	26
3.5	Type hierarchy for some basic Matrix types	28
3.6	Basic Matrix sign	29
3.7	Containment relations for verbs in the Matrix (based on Hellan, 2003)	30
3.8	Matrix types for lexical rules	31
3.9	Inflectional rule and type from BiTSE	32
3.10	Matrix $mrs$ and $hook$ types	32
3.11	Translation equivalence from Copestake et al. (1995)	34
	Basic MRS transfer rule from LOGON (Lønning et al., 2004)	34
3.13	Example of a MRS transfer rule in LOGON (Lønning et al., 2004)	35
6.1	Examples of types used to constrain language in BiTSE	46
6.2	Type hierarchy for basic verbs	47
6.3	Type and lexical entry for transitive verbs like "hunt"	47
6.4	Types and rules for head-complement phrases	49
6.5	Type hierarchy for nouns	51
6.6	Base types for verbs with an empty or a contentive complement	53
6.7	Matrix type for intersective modifiers	54
6.8	BiTSE type for adjective lexemes	55
6.9	Partial type for preposition lexemes	56

6.10	MRS for the question "Sover Kim?" ("Does Kim sleep?")	58
7.1	BiTSE type for pure reflexive verbs	61
7.2	Alternative lexical entries for a verb with an empty prep. complement	63
7.3	BiTSE type for intransitive phrasal verbs	65
7.4	MRSs for sentences containing a raising verb and a scopal adverb	71
7.5	Matrix and alternative types for scopal modifiers	72

## List of Tables

2.1	Lexical-semantic MT Divergences (based on Dorr, 1994)	11
5.1	Structural divergences	40
5.2	Conflational divergences	
5.3	Head inversion divergences	41
5.4	Categorial divergences	42
5.5	Internal divergences	42
6.1	Examples of empty and contentive constituents	52
7.1	Examples of Swedish particles of different word classes	64
7.2	The meaning of "gå" ("go") with different particles	66
7.3	Ditransitive verbs with different valence patterns	76
8.1	Size of the different language parts of BiTSE	77
8.2	VFDs and their current coverage	
A.1	Verb frames in English	89
	Verb frames in Swedish	
В.1	Sample sentences and the translations of them proposed by BiTSE	93

## List of acronyms and terms

AVM Attribute value matrix

BiTSE Bilingual Translation grammar for Swedish and English

CBL Constraint-based lexicalist grammars
DELPH-IN Deep Linguistic Processing with HPSG

EP Elementary predication

HPSG Head-driven Phrase Structure Grammar

LKB Linguistic Knowledge Builder – grammar development platform Matrix LinGO Grammar Matrix – language-independent starter kit

MRS Minimal Recursion Semantics –

underspecified, flat meaning representation language

MT Machine translation

NLP Natural language processing

SL Source language

TFS Typed feature structure

TL Target language

VFD Verb frame divergence

## 1 Introduction

Translation between two languages is a very difficult task where many different types of knowledge are needed, such as lexical and grammatical knowledge and knowledge about the world. It is a difficult task for humans, but machine translation (MT), when a computer translates, is even more difficult since computers somehow have to be given all the knowledge they need. One way to give computers grammatical knowledge is to encode the grammars of the languages one wants to translate between in a machine-readable way. This is in itself a difficult task for one language, since grammars by nature are very complex. To achieve MT this is still not enough, since grammars of different languages are different, and have to be coordinated somehow.

English and Swedish are two languages that are quite closely related, and thus their grammars have many similarities. Even so there are many things that are different between Swedish and English. One such case is that definite form is formed by adding the word "the" in English while it is formed by adding a suffix in Swedish. Another case is differences in verb frames, verb frame divergences (VFDs), which is the focus of this study.

A verb frame consists of a verb and its arguments. A verb frame divergence is when two verb frames with the same meaning have different structures, some examples of which can be seen in (1).

(1) a.  $Jag\ kommer\ ih\mathring{a}g$  det.  $I\ remember\ it$ . I come in mind it.

b. Jag vill ha det. I want it.

I want have it.

c. Jag hämnas på honom. I revenge myself on him.

I revenge on him.

In this thesis I investigated machine translation between Swedish and English with a focus on verb frames divergences. In order to do this I classified VFDs between Swedish and English, and constructed a computational grammar to encode them. This grammar functions as the core of a small MT system, where translation takes place by parsing a Swedish or English sentence to a common semantic level from which all equivalent sentences in both languages can be generated.

For this study I have adopted the models and frameworks of the DELPH-IN<sup>1</sup> collaboration. Practically that means that I am using the linguistic frameworks Head-driven Phrase Structure Grammar (HPSG) and Minimal Recursion Semantics (MRS). I am also using the Linguistic Knowledge Builder (LKB) grammar developing environment and the LinGO Grammar Matrix grammar base.

## 1.1 Objectives

The general objective of this mainly explorative thesis was to investigate machine translation between Swedish and English with a focus on verb frame divergences. The objectives can be coarsely divided into three parts, a descriptive, a practical and a machine translation theoretical.

In the descriptive part the goal was to describe and classify VFDs between Swedish and English. I also tried to find solutions of how to solve these differences in a machine translation system using a common semantic representation, or interlingua, for Swedish and English. That included finding a semantic representation for each VFD and mapping syntactically different clauses to this representation.

The core of the practical part was to construct BiTSE, the Bilingual Translation grammar for Swedish and English. It is a HPSG-based grammar covering the core of the languages as well as many of the VFDs that I identified possible solutions for. Constructing this grammar allowed me to experiment with different possible VFD-solutions. The grammar is based on the LinGO Grammar Matrix, a language independent grammar base. The Matrix is under constant development, and thus BiTSE could contribute to the Matrix project by finding out how well Swedish fits into the Matrix. I also wanted to investigate how well a grammar design with two languages in one grammar would work. This design also shows the similarities and differences between Swedish and English in a clear way.

No complete MT end-user application was built, but the ideas and the grammar can be integrated into a complete MT system in the future. The grammar could also be used in other areas of natural language processing (NLP) where a semantic representation of Swedish and/or English are needed.

In the machine translation theoretical part I tried to see how far the MT approach used would work. I wanted to find out in what cases the approach with a common semantic representation (interlingua) would not work, or be cumbersome, and at what point some other approach, most likely semantic transfer, would be better or even necessary.

The main objectives of the thesis can be summarized in the following points:

<sup>&</sup>lt;sup>1</sup>Deep Linguistic Processing with HPSG, see www.delph-in.net

#### • Descriptive

- Collect and classify verb frame divergences between Swedish and English

#### • Practical

- Implement a parallel Matrix-based grammar for Swedish and English covering many of the described verb frame divergences
- Find out what parts of the grammar are the same for Swedish and English,
   and what the differences are

#### • Machine translation theoretical

- Find out what cases of verb frame divergences can be handled by a common semantic representation (MRS)
- Find the border where this approach does not work any more and some other approach would be needed

#### 1.1.1 Delimitation

The translation was done on sentence level and the main focus was on propositions in active form, although some attention was also given to polar questions. Issues like topicalisation and negation were not handled. Another phenomenon that was not handled is non-local dependencies, such as WH-questions and relative clauses, that usually receives gap analyses.

Ambiguity was not a focus. Sentence pairs that are considered equivalent have been chosen without considering alternative translations. In the cases where several translations are proposed by BiTSE for the same sentence no ranking between the alternatives was made.

There are many other complexities of machine translation that were not handled in this study such as efficiency and robustness. The only knowledge in the system is syntactic/semantic and lexical, no other knowledge such as world knowledge or discourse factors were used.

### 1.2 Method

The work was to a large extent explorative and experimental in nature, and the later parts of it depended on the findings in the previous parts. The work could be divided into three parts: empirical investigation, implementation, and study and development of linguistic theory.

The empirical part was to identify and classify VFDs between English and Swedish. In order to do this I created lists of all verb frames for Swedish and English respectively that I could identify. For the Swedish verb frames I used Gambäck (1997) as a starting point, and also used a grammar book (Jörgensen & Svensson, 1987). For the English verb frames I used grammar books (Svartvik & Sager, 1977; Quirk et al., 1985), a classification of English verb alternations (Levin, 1993) and a dictionary (Hornby et al., 1963) with a useful grammar part. I also used introspection, and scanned a number of English and Swedish texts to find verb frames therein. I then contrasted the identified verb frames and tried to find the possible English alternative frames for each Swedish frame. For the classification I also used existing classifications for divergences (Dorr, 1993; Catford, 1965) which I compared my findings to.

In order to implement solutions for the VFDs I needed a base grammar so I had somewhere to fit them into. I thus constructed BiTSE, the bilingual translation grammar for Swedish and English. BiTSE is based on the LinGO Grammar Matrix (Bender et al., 2002) and based on the assumptions of the Matrix and generally accepted linguistic analyses for the phenomena covered by it, mainly from work within HPSG (e.g. Pollard & Sag, 1994; Bender, 2004, 2005; Sag et al., 2003) but also from general grammar books (e.g. Svartvik & Sager, 1977; Jörgensen & Svensson, 1987) and from work within other frameworks such as Lexical Functional Grammar (e.g. Toivonen, 2003). BiTSE was constructed in three different parts, one that is common for Swedish and English, and one separate part for each language, clearly showing the common and distinctive parts between the languages.

The VFD categorisation was used as a base for the work with finding solutions to VFDs. I also based the solutions on existing linguistic theories, from the same types of sources as for the core grammar. I also studied previous work on machine translation, particularly work where MRS had been used for semantic transfer (e.g. Lønning et al., 2004; Copestake et al., 1995). The process was iterative, and during the implementation more VFDs were found and the classification had to be revised. During this process I could also find out what was easy and hard to implement.

### 1.3 Conventions

Typed feature structures (TFSs) are frequently used a lot in this report. They are shown using two formats, AVM, described in Section 2.4.1 and TDL described in Section 3.1.1. TFSs contains types, features and values. Types are written in **bold italics**, features in SMALL CAPITAL LETTERS and values in *italics* in the text. In LKB rules are used; they are written in *italics*.

Standard conventions for linguistic examples have been used with grammaticality judgements shown by an asterisk marking a non-acceptable example, as "\*he sleep", and a question mark marking doubtful examples.

Swedish examples have been glossed to English when the structure is different than in the English translation. A natural translation is also given. In the glossing each word is translated with its closest English counterpart, and when no possible translation of a word have been found the word class in capital letters is used, for an illustration see (2).

(2) a. Mina krafter stod mig bi.

My powers stood me PART.

My strength held out.

### 1.4 Thesis outline

**Chapter 1** is this introduction.

**Chapter 2** presents the theoretical background of the thesis: some basic grammatical concepts, translation divergences, machine translation and Head-driven Phrase Structure Grammar.

**Chapter 3** introduces the DELPH-IN resources and frameworks that are used, and describes the DELPH-IN approach to MT.

**Chapter 4** revisits the objectives, further specifying them using the concepts from chapter 2 and 3.

**Chapter 5** describes the results of the verb frame categorisation.

**Chapter 6** describes the design and the core of the BiTSE grammar that was developed.

**Chapter 7** discusses solutions for the verb frame divergences that were found and also describes the implementations that were incorporated into BiTSE.

**Chapter 8** contains discussions of the work in this thesis and the implications of it, the contributions of the thesis, some proposals for future work and the conclusion.

**Appendix A** contains the classification of verb frames for Swedish and English.

**Appendix B** contains some examples of sentences and their translations as proposed by BiTSE.

## 2 Theoretical background

This chapter provides a theoretical background by introducing some basic grammatical concepts, translation divergences, machine translation and the grammatical theory Head-driven Phrase Structure Grammar.

## 2.1 Basic grammatical concepts

There are two basic ways to connect one constituent with another, either as an argument or as a modifier. Arguments, which are the subject and the complements, are needed to make the meaning of a verb complete (Svartvik & Sager, 1977). The arguments of a constituent are called the argument frame, or valence. For verbs this is the verb frame.

The possible arguments for English verbs, loosely based on Quirk et al. (1985), are shown in Figure 2.1 and 2.2, with a typical example of each type of verb. Figure 2.1 shows non-clausal arguments. The examples of possible combinations of arguments are not exhaustive. Figure 2.2 shows the possible types of clausal arguments. These can also be combined with particles and reflexives. Some verbs have several different verb frames, the verb "want" for instance can have either an object, "I want a book", or an infinitival clausal complement, "I want to swim". The possible arguments for Swedish are quite similar, the non-clausal complements have the same elements, but there are differences in the way the different complements can be combined. There are also some differences in the possible clausal complements. Appendix A contains a full inventory of the identified verb frames in Swedish and English.

Modifiers, or adjuncts, are constituents that can be appended to other constituents adding new information to its meaning. Sag et al. (2003) notes that it is hard to describe the distinction between modifiers and arguments formally, but gives the following intuition: "complements refers to the essential participants in the situation that the sentence describes, whereas modifiers serve to further refine the description of that situation" (p. 98). Modifiers are usually divided into two groups, intersective and scopal, which differs in the scope. Intersective modifiers always have the same scope as the constituent it modifies, and scopal modifiers can have higher scope.

Verbal modifiers are often called adverbials, although this can also include the obligatory adverbials that are arguments of some verbs. Svartvik & Sager (1997) describes English

subj	verb	part	refl	obj1	obj2	pred	obl.adv.	<pre>prep.compl</pre>
I	sleep	_	_	-	-	-	-	-
I	threw	up	_	-	-	-	-	-
I	saw	_	_	him	_	-	_	_
I	depen	d-	_	_	_	_	_	on him
It	seems	_	_	_	_	ready	_	_
I	live	_	_	_	_	_	here	_
I	sent	_	_	him	it	_	_	_
I	told	_	_	it	_	_	_	to him
It	made	_	_	him	_	sick	_	_
I	put	_	_	it	_	_	down	_
I	perju:	red	myself	_	_	_	_	_
I	reven	ged	myself	-	_	-	_	on him

Figure 2.1: Non-clausal arguments of English verb frames

finite	bare	hope it works
	that	think that it works
	WH	guess <i>why it works</i>
non-finite		
subj. less	WH-inf	learn what to say
	to-inf	continue $\it to work$
	pres. participle	enjoy ${\it working}$
	bare inf	$\verb"can" work"$
with subj	to-inf	want $\mathit{him}$ to $\mathit{work}$
	pres. participle	dislike him working
	bare inf	see $\mathit{him}$ work
	past. participle	keep <i>him posted</i>

Figure 2.2: Clausal arguments of English verb frames

adverbials as a diverse class, which can be realised by different constituents and have different functions as well as having different positions in a clause. They divide the meaning of modifiers into manner, place, time, degree, modality and contingency. Examples of these can be seen in (3). Modifiers thus expresses something that is outside of the main meaning of the verb it modifies. Modifiers can in some cases be superficially like arguments.

- (3) a. She sings beautifully (manner)
  - b. I walked on the street (place)
  - c. I played football yesterday (time)
  - d. She nearly fell (degree)
  - e. He reluctantly washed the dishes (attitude)
  - f. I yet managed to succeed (contingency)

## 2.2 Translation divergences

Translation takes place from a source language (SL) text to a target language (TL) text. In many cases there are differences between the SL text and the TL translation. A translation equivalence is when a pair of texts from a translation have neither syntactic nor semantic differences. Barnett et al. (1991) divides the non-equivalent cases into two groups: divergences and mismatches. In a translation divergence both texts have equivalent semantics but different syntax. A translation mismatch occurs when the semantics of the source and target text are different. Divergences and mismatches are usually identified at smaller units than a complete text, generally on sentence level.

One of the first to categorize translation divergences, or shifts in his terminology, was Catford (1965) who divides them into two groups, level shifts and category shifts. A level shift occur when there has been a change between the level on which a certain feature has been realised. This often happens between the syntactic and lexical levels, as in (4) where English realises definite form on the syntactic level by adding the determiner "the" and Swedish realises it on the lexical (morphological) level by adding the suffix "-et".

Catford defines category shifts as "departures from formal correspondence" (p. 143). He divides them into four types:

structure-shifts, changes in word order

Idag simmar jag Today I swim
Today swim I
(Adv V NP) (Adv NP V)

class-shifts, changes of syntactical class

Jag hälsar på henne I greet her
I greet at her
(prep. obj) (direct object)

unit-shifts, changes in number of syntactical units

 $Jag\ kommer\ ih\mathring{a}g \qquad det \qquad \qquad I\ remember\ it$  I come in.mind it (verb+particle) (single verb)

intra-system-shifts, happens when two languages have approximately the same systems for a specific feature, but different distributions of it.

sax scissors (singular) (plural)

Since Catford there have been many other attempts at classifying translation divergences. Merkel (1999) divides shifts into obligatory and optional. Obligatory shifts are those that are necessary due to the difference in grammar between languages, such as the ones classified by Catford. Optional shifts are those that translators make that are not forced for grammatical reasons, but chosen anyway for various reasons such as getting a better text flow or adapting the text to the target language culture. Optional shifts also include mismatches. Some examples of optional shifts from Merkel are (5a) (p. 184) which is an addition and (5b) (p. 181) which is a clause to non-clause.

(5) a. Datasheets are the easiest type of subforms

Datablad är den **snabbaste** och enklaste typen av underformulär

Datasheets are the fastest and easiest type of subforms

b. But suddenly the music stops and a terrorist bomb is reported

Men plötsligt tystnar musiken för en rapport om en terrorbomb

But suddenly silences music.the for a report about a terror.bomb

(Dorr, 1993, 1994) has classified obligatory divergences specifically for an MT application. She divides divergences into two main groups, syntactic and lexical-semantic. Syntactic divergences are characterized by grammatical properties within each language that are mainly independent of the actual lexeme used, while lexical-semantic are determined lexically. Both types of divergences are further subdivided. Some examples of syntactic divergences given by Dorr are: constituent order, long distance movement and null subject.

The lexical-semantic divergences are divided into seven categories, which can be seen in Table 2.1. Dorr also suggests systematic solutions for these classes of divergences.

Divergence	Description	Example			
Thematic	The theme is	Me gusta Mary I like Mary			
	realised as different	Me pleases Mary			
	constituents				
Promotional	The head is	John brukar sova John usu			
	switched up	John tends sleep	sleeps		
Demotional	The head is	Ich esse gern	I like to eat		
	switched down	I eat likingly			
Structural	Different structures	Jag hälsar på henne	I greet her		
	of the same	I greet at her			
	constituent				
Conflational	Incorporation of	Yo le di puñaladas a John	I stabbed John		
	necessary	I gave knife-wounds to			
	arguments of a	John			
	given action				
Categorial	Change of category	Ich habe hunger	I am hungry		
	of a constituent	I have hunger			
Lexical	Use of	John forzó la entrada al John broke			
	non-equivalent	cuarto	the room		
	lexemes	John forced the entry to			
		the room			

Table 2.1: Lexical-semantic MT Divergences (based on Dorr, 1994)

Both promotional and demotional divergences involve a main verb corresponding to an adverbial modifier, i.e. a head inversion. There is however a slight difference in which element *triggers* the head inversion. In a promotional divergence a main verb is always replaced by a modifier in the other language, the Swedish main verb "brukar", for instance,

is always translated with the English adverb "usually", but "usually" does not have to be translated with a verb in Swedish, an adverb like "oftast" is also an option. In a demotional divergence a modifier in one language is always replaced with a main verb in the other, in the example in Table 2.1 the German adverb "gern" is always translated by English "like", but "like" does not have to be translated by "gern", "I like cars" can be translated as "mir gefällt autos" ("me pleases cars") for instance.

Lexical divergences most often occur together with other divergences<sup>1</sup>. A lexical divergence means that a verb that does not literally correspond to the source language verb is chosen, but the verb phrases are still equivalent due to other differences between them, such as a conflated argument in one language. In Table 2.1 lexical divergences also occur in the examples of thematic, conflational and categorial divergences.

A difference between classification schemes is if they see the divergences as directional or not. Merkel's terms, e.g. deletion and addition, are clearly directional, going from English to Swedish in (5a) is an addition, but going from Swedish to English is a deletion. Dorr's and Catford's categories on the other hand are non-directional, they refer to a general difference between a sentence pair, regardless of which language is the source language.

Verb frame divergences are divergences where only the elements that are part of verb frames are of interest.

### 2.3 Machine translation

Hutchins (2003) defines machine translation (MT) as: "computerized systems responsible for the production of translations with or without human assistance" (p. 501). Machine translation has been a field of research since the late 1940s, at what time researchers were very optimistic of quickly being able to get fully automated high quality machine translation, FAHQMT Arnold et al. (1994). It turned out to be more difficult then they thought.

MT is often divided into two different branches, unassisted and assisted MT. Unassisted MT is when the computer performs the whole translating process on its own with no involvement from humans. The ultimate goal of this branch used to be FAHQMT. However, as Kay (1996) puts it: "Few informed people still see the original ideal of fully automatic high-quality translation of arbitrary texts as a realistic goal for the foreseeable future." There are no clear replacement of this goal, but one suggestion is "understandability quality" (Sågvall Hein, 2005, my translation).

Today there are no all purpose systems of ultimate quality, but there are systems in limited domains such as Taum-meteo for weather reports which has been operational since 1977

<sup>&</sup>lt;sup>1</sup>Dorr claims that they always occur together with other divergences, if only lexical-semantic divergences are considered.

(Isabelle & Bourbeau, 1985). There are also MT-systems that are good enough for gisting, i.e. to get an overall picture from an imperfectly translated text.

Assisted MT involves a human in some stage of the translation process. It can be further subdivided into two branches, human aided machine translation, where a person improves the quality of a translation, usually by pre- and/or post-editing the text and machine aided human translation where a human translator gets support by a computer e.g. by a translation memory.

MT systems can contain different numbers of languages, they can be bilingual or multilingual. Another distinction is between unidirectional and bidirectional systems. Unidirectional systems are usually bilingual and can only translate one way, between one source language and one target language only. Bidirectional systems can translate in both directions. Bidirectional systems can be reversible, i.e. use the same modules for translating both ways between the languages, or they can consist of several unidirectional systems put together.

MT systems are also classified depending on architecture. The classical one, which is still common, is rule-based MT, where translation is based on linguistic rules, somehow encoded for computers. The other main field is empirical MT which includes statistical MT and example-based MT, which are both based on parallel corpora. There are also MT frameworks which do not fit into these two groups neatly, such as neural-net-based MT (e.g. (Ishikawa & Sugimura, 1992)). Rule-based and empirical MT systems have different strengths, something that have been taken advantage of in hybrid systems (Somers, 2003). Hybrid systems mix different architectures so that they all deal with the problems they are most suited to. Another type of hybrid system is multi-engine MT, which runs the text through different MT systems and chooses the translation it ranks as best (e.g. Hogan & Frederking, 1998).

Rule-based MT can be done at different depths of analysis broadly divided into *interlingua*, transfer and direct translation, illustrated in Figure 2.3.

Interlingua based systems uses a common language independent representational level, the interlingua. Translation takes place in two stages, The SL is analysed to the interlingua, from which the TL is generated. The interlingua can be very different in different systems. Hutchins (2003) mentions interlinguas based on: logical artificial languages, natural auxiliary languages such as Esperanto, general semantic primitives and supposedly universal vocabularies. The obvious advantage of interlingua systems is that it is a good structure for a multilingual system and that it is simple to add new languages to the system, only modules to go from the new language to the interlingua and back are needed. The main disadvantage is that it is hard to find a general interlingua for many languages. It is however often possible to find an interlingua for a limited set of languages, which might be good enough in many cases.

A transfer-based system operates in three stages. First the SL is analysed to a language

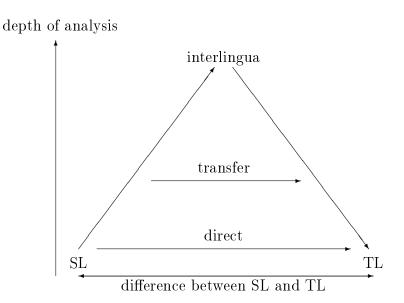


Figure 2.3: MT at different depth of analysis

dependent representation. Then transfer takes part, which changes the SL representation into a TL representation. The last stage is generation of the TL from the TL representation. The language dependent representation can be either syntactic or semantic. Syntactic transfer was common in early MT systems (Hutchins, 2003). An example of a syntactic transfer system between Swedish and English is MATS (Sågvall Hein et al., 2003). An example of semantic transfer is the LOGON project (Lønning et al., 2004), for translation between Norwegian and English, where MRS is used for transfer. A transfer system for two languages usually has three dictionaries, one monolingual each for analysis of the SL and generation of the TL, and one bilingual transfer dictionary relating the two languages (Hutchins, 2003). A transfer architecture is generally easier to build than an interlingua system, but it is harder to add more languages to the system since that means that besides analysis and generation parts for that language a different transfer part will have to be built for every other language in the system, and in case the transfer is not bidirectional two transfer parts for each language pair are needed.

One big advantage of both transfer and interlingua systems is that the output tends to be grammatical, since it has a generation part for the TL, based on the TL grammar. A disadvantage is that if the system cannot analyse a sentence it might not be able to offer any translation for it, i.e. the system might not be robust. Interlingua and transfer are often called deep translation, since the linguistic analysis is deep. This is opposed to

direct and empirical translation which are often called shallow.

In a direct translation system the source language is directly transformed into the output language with no, or only shallow analysis, for instance by translating the words and rearranging them through reordering rules for the language pair. Direct systems typically consists of a bilingual dictionary and a single program that does the analysis that needs to be done (Hutchins, 2003). An advantage of this is that it is robust, i.e. always gives some kind of output. Unfortunately the output is ungrammatical in many cases because it is not based on a grammar of the TL, but on rearranging SL text.

Since the mid-1980s another branch of rule-based MT has emerged, unification-based or constraint-based MT (Hutchins, 2003). These systems are based on unification-based or constraint-based theories of grammar, such as Head-driven Phrase Structure Grammar, which will be described in the next section. These system are also often lexical-based, which means that most information is part of the lexicon, or dictionaries, instead of part of external grammatical rules.

### 2.4 Head-driven Phrase Structure Grammar

Head-driven Phrase Structure Grammar (HPSG) was first described by Pollard & Sag (1994), based on their previous work. It has since then been applied to a wide range of languages and a wide range of specific phenomena by a multitude of researchers, and in this process it has been both extended and revised.

HPSG is one of several grammars that belong to the group of constraint-based lexicalist grammars (CBL). The main characteristics of CBLs are that they are surface oriented, constraint-based and strongly lexicalist (Sag et al., 2003).

- A surface oriented grammar only uses information related to the words of a sentence in surface order and information that can be derived from this; no additional abstract structures are used.
- A constraint-based grammar expresses the principles of the grammar as constraints on lexical entries and grammar rules that interact. All structures are of the same sort unlike in many other types of grammars where there are transformational rules that derive one type of structure from another.
- A strongly lexicalist grammar holds most of the grammatical and semantic information in the lexical entries. These correspond in turn to each word in a string which is what drives the derivation of the syntactic and semantic structure of the sentence.

Except for HPSG other well known CBL theories of grammar are Categorial Grammar, Construction Grammar, Lexical Functional Grammar and Dependency Grammar.

### 2.4.1 Typed feature structures and attribute value matrices

HPSG uses typed feature structures (TFSs) to model linguistic objects. This section will describe a way to show TFSs, using attribute value matrix (AVM) format to depict it. In the next section the meaning and use of TFSs will be described. Figure 2.4 shows an example of a typed feature structure in AVM format. TFSs consists of features, or attributes, written with capital letters on the left hand side, which can have different values. The values can be either atomic, and are then written with italics to the right of the corresponding feature like non for the AUX(iliary) feature, or complex, that is in turn another TFS, like the feature for CONT(ent). The fact that the feature structures are typed means that each feature has a type, written in italics to the top left of the feature structure. The type is sometimes left out, as in the TFS that is the value of VAL. Two values can be structure-shared or co-indexed, shown in small boxes, i.e. I, which means that they share the same token value<sup>2</sup>. Abbreviations for phrases are used, in Figure 2.4 NP and PP, which are abbreviations of TFSs modelling a noun phrase and a preposition phrase. Lists are notated "< >" and difference lists are notated "<!!>". The symbol "|" is used to show a path; "SYNSEM LOCAL local" is equivalent to "SYNSEM [LOCAL local]". Partial feature structures, where features that are irrelevant when illustrating a specific point are removed, are often used.

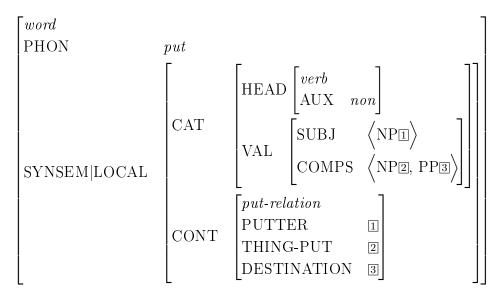


Figure 2.4: Typical HPSG lexical entry for the word "put"

<sup>&</sup>lt;sup>2</sup>This is often referred to as unification informally

### 2.4.2 The grammatical theory

The basic unit in HPSG is the sign. All linguistic objects, e.g. words, phrases and sentences, are signs. A sign minimally consists of the two attributes PHON, which is some representation of the phonetics of a sign, and SYNSEM, which contains information of the syntax and semantics of the sign. As an illustration a partial lexical sign for the word "put" can be seen in Figure 2.4. PHON is here represented simply as a list with the text string "put", but it could be elaborated further if needed. SYNSEM information is divided into LOCAL and NONLOCAL, for simplicity the latter is not shown. LOCAL is in turn divided into CAT(egory) and CONT(ent). CAT contains syntactic information about the sign and is divided into HEAD, information about the sign itself, here that "put" is a non-auxiliary verb and VAL(ence), which states what arguments a word takes. CONT constitute the semantics, it contains a semantic relation for the word in question, with the roles for the arguments of the word bound to, or structure-shared with, the different constituents of the valence.

There are also phrasal signs that in addition to the features of lexical signs have daughters (DTRS) of different types. Most phrases are headed phrases, and the head of the phrase is the head-daughter. As an illustration Figure 2.5 shows a TFS for the verb phrase "Bo swims", again with abbreviations for TFSs for phrases. Uszkoreit et al. (2000) states that a HPSG sign is "a grammatically sanctioned correspondence aligning certain phonological, syntactic and semantic information" (p. 221). In Figure 2.5 it aligns the phonological sequence "<Bo, swims>", the syntactic category sentence and the proposition that an individual named "Bo" does the act of swimming.

Figure 2.5: Typical TFS for the simple phrase "Bo swims"

Uszkoreit et al. (2000) points out that a key feature of HPSG is that all kind of linguistic

information is stored in a consistent manner, as TFSs. This is unlike many other approaches which uses for instance derivations of phonology or semantics from some skeletal syntactic structures. In HPSG words and phrases are different size constellations of the same related information. They further note that this seems to correspond to what we know about human sentence processing, it uses various information levels flexibly and interleaved, with no sign of any autonomous syntactic manipulations.

The TFSs are the models of the theory. The theory also includes a number of principles or constraints that models must fulfil in order to be accepted. Two such principles are the head feature principle and the subcategorization principle. The head feature principle states that the HEAD value of a headed phrase is structure-shared with the HEAD value of the head daughter. The subcategorization principle states that the arguments of the head daughter in a headed phrase is the concatenation of the arguments of the phrase and of the SYNSEM values of the non-head daughters. These principles can also be implemented by TFSs. Figure 2.6 shows the TFS for the head feature principle which co-indexes the HEAD value of the phrase and its head daughter.

Figure 2.6: TFS for the head feature principle

The specific feature structures for lexemes and rules are not stored for each particular entry, but organized into a multiple inheritance hierarchy. This is utilized to generalize phenomena across different categories. Figure 2.7 shows an example of this concerning part of speech and argument selection. Thus there is one basic type for every group of words that have the same features and values for e.g. head, valence and alternation patterns, and only this type and word specific information such as phonetics and semantic roles, are stored in the lexicon.

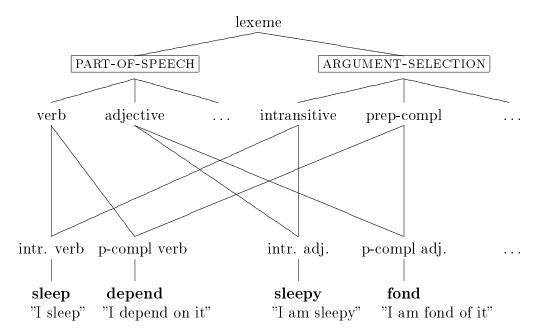


Figure 2.7: Multiple inheritance hierarchy showing part of speech and argument selection

# 3 Resources and frameworks

The Linguistic knowledge builder (LKB), a grammar development system, was used for developing BiTSE and for parsing and generation. BiTSE is based on the LinGO Grammar Matrix and the semantic representation used as interlingua is Minimal Recursion Semantics (MRS). These will be further introduced in this chapter. LKB and the Matrix are developed by the LinGO laboratory<sup>1</sup> and all three are formats and tools adopted by the DELPH-IN collaboration. LKB and the Matrix are and available as open-source. The general approach to MT used by DELPH-IN, semantic transfer on MRS, will also be discussed in this chapter.

# 3.1 Linguistic Knowledge Builder

LKB<sup>2</sup> is a development environment for typed feature structure<sup>3</sup> (TFS) grammars and lexicons. It has mainly been used within the HPSG framework, but it can also be used by other grammatical frameworks that use TFSs (Copestake, 2001). LKB is based on unification and can be used both for parsing and generation with the same grammars. The generator takes a semantic representation in form of MRS (see Section 3.2) or some similar formalism as its input. LKB is specifically useful for interactive grammar development since it has many helpful features such as consistency checks for the type hierarchy and interactive unification. It also has a useful graphical user interface with several possible views available, Figure 3.1 shows a screen shot of some of its views.

Copestake describes unification informally as "the combination of two TFSs to give the most general TFS which retains all the information which they individually contain" (2002, p. 54). If it is not possible to combine the TFSs, e.g. because some feature is not compatible, unification fails. Figure 3.2 shows the unification of two TFSs. This unification is valid if np is an ancestor of phrase, and invalid else.

Tseng (2003) notes that there are differences between theoretical HPSG, as of Section 2.4, and a practically implemented HPSG for LKB. Two of the differences he notes is the

<sup>&</sup>lt;sup>1</sup>CSLI Linguistic Grammars Online laboratory at Stanford University, see http://lingo.stanford.

<sup>&</sup>lt;sup>2</sup>Information and download available at http://wiki.delph-in.net/moin/LkbTop

<sup>&</sup>lt;sup>3</sup>LKB actually uses TDFS, which is TFS extended with defaults

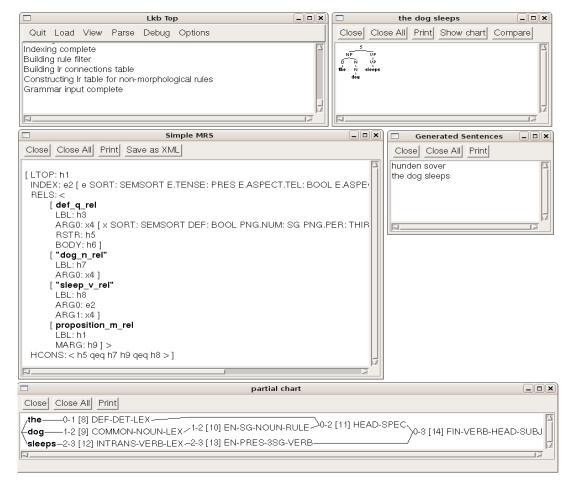


Figure 3.1: Some screenshots of LKB, from top left: main window, tree representation, MRS, generated sentences and partial chart. The last four are for the sentence "the dog sleeps"

$$\begin{bmatrix} word \\ PHON & "Kim" \\ CATEG & phrase \end{bmatrix} \quad \sqcup \quad \begin{bmatrix} word \\ CATEG & np \end{bmatrix} \quad = \quad \begin{bmatrix} word \\ PHON & "Kim" \\ CATEG & np \end{bmatrix}$$

Figure 3.2: Example of unification

absence of negation and disjunction in LKB and the fact that rules need a fixed number of daughters with a set linear order. Even though disjunction can not be used, it can in most cases be expressed through the type hierarchy (Flickinger, 2000).

The basic components of a grammar in LKB as described in Copestake (2001) are:

The type system is the core of the grammar. The types are organized in a multiple inheritance hierarchy expressing constraints on what they model.

**Lexical entries** constitute the lexicon and define a relationship between a string representation of a word and some linguistic representation of the word based on types from the type system.

**Grammar rules** are TFSs that define how phrases are obtained from words and other phrases, and how to obtain clauses from phrases. The rules are often quite simple with most of its information in the type for the rule.

**Lexical rules** are mainly used for inflection, but also for phenomena that only covers lexemes, such as subject-verb inversion and different verb frame possibilities. These are also based on types from the hierarchy.

The start structure is an instance of a type, and it states what the top level of a valid parse should be. It is often some structure equivalent to a sentence. Multiple start structures are allowed.

Parse node labels are not necessary, but they are used to name the nodes of the trees constructed in the parse, to allow the grammar writer to check the trees easier.

The grammar and lexicon consist of typed feature structures. To be able to write these conveniently it is necessary to have some description language for TFSs that can be entered with a normal text editor. It is also necessary to be able to describe inheritance. This is not possible with AVMs for instance, since it for instance uses long vertical bars and little boxes around text. LKB supports multiple description languages, for instance TDL, Type Description Language (Copestake, 2001). TDL is also used in the Matrix and in BiTSE and will be used in this report.

## 3.1.1 Type definition language

The version of TDL described here is that of Copestake (2001). Figure 3.3 shows an example of a typed feature structure in TDL notation and the corresponding AVM, except for inheritance. The TDL syntax is in many regards the same as for AVMs. The main differences are that the types are placed outside the feature structure in TDL, and that the conjunction symbol "&" is explicitly used. In AVMs conjunction is implicit. A TDL description of a type or entry starts with the name of the type, then the symbol ":=", the definition of the type, including the types it inherits from and new constraints, and the end symbol "." Co-indexation is marked by tags beginning with "#", instead of small boxes for AVMs. Paths use the symbol "." instead of "|". It is possible to add new features to types that have been defined earlier with ":+". Comments in the code files are lines that start with the symbol ";". In the figures of this report italic text will be used for comments, for better readability.

Figure 3.3: Example TFSs in TDL (top) and AVM (bottom) notation

## 3.2 Minimal Recursion Semantics

In classical HPSG, as described in Section 2.4, semantics are handled by the CONTENT value. MRS is more elaborated than the content theory and can be substituted with CONTENT. This is fully supported by LKB and MRS is the standard input format of its generator. MRS is also the semantic representation used in the Matrix.

MRS is "a framework for computational semantics that is suitable for parsing and generation and that can be implemented in typed feature structure formalisms" (Copestake

et al., 2003, p. 1). It is not a semantic theory in itself but rather a language for describing semantic structures. It can model semantics using different underlying object languages, Copestake et al. (1995) suggests predicate calculus and discourse representation theory (Kamp, 1981). It was designed to enable semantic composition using only unification of TFSs (Flickinger & Bender, 2003). It is a flat representation, but includes a treatment of scope that allows underspecification of scope.

The primary unit in MRS<sup>4</sup> is the elementary predication (EP), which is basically a single relation, and its associated arguments such as "hunt(x,y)" for the verb "hunt". That MRS is flat means that the EPs are all handled on the same level, not for instance with a tree representation. Thus a MRS structure contain a bag of EPs, i.e. a conjunction of EPs with no inherent order. To allow scope to be handled each EP has a handle, which identifies it, and quantifiers have handles as arguments. (6a) shows an example of an MRS for the sentence "every big horse sleeps" from Copestake et al. (2003). For the quantifier "every", its first argument, x, stands for the bound argument, its second argument, h1, for its restriction and the third, h2, for its body. There is an implicit conjunction between the EPs which means that (6b) is equivalent to (6a).

(6) a. h0:every(x,h1,h2), h1:big(x), h1:horse(x), h2:sleep(x)b.  $h0:every(x,h1,h2) \wedge h1:big(x) \wedge h1:horse(x) \wedge h2:sleep(x)$ 

A complete MRS structure also contains handle constraints (hcons), and a top handle. Rephrasing (6a) including handle constraints and top results in (7). Handle constraints enable underspecification of scope. They contain a bag of constraints on scope. The constraint used is called qeq, equality modulo quantifiers, and relates a handle to a label. The handle can either be directly filled by the label, or other quantifiers can appear between them. Top is a handle that corresponds to the top relation in the entire MRS.

(7) Top: h1
Rels: h0:every(x,h1,h2),h1:big(x),h1:horse(x),h3:sleep(x)Hcons: h2 qeq h3

MRSs can also be described as feature structures. Figure 3.4 shows an AVM version of (6a). The numbers assigned to the arguments are different, but that has no significance, what matters is how the handles are combined. The TFS version also has event relations for the verb and adjective, and at top level. The event carries event variables that are not shown in Figure 3.4. The event variables carry information about tense, aspect and mood. Likewise indeces of entities, usually nouns, carry information about person,

<sup>&</sup>lt;sup>4</sup>The abbreviation MRS will be used both for the semantic theory, and for individual MRS structures, i.e. the semantic representations used for a word, phrase or sentence.

number, gender and definiteness. These are not shown in Figure 3.4, but can be added to a MRS when needed. For clarity co-indexed entries will be marked both with a number and a letter indicating its type, using "h" for handle, "e" for event, "x" for entities and "i" for individuals, that is arguments that can be either event or entities.

$$\begin{bmatrix} mrs \\ \text{LTOP} & h12 & h \\ \text{INDEX} & e2 & e \\ & \begin{bmatrix} -every\_q\_rel \\ \text{LBL} & h6 & h \\ \text{ARG0} & x7 & x \\ \text{RSTR} & h9 & h \\ \text{BODY} & h8 & h \end{bmatrix}, \begin{bmatrix} -big\_a\_1\_rel \\ \text{LBL} & h10 & h \\ \text{ARG0} & e11 & e \\ \text{ARG1} & x7 \end{bmatrix}, \\ \begin{bmatrix} -horse\_n\_1\_rel \\ \text{LBL} & h10 \\ \text{ARG0} & x7 \end{bmatrix}, \begin{bmatrix} -sleep\_v\_1\_rel \\ \text{LBL} & h12 & h \\ \text{ARG1} & x7 \end{bmatrix}$$

$$\text{HCONS} \left\langle \begin{bmatrix} qeq \\ \text{HARG} & h9 \\ \text{LARG} & h10 \end{bmatrix} \right\rangle$$

Figure 3.4: MRS for "every big horse sleeps" using AVM-notation

In the feature structure representation the arguments for each EP have names such as ARGO, ARG1, RSTR and BODY. In standard HPSG each word has its own specific feature for relations, for instance HUNTER and HUNTED for the word "hunt" and EATER and EATEN for "eat". This is avoided in MRS in order to allow generalisations between words with the same argument structure, for instance when binding semantic arguments to syntactic arguments. This way words like "hunt" and "eat" can have the same basic type, and only specify their string representation in their individual entries. Usually the relations are named ARGX, where X starts at 0, but some specific types of EPs receive special argument names, for instance RSTR and BODY FOR quantifiers. The argument names have standard uses, ARG1 for verbs is for instance always mapped to the subject.

#### 3.3 The LinGO Grammar Matrix

The LinGO Grammar Matrix<sup>5</sup> is a HPSG-based grammar base, or "starter kit", for developing grammars in different languages in the LKB system<sup>6</sup>. On their homepage<sup>7</sup> the Matrix developers' group states their main objectives as:

to leverage the expertise in grammar engineering embedded in the broad-coverage grammars to create a resource available to other efforts;

to build and test a set of hypotheses about linguistic universals, from a data-driven, bottom-up perspective;

to facilitate the development of grammars for different languages which produce semantic representations in a common format (MRS), such that they may be used interchangeably with the same backend software in NLP systems;

in the long term, to create a tool that allows field linguists to easily build implemented grammars as they research a language, to test hypotheses and encode their results; and

to facilitate the exchange of data and analyses of a wide range of phenomena across diverse languages. (Bender et al., 2005)

The Matrix is based on grammars for English (LinGO ERG<sup>8</sup>) and Japanese (JACY<sup>9</sup>). It is continuously under development and thrives on input from grammars that are being developed based on it, for instance for Norwegian, Spanish, Greek and German.

The main part of the Matrix is a type hierarchy of basic features that are common for languages in general. In the current version there are 258 basic types and another 501 types with cross-classifications of head-types. It is not in itself a finished grammar, but a base to build other grammars on. The Matrix types are hypothesized to be necessary, or at least useful for any large scale Matrix-based grammar (Hellan, 2003). Among the types in the Matrix are the following categories (Bender et al., 2002):

• types for basic feature geometry and technical devices, like list manipulation (low-level types)

<sup>&</sup>lt;sup>5</sup>Version 0.8 was used in BiTSE, and is the version referred to in this report. The Matrix is available for download at http://lingo.stanford.edu/ftp/

<sup>&</sup>lt;sup>6</sup>Matrix-based grammars can also be used in other compatible TFS-based systems, such as PET, see http://wiki.delph-in.net/moin/PetTop

<sup>7</sup>http://www.delph-in.net/matrix/

<sup>&</sup>lt;sup>8</sup>See http://lingo.stanford.edu/erg.html

<sup>&</sup>lt;sup>9</sup>See http://www.delph-in.net/jacy/

- types for modelling semantics by using MRS, including types for MRS structures, constraints for the propagation of semantic information and a provision to let grammar rules make semantic contributions
- general classes of rules
  - constant and inflectional lexical rules
  - phrase structure rules of several types such as headed versus non-headed and unary versus binary
  - rules that includes basic HPSG principles like the head feature principle
- types for basic constructions like head-complement and head-subject
- head types for basic constituents, and basic phrase types for them (in later versions than 2002)

The Matrix also includes configuration files for LKB.

As mentioned before all types are organised in a hierarchy. Figure 3.5 shows a part of this hierarchy. It has a single root element \*top\*. Below \*top\* are three different types: tree-node-label, sort and avm. tree-node-label is the top type for the labels that are used to name the nodes in parse trees. They are not a part of the grammar itself, rather a help for the grammar writer/evaluator, and will not be discussed further. sort is the top type for all atomic types, that is, types without features, such as tense, semsort (semantic sort, like animate or time) and mood. avm is the top type for types which are feature structures, or attribute value matrices. Most types are avms, for instance the types for sign, synsem and mrs.

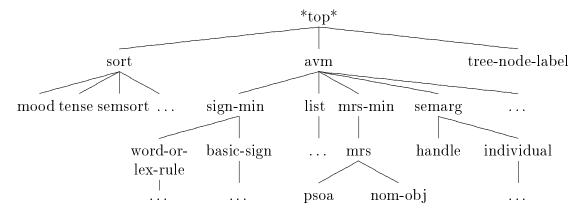


Figure 3.5: Type hierarchy for some basic Matrix types

In the Matrix signs are slightly different from those in standard HPSG as described in Section 2.4. Figure 3.6 shows a basic Matrix sign. The feature for PHON is changed to STEM and has a list of strings as its value. The valence is extended to four lists: SUBJ, SPR, COMPS and SPEC. The first three are used by heads to select different types of arguments, and SPEC is used for specifiers to select their head.

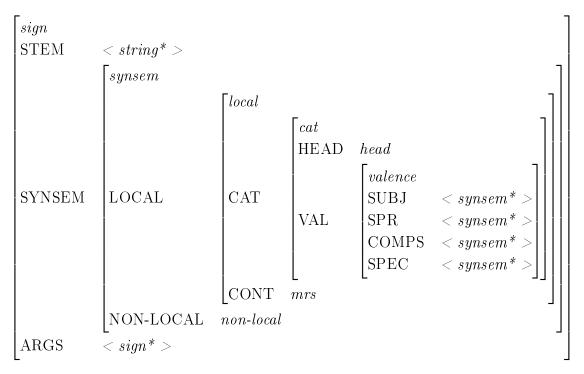


Figure 3.6: Basic Matrix sign ( $\langle X^* \rangle$  stands for a list with 0 or several elements of type X)

The types for rules all contain one or two daughter constituents. Figure 3.7 shows the containment relations for verbal types. The two lower rule types are lexical rules, and the others are grammar rules, that join a verb or verb phrase with another constituent.

As an example of the types in the Matrix I will discuss lexical rules in some more detail. The Matrix types for lexical rules are shown in figure 3.8. There is a base type for all lexical rules which inherits from other basic rule types. This type describes what is common for all lexical rules, for instance how the MRS information should be combined. Lexical rules can then be described based on two dimensions, if they change a lexeme into a word or into a lexeme, and if they change the spelling or not. Thus there are four types that inherits from the basic lexical type and expresses each of these things. There are then four other types which combine the values from the two dimensions pairwise. These last four types are then used by every actual lexical rule in a Matrix-based grammar, they express the basic function of the rule, and information has to be added on what the rule actually changes, and if it changes the spelling, how it does that.

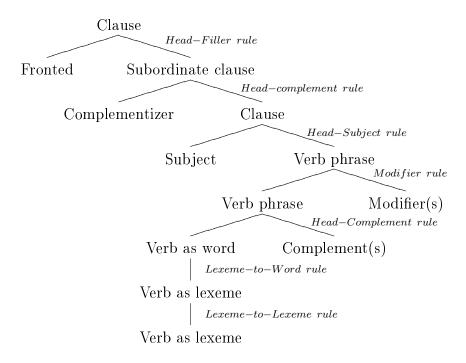


Figure 3.7: Containment relations for verbs in the Matrix (based on Hellan, 2003)

Figure 3.9 shows an example of how to use Matrix types. It shows a BiTSE rule and its type for forming present tense for Swedish verbs of the er-conjugation, which means that it should end with "er". The rule is an instance of a type called *pres-verb-lex-rule*, which in turn inherits from the Matrix type *infl-ltow-rule* described above, which means that the spelling is affected and that applying the rule changes the lexeme to a word. The rule more specifically changes the value of TENSE to *present* and the value of VFORM to *finite*, since present tense is finite. The actual rule then inherits from this type. Inflectional rules also has a line that starts with "%suffix" which shows how to change the spelling, if it ends with a vowel ("!v"), the vowel is removed and the suffix "er" is added, else the suffix "er" is simply added. The rule is also constrained to work only on verbs in the er-conjugation, by stating that the verb should have the value *pres-er* for its CONJUG feature.

MRS is the semantic representation used in the Matrix. The type mrs appears as the value of CONT(ent) in lexical and phrasal signs and of constructional content (C-CONT) in rules. C-CONT is the semantic contribution that is added by a rule. Figure 3.10 shows the mrs type in the Matrix. Rels is a difference list of relations, HCONS a difference list of handle constraints, or qeqs, MSG is a message, which is used for the propositional content of a sign, i.e. if it is a question, proposition or command. HOOK is a feature structure of type hook and it contains the information within the MRS that can be accessed externally.

```
lex-rule := phrase-or-lexrule & word-or-lexrule &
  [ NEEDS-AFFIX bool,
    SYNSEM.LOCAL.CONT [ RELS [ LIST #first,
                               LAST #last ],
                               HCONS [ LIST #hfirst.
                                       LAST #hlast ] ],
                               DTR #dtr & word-or-lexrule &
                                  [ SYNSEM.LOCAL.CONT [ RELS [ LIST #first,
                                                               LAST #middle ],
                                                        HCONS [ LIST #hfirst,
                                                                LAST #hmiddle ] ],
    C-CONT [ RELS [ LIST #middle,
                   LAST #last ],
                   HCONS [ LIST #hmiddle,
                           LAST #hlast ] ],
    ALTS #alts,
    ARGS < #dtr > ].
lexeme-to-word-rule := lex-rule &
  [ INFLECTED +,
    KEY-ARG #keyarg,
       SYNSEM #synsem,
    DTR [ INFLECTED -,
          KEY-ARG #keyarg,
             SYNSEM #synsem ],
    C-CONT.RELS <! !> ].
lexeme-to-lexeme-rule := lex-rule &
   [ INFLECTED #infl,
    SYNSEM.LOCAL.CAT.MC #mc,
    DTR [ INFLECTED #infl,
          SYNSEM.LOCAL.CAT.MC #mc ] ].
inflecting-lex-rule := lex-rule &
  [ NEEDS-AFFIX + ].
constant-lex-rule := lex-rule &
  [ STEM #stem,
    DTR [ STEM #stem ]].
const-ltol-rule := lexeme-to-lexeme-rule & constant-lex-rule.
infl-ltol-rule := lexeme-to-lexeme-rule & inflecting-lex-rule.
const-ltow-rule := lexeme-to-word-rule & constant-lex-rule.
infl-ltow-rule := lexeme-to-word-rule & inflecting-lex-rule.
```

Figure 3.8: Matrix types for lexical rules

Figure 3.9: Inflectional rule and its type for present tense of Swedish verbs from BiTSE

This is the information that is used when MRSs from different signs are combined to build larger MRSs. The *hook* type is shown in Figure 3.10. Its feature LTOP stands for local top handle and is co-indexed with the handle with the highest scope in the MRS. XARG is linked to the only argument in a phrase that can be controlled externally. Its most common use is to keep track of the subject for raising and control verbs like "try" and "continue". INDEX contains semantic information about the sign, and is identified with the index of the semantic head daughter. It has two subtypes, *ref-index* (referential index) and *event*. If it is a *ref-index* it contains information about *person*, *number* and *gender*, *png*, and if it is an *event* about *tense*, *mood* and *aspect*, *tam*. There are subtypes of *mrs* that constrain INDEX: *psoa* (parameterized states of affairs) constrain it to be an *event*, and *nom-obj* to be an *index*.

```
mrs := mrs-min & hook := avm &

[ HOOK hook,

RELS diff-list,

HCONS diff-list,

MSG basic_message].
```

Figure 3.10: Matrix mrs and hook types

Unfortunately the Matrix is not very well documented at the moment, except for source code comments. But there is a useful user's guide to MRS in the Matrix (Flickinger et al., 2003).

## 3.4 The DELPH-IN MT architecture

The general architecture proposed for MT using DLEPH-IN resources is semantic transfer, which will be discussed here. An alternative, where more than one language are included in one grammar will also be discussed.

#### 3.4.1 Semantic transfer

Copestake et al. (1995) describe how MRS can be used for translation, using a design based on semantic transfer of MRS. They suggest a transfer component that works on MRSs to produce output that the target grammar can accept. It is possible that the transfer component can output more than one form, some of which may be unacceptable by the generator. When several forms are output they will be ordered by a control mechanism that is distinct from both the transfer component and the generator.

Their suggestion also allows some interlingual predicates that are common for all languages such as a negation-relation. Flickinger et al. (2005) points out that a similarity between semantic transfer and interlingua is that they share "the assumption that translation is at its core a semantic activity" (p. 165). This differs from syntactic transfer that focus on syntactic properties instead. Flickinger et al. (2005) further note that semantic transfer differs from interlingua in that it emphasizes "that different languages carve up reality differently" (p. 165). They think that these differences should be handled for one language pair at the time.

Copestake et al. (1995) states that two main advantages of using MRS in translation is that it allows semantic underspecification of various types, and that it is flat. One example of underspecification is underspecification of quantifier scope. This because it is often unnecessary to resolve quantifier scope in translation since it usually has no effect in translation and it is hard to resolve when parsing (Copestake et al., 2003). Other cases where MRS allow underspecification is lexical ambiguity, for which the type hierarchies of relation types can be used, and attachment of prepositional phrases.

The transfer component suggested by Copestake et al. (1995) is based on setting up symmetric and bidirectional transfer equivalences between each pair of languages. A transfer equivalence consists of a relationship between a set of relations from each language with correct mapping of the co-indexation of the relations. Figure 3.11 shows an example of a translation equivalence. The names used are different from later versions of MRS, ARG is later called ARGO, and HANDEL is called LBL. The fact that MRS is flat allows transfer rules to be simplified compared to more conventional semantic representations, by allowing a rule to apply anywhere in the structure, as long as the relations between handles and arguments are preserved (Bond et al., 2005). Transfer rules are expressed

as feature structures and classes of them could be described using the type hierarchy to express generalisations.

$$\left\langle \begin{bmatrix} schlecht\_rel \\ ARG & elevent \end{bmatrix} \right\rangle \iff \left\langle \begin{bmatrix} neg\_rel \\ ARG & \boxed{h2} \end{bmatrix}, \begin{bmatrix} good\_rel \\ HANDEL & \boxed{h2} \\ ARG & ell \end{bmatrix} \right\rangle$$

Figure 3.11: Translation equivalence from Copestake et al. (1995), relating German "schlect" with English "not good"

A large scale project where semantic transfer with MRS is used is LOGON, which focus on translation between Norwegian and English (Oepen et al., 2004; Lønning et al., 2004). The main architecture is: analysis of Norwegian to MRS using the Norwegian LFG grammar NorGram<sup>10</sup>, MRS transfer, much like described above, and generation to English using the LinGO ERG<sup>11</sup>. The base type for transfer rules on MRSs as described by Lønning et al. (2004) is shown in Figure 3.12. The first three features, INPUT, CONTEXT and FILTER are unified against the incoming mrs, and if successful INPUT is replaced by OUTPUT. CONTEXT and FILTER are optional features and are used to constrain the rule applications to specific contexts. Lønning et al. (2004) then show an example of a rule and an instance of it for regular verbs, shown here in Figure 3.13.

Figure 3.12: Basic MRS transfer rule from LOGON (Lønning et al., 2004)

The LOGON system is unidirectional, and only translates from Norwegian to English, due to the design with different grammars for analysis and generation. However, Bond et al. (2005) notes that in the general design the HPSG for each language can be used both for parsing and generation. This complicates grammar construction somewhat, but means that the construction of one grammar gives two modules, one for parsing and one for generation. The transfer rules are also reversible, except for CONTEXT and FILTER information when applicable.

The Matrix contains base types for transfer rules, similar to those used in the LOGON project. This part of the Matrix are however still under development and not ready for

<sup>10</sup> see http://www.hf.uib.no/i/LiLi/SLF/Dyvik/norgram/

<sup>&</sup>lt;sup>11</sup>See http://lingo.stanford.edu/erg.html

```
arg1_v_mtr := mrs_transfer_rule &
  [ INPUT.RELS < [ LBL #h, ARGO #e, ARG1 #x ] >,
     OUTPUT.RELS < [ LBL #h, ARGO #e, ARG1 #x ] > ].
bjeffe :=arg1_v_mtr &
  [ INPUT.RELS < [ PRED "_bjeffe_v_rel" ] >,
     OUTPUT.RELS <[ PRED "_bark_v_rel" ] > ].
```

Figure 3.13: Type and instance of a MRS transfer rule for the verb "bark"/"bjeffe" in LOGON (Lønning et al., 2004)

use yet, but there are plans to extend it in future Matrix versions. However, the use of the current Matrix version when building grammars allready facilitates translation by normalizing the type hierarchy and by standardising the names of common relation types, for instance for messages (Bond et al., 2005).

Bond et al. (2005) discusses the open source resources for MT made available by the DELPH-IN collaboration and the general strategies used, including the basic ideas presented in this section. They also raises some proposals for future work, including "How much of the semantic representation can be shared between languages (and thus require little or no transfer)?" (p. 20).

## 3.4.2 Multilingual grammar design

A very different architecture for MT has been suggested by Søgaard & Haugereid (2005). Their work was part of a project to create a grammar Matrix for the Scandinavian languages, building on top of the LinGO Grammar Matrix. What they did was to include the three Scandinavian mainland languages in one single grammar. To be able to tell constructions of different languages apart they introduced a feature for LANGUAGE, which they organised in a hierarchy of the three languages. The LANGUAGE feature is not semantic, and the MRS that is obtained when parsing a sentence from any language is language independent, so generating from it gives all equivalences in all three languages. Thus, as Søgaard & Haugereid (2005) points out, this grammar design gives a grammar that in a sense also is a small MT-system. In this system no transfer takes place, and the MRS representation functions as an interlingua for the three languages.

# 4 Objectives revisited

Having discussed the general concepts of this study I describe the objectives a bit more thoroughly here using some of the concepts presented in the background chapters. The three main objectives of this study are:

- Descriptive: to find out which verb frame divergences exist between Swedish and English, and how they can be classified
   I tried to identify as many cases of VFDs as possible and to fit them into Dorr's (1993) classification, which is targeted at machine translation. This is a base that allows me to handle the divergences in each group uniformly.
- 2. Practical: to construct a bilingual grammar for Swedish and English

  The basic grammar design by Søgaard & Haugereid (2005) with several languages
  in one grammar was adopted in constructing BiTSE. This shows how much of the
  grammars of Swedish and English are the same, and reduces the redundancy of
  having the part of the grammar which is actually the same in two separate grammars.

  It also forces the grammar to have a common MRS representation for equivalent
  sentences. Developing a Matrix-based grammar that includes Swedish could also
  potentially contribute to the Matrix endeavour by seeing if Swedish fits into the
  Matrix.
- 3. Machine translation theoretical: to find out what cases of verb frame divergences can be handled by an English-Swedish interlingua, and when it is not as suitable.

  The major DELPH-IN approach to MT is semantic transfer with some elements of interlingua. The bilingual grammar design allows me to find out if it is possible to use an interlingual approach instead. I will thus bring the suggestion of Bond et al. (2005), to find out how much information can be shared by languages, to an extreme by trying to share all information for Swedish and English. In doing this I can also find out when a common representation does not seem to work, or at least be very complicated. I will both describe theoretical solutions, or tentative solutions to the identified VFDs and implement a subset of them in the grammar. I am not trying to find a general interlingua, only one that is valid between Swedish and English.

# 5 Categorization of verb frames

This chapter discusses the classification of the specific VFDs into broader categories of divergences based on the results of the survey that was carried out to identify verb frame inventories of Swedish and English and their possible corresponding translations.

# 5.1 Verb frame divergence categories

The initial survey of verb frame divergences in Swedish and English resulted in inventories of Swedish and English verb frames, which can be seen in Appendix A. Also a comparison was done to try to find the possible English verb frames corresponding to each specific Swedish verb frame found. This was then mostly used as a basis for categorizing VFDs into groups with similar differences.

A comparison of the identified categories and Dorr's categories (Dorr, 1993) showed that the identified divergences fitted well into Dorr's categories. For one of Dorr's lexical-semantic categories, thematic, there were no instances in the survey, though it is still possible that there can be some rare cases of them, in which case thematic divergences would have to be added. Lexical divergences always co-occur with some other divergence and will not be handled separately in this study. Dorr's demotional and promotional categories have been put together to one category: head inversion.

The divergences that correspond to Dorr's syntactic divergences have not been further divided at the top level. These are, as Dorr points out, different to the other categories, which are all lexical semantic.

This classification only covers VFDs. The categories can be used for divergences concerning other constituents than verbs as well, but if a broader field were to be covered the number of categories might have to be expanded.

The five broad divergence categories used: structural, conflational, head inversion, categorial and syntactic, will be further described in this section, and examples of them will be shown.

## 5.1.1 Structural divergences

This category is the same as for Dorr. It occurs when the same logical constituent in the two languages have different structures. The identified cases of structural divergences can be seen in Table 5.1. In the cases with a reflexive pronoun or a particle, they are seen as forming a unit with the verb, which is thus structurally different to a simple verb. Common to all identified cases are that they are discontinuous, i.e. other constituents can in some cases appear between the verb and the other constituent. Especially in Swedish it is common to find examples which combines two or more of these divergences, as in (8).

(8) a. Jag sätter mig upp mot Bo. I defy Bo.

I sit myself up aginst Bo.

Description	Example
prep compl. – NP obj	lita på någon – trust someone
	trust on someone
refl. verb – verb	lära sig – learn
	learn oneself
part. verb – verb	komma ihåg – remember
	come to.mind
plain inf. – inf. +	behöva sova - need to sleep
marker	need sleep(INF)

Table 5.1: Structural divergences

# 5.1.2 Conflational divergences

A conflational divergence occurs when a constituent in one language is implicit in the other. The identified cases of conflation can be seen in Table 5.2. A special case of conflational is when a constituent is optional in one language and not in the other, and thus behave as conflated when it is not present. There are also cases which borders on belonging to the conflational VFD group, like (9) where the possessor of the "head" is missing in Swedish. Here however, it is not a complete constituent that is missing, both in English and Swedish the verb takes a NP as complement, but the structure of the NP is different.

(9) jag skadade knät I hurt my knee.
I hurt knee.the

Description	Example
English explicit body	snyta sig - blow one's nose
part	blow.nose oneself
Swedish explicit	fatta beslut – decide
compl.	take decision
English explicit	svara - give an answer
compl.	answer
English optional refl.	raka sig – shave [oneself]
	shave oneself
Swedish optional	hitta [vägen] - find the way
compl.	find.way

Table 5.2: Conflational divergences

## 5.1.3 Head inversion divergences

This category covers the two Dorr categories promotional and demotional, which occurs when a main verb in one language corresponds to a modifier in the other. Dorr's distinction between the two has not been considered necessary in this study. Dorr only mentions adverbial modifiers, but since examples containing particle modifiers were identified they are also included into this group. The identified cases of head inversion can be seen in Table 5.3.

DescriptionExamplemain verb – adverbial<br/>modifierspacka klart – finish packing<br/>pack readyparticle modifiers –<br/>main verbspringa på – keep running<br/>run on

Table 5.3: Head inversion divergences

# 5.1.4 Categorial divergences

This category is the same as for Dorr and it covers divergences where semantically corresponding constituents have different syntactic categories in the two languages. The identified cases of categorial divergences can be seen in Table 5.4. In all the identified cases a copula verb and its predicate in one language are expressed in some other way in the other.

Description	Example
pred (adj) – NP	vara förkyld - have a cold
	be colded
pred (adj) – verb	vara skyldig - owe
	be guilty
verb+part+refl -	smutsa ner sig - get dirty
pred (adj)	dirty down oneself
verb+refl - pred (past	gifta sig - get married
participle)	marry oneself

Table 5.4: Categorial divergences

#### 5.1.5 Syntactic divergences

Syntactic divergences corresponds to Dorr's syntactic category. This category comprises divergences that arise from syntactic, usually language internal, differences of distribution of certain features. The category contains Catford's intra-system shifts in regard to syntactical factors. An example of this are ditransitive verbs which have the same two patterns in Swedish and English, but where the distribution is different. The identified cases of internal divergences can be seen in Table 5.5. The other case in the table, English particle placement, gives rise to a diversion in one of its possible word orders. This would be classified as a word order syntactic divergence by Dorr. It could however be discussed if it should not rather be seen as a syntactic language internal factor.

Table 5.5: Internal divergences

Description	Example
Different ditransitive	I told Bo a riddle - *Jag berättade Bo en gåta
alternations	I told a riddle to Bo - Jag berättade en gåta för Bo
English particle	I washed out the bath - Jag tvättade ur karet
placement	I washed the bath out - *Jag tvättade karet ur
alternations	

# 6 The BiTSE grammar

This chapter will describe the design and the core of BiTSE, the Bilingual grammar for Translation between Swedish and English, that was developed as part of the thesis.

# 6.1 Basic design

BiTSE is a bilingual grammar that contains both Swedish and English. Adopting the DELPH-IN framework it is based on the Matrix. One advantage with basing it on the Matrix is that much of the basis of what I needed was already done. Another good point about this choice is that using the Matrix for yet another language might further contribute to developing the Matrix. Using the Matrix also makes the grammar, at least to some extent, compatible with other DELPH-IN grammars and resources.

An alternative would have been to use existing grammars, possibly LinGO ERG for English and SweCore<sup>1</sup> for Swedish. This would have meant that I had to make sure that the two grammars gave the same semantic output for sentences I find equivalent, something which is not always the case for these grammars, even when not taking relation names into consideration. Another problem is that the coverage is not the same in the two grammars, ERG is a broad coverage grammars while SweCore only covers the core of Swedish grammar. It might also be the case that the analyses that I want to use for VFDs are not consistent with the existing grammars. All of this would have meant that I would have had to make changes to these existing grammars, something that would have been especially difficult in the ERG, because of its size and detail.

# 6.2 The core grammar

The very core of BiTSE is based on exercises that are part of a lab course in a grammar engineering course held by Emily Bender, one of the Matrix developers, at the University of Washington in 2004 and 2005, which is available on-line (Bender, 2004, 2005). The coverage of the core grammar is:

<sup>&</sup>lt;sup>1</sup>Developed by Lars Ahrenberg based on the Norwegian NorSource grammar

- Verb phrases with:
  - Intransitive verbs
  - Transitive verbs
  - Simple raising verbs
  - Verbs with embedded declarative and interrogative clauses as complements
- Noun phrases
- Polar questions
- Prepositional phrases as modifiers
- Adjectival modifiers
- Full inflection for all lexemes in the grammar for:
  - Past tense verbs
  - Swedish present tense verbs (English present tense needs no inflection)
  - Plural nouns in English and Swedish
  - Definite nouns in Swedish
  - Concord inflection of Swedish adjectives

The above is the core of BiTSE. To accommodate VFDs, which will be described in the next chapter, the following parts were added:

- Ditransitive verbs
- Idiomatic particles
- Empty prepositional complements
- Reflexive pronouns
- Some scopal adverbs
- Some types of optional complements

The purpose of this grammar is to find solutions of how to handle phenomena in Swedish and English. Thus the focus have been on the grammar itself, and the lexicon is quite small, for most cases there is only one example of a word for each type of grammar phenomena. All together there are around 65 lexemes in each language.

The description of the core grammar describes its current status. If more features were to be added, it might induce changes to the core as well, since assumptions used there might have to be revised. In some cases the types that are shown have been somewhat simplified to illustrate the discussed phenomenon clearer.

The largest part of the grammar is the description of types. They are organized in a hierarchy, and to a large extent based on the types of the Matrix. The specific types needed for the handling of different phenomena will be described in connection with the description of that phenomenon.

## 6.2.1 Constructions for language

To include more than one language in a grammar, a feature that constrain the language of SIGNs had to be added in addition to the basics of the Matrix. Following the basic idea of Søgaard & Haugereid (2005), a feature for language, called LANG, with the values for Swedish (sw) and English (en), was added to SIGNs to represent what language the SIGN belongs to. Constraints then had to be added on all rules to make them work on a single language at the time, and for language specific rules to work on only one language. Figure 6.1 shows some sample types for language handling. LANG is not a semantic feature, which makes the semantic representation (MRS) language independent, resulting in generation giving all equivalent sentences in both languages. Thus, as Søgaard & Haugereid (2005) points out, this grammar design gives a grammar that in a sense also is a small MT-system.

All equivalent relations must have the same names in order to have all equivalent sentences generated. To achieve this all relations have English names, such that each English word generally has a relation with the same name as the word, and each Swedish word has a relation with the corresponding English name.

# 6.2.2 Verb phrases

Some basic types of verbs were added to the grammar as a starting point, examples of these can be seen in (10). The coverage for basic verb phrases is the same for Swedish and English.

(10) a. I sleep

```
addition of a language feature to SIGNS
sign :+ [ LANG lang ].

definitions of the type lang
lang := sort.
sw := lang.
en := lang.

basic rule for binary rules that apply to one language only
binary-lang-agree-phrase := binary-headed-phrase &
   [ LANG #lang,
        HEAD-DTR.LANG #lang].

basic rule to constrain headed rules to work only for Swedish
swedish-only-rule := headed-phrase &
   [ LANG #sw,
        HEAD-DTR.LANG #sw & sw ].
```

Figure 6.1: Examples of types used to constrain language in BiTSE

- b. I hunt him
- c. I can sleep
- d. I hope that he sleeps
- e. I wonder whether he sleeps

The base type for verbs were inherited from the Matrix and is further constrained to have a NP subject, and to get the correct semantics. Some features were also added to the head type verb: two boolean features for INVersion and AUXiliary verbs, which are used for instance in connection with questions, INV, indicating inverted word order, with the verb before the subject, and AUX used to tell auxiliary verbs apart form other verbs. Another feature VFORM is added for the temporal form of verbs, this is further subdivided into finite and infinite, which is in turn divided into subtypes for infinitive, base, present participle, past participle and perfect participle. Only infinitive is currently used in BiTSE. The finite tenses are handled by the semantic feature TENSE, which is also used for clauses. Currently present and past tense are part of BiTSE.

The type hierarchy for the basic types of verbs can be seen in Figure 6.2. All verb types are inherited from verb types in the Matrix and further constrained for relevant features.

The semantic relations between verbs and their arguments are also constrained in order to form correct semantics. The subject of all verbs is constrained to be of the nominative case, and the objects to be accusative. As an example of this a somewhat simplified type for simple transitive verbs, trans-verb-lex, like "hunt" is shown in Figure 6.3, which also shows the lexical entry for "hunt". This inherits from other BiTSE and Matrix types, which carry information relevant to a broader class of signs, such that the binding of the arguments to the semantic roles of the verb's relation. As could be seen most information is in the types of the hierarchy, and the lexical entries only contain the specific information for the word, the word string, the language it is in, and the relation for the word.

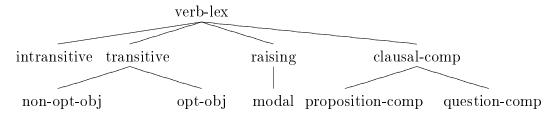


Figure 6.2: Type hierarchy for basic verbs, the full names are not shown, but can be inferred from the nodes above, e.g. opt-obj => trans-opt-obj-verb-lex

Figure 6.3: Type and lexical entry for transitive verbs like "hunt"

The *trans-verb-lex* type in Figure 6.3 also shows the handling of optional complements. Verbs like "hunt" can leave their object out, in which case it is left underspecified what is being hunted. This is marked by adding a positive value for the feature OPT for the object. A rule to remove optional complements were also added which only applies to

complements marked OPT +, and which removes the complement from the COMPS list and leaves the semantic relation of the verb underspecified.

To combine verbs with their arguments there are two basic grammar rules, *head-comp* and *head-subj*, which are both inherited from the Matrix. Figure 6.4 shows the rule instantiation, *head-comp* and the basic Matrix type for head-complement phrases. *Head-comp* is used to add a complement to a head, not only for verbs but for all other constituents that have complements. It inherits most of its content from the Matrix base type, which contains information to keep the semantics correct, and that moves the first complement off the comps list and co-indexes it with the non-head daughter. It also copies all information that has not changed from the head daughter to the top SYNSEM. *Head-comp* also inherits from the Matrix type *head-initial*, that constrain the head to come before its complement and, as all BiTSE phrases, from the BiTSE type *binary-lang-agree-phrase* that constrains the head and complement to have the same language. The rule instantiation does not posit any new constraints, and is an example of the fact that most information is kept in the type hierarchy.

**Head-subj** is used to add a subject to a head and it also inherits from a Matrix base type and **binary-lang-agree-phrase**, but also from **head-final**, which constrains the head to come after the subject. It is constrained to have a head with an empty COMPS list, which means that the subject is always added after all complements have been added. Finally it inherits from **declarative-clause**, which turns the phrase into a declarative clause by adding a proposition relation with the head verb as its argument.

Raising verbs, for instance the modal auxiliary "can", in a sentence like "I can swim", inherits the Matrix type for transitive raising verbs and are constrained for its verbal complement to have VFORM infinitive. Verbs taking clausal complements, for instance "hope" in a sentence like "I hope that he sleeps", are constrained to have a clausal complement of VFORM finite. There are two types of clausal complements, declarative and interrogative. Thus the complement is also constrained to be declarative or interrogative, or underspecified in case it allows both, like "know". Clausal complements must currently start with a complementizer, either "that"/"att" for declarative embedded clauses or "whether"/"om" for interrogative clauses. Complementizers do not have a semantic relation, but are seen as purely grammatical, and they take a finite clause as their only complement.

All verbs are marked as noninflected in the lexicon, and a number of lexical rules handles inflection. There are constant rules for infinitive form and English present tense that do not need any inflection, and only marks the verb with the correct tense and verb form. All other tenses and verb forms have inflectional rules. For Swedish verbs a head feature for conjugation is added, which is used to divide verbs according to their inflectional patterns, currently in the present and past tenses. There are a Swedish inflectional rule for each tense-conjugation combination, see Figure 3.8 for an example, and an English inflectional

```
The instantiation of the rule
head-comp := head-comp-phrase.
The BiTSE type
head-comp-phrase := basic-head-comp-phrase \& head-initial \&
                    binary-lang-agree-phrase &
 [ SYNSEM phr-synsem & [ LOCAL.CONT.MSG #msg ],
   HEAD-DTR.SYNSEM.LOCAL.CONT.MSG #msg ].
The Matrix type
basic-head-comp-phrase := head-valence-phrase & head-compositional &
                          binary-headed-phrase &
 [ SYNSEM canonical-synsem &
          [ LOCAL.CAT [ MC #mc,
                        VAL [ SUBJ #subj,
                              COMPS #comps,
                               SPR #spr ],
                        POSTHEAD #ph ],
            LIGHT #light ],
   HEAD-DTR.SYNSEM [ LOCAL.CAT [ MC #mc,
                                  VAL [ SUBJ #subj,
                                        COMPS < #synsem . #comps >,
                                        SPR #spr ],
                                  HC-LIGHT #light,
                                 POSTHEAD #ph ]],
   NON-HEAD-DTR.SYNSEM #synsem & canonical-synsem,
   C-CONT [ RELS <! !>,
            HCONS <! !> ] ].
```

Figure 6.4: The head-complement rule, its BiTSE type, and the basic head-complement Matrix type.

rule for each tense. For irregular verbs their inflected forms are added in a special file containing all irregular inflectional forms, for all types of words.

## 6.2.3 Noun phrases

Three major types of nouns are implemented: common nouns, personal pronouns and proper nouns. In BiTSE and generally in HPSG pronouns are seen as nouns, since they have a similar function (Pollard & Sag, 1994), and not, as sometimes is the case, as a word class of their own (see e.g. Jörgensen & Svensson, 1987). Examples of the types of noun phrases covered can be seen in (11). The coverage is the same for English and Swedish. Some of the noun phrase phenomena that are not handled are: quantifiers such as "every", relative clauses and mass nouns such as "milk".

- (11) a. I
  - b. me
  - c. dogs
  - d. Tom
  - e. the big dogs
  - f. those dogs
  - g. the lion in the story
  - h. hunden (dog.the)

The type hierarchy for nouns can be seen in Figure 6.5. A base type for nouns are inherited from the Matrix, adding the constraint that all nouns should have a specifier. This is seen as true even for pronouns and proper nouns which normally do not have specifiers, since there are possible cases when they do, for instance "The John that I know". The grammar rules to achieve this are however not implemented. Common nouns are constrained to be of third person as default, and for its specifier to be regular. In Swedish common nouns have gender, and thus there are subtypes for neuter and non-neuter Swedish nouns.

For all pronouns a special pronoun relation is introduced by the pronoun and its specifier is constrained to have a pronoun quantifier relation. The types for pronouns are then further subdivided into nominative, accusative and reflexive, that are used by verbs to constrain subjects to be nominative and objects to be accusative or reflexive. Verbs can also choose reflexive pronouns as complements using this feature. Information about person, number and gender are then added on each pronoun entry in the lexicon.

Proper nouns are divided into feminine, masculine and thing, which for instance are used for geographical names. Only proper nouns that are the same in both languages are currently included in BiTSE, and they are underspecified for language. The semantic

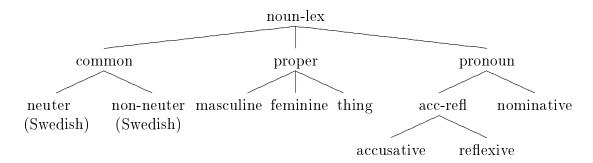


Figure 6.5: Type hierarchy for nouns, the names are simplified for readability

relation for proper nouns is named-relation, which have a CONSTANT ARGUMENT, CARG, which contains the string value of the entity the proper noun refers to. Proper nouns which are different in the two languages, such as "Lisbon" - "Lissabon" are not handled yet. To handle them it is either possible to let CARG refer to some interlingual name of the proper noun, for instance the name in the original language, i.e. "Lisboa" for "Lisbon", or an arbitrary name such as "city486". This practically works, but might not agree with the idea behind the named-relation. Another option would be semantic transfer, where the mapping of name has to be specified as transfer equivalences.

There are a number of lexical rules for common nouns. There are inflectional rules for adding plural for both languages, and for definiteness for Swedish. There is also a constant rule for adding singular. English plurals are generally constructed by adding the suffix "-s", all other cases are considered irregular. Swedish inflection is a bit more diverse, and because of this a feature DECL(ination) is introduced, which divides Swedish nouns into five classes within which inflection is similar. Then there are different rules for each declination, that contain the correct inflection for each group. The distinction between the five declinations is based on Jörgensen & Svensson (1987), but others have different suggestions of declination classes, and this might have to be updated if a larger coverage of nouns will be required in the future.

The type for determiners is simply inherited from the Matrix, only constraining all elements of the valence to be empty. Information about number and relation type (indefinite, definite or demonstrative) are then added for each entry in the lexicon. Determiners are used as specifiers of nouns.

Noun phrases containing a specifier are built by using the *head-specifier-rule*, that combines a noun with a specifier. The specifier constrain the noun phrase that it specifies to agree on information about person, number and gender (PNG).

For noun phrases without specifiers a lexeme-to-lexeme-rule called *covert-det-phrase* was created that adds the implicit quantifier relation. This rule has four subtypes, one for pronouns, that adds a *pronoun-quantifier-relation*, one for proper nouns that adds a *definite-quantifier-relation*, one for plural nouns, that adds an *indefinite-quantifier-relation*, and one for Swedish definite nouns, where definiteness is achieved by adding a suffix instead of a specifier as in English, which receives a *definite-quantifier-relation*. All other noun types in BiTSE are hypothesized to obligatory take a specifier.

Noun phrases can also have prepositional or adjectival modifiers, these will be described in Section 6.2.5 about modification.

## 6.2.4 Empty and contentive constituents

Lexemes are divided into two categories depending on their semantic contribution: *empty*, which do not make any semantic contribution by themselves and *contentive*, which make a semantic contribution and thus contain a semantic relation<sup>2</sup>.

Verbs, nouns and adjectives are generally contentive and grammatical words like complementizers like "att"/"that" are empty. Prepositions, particles and pronouns can be of either category, examples of this variation can be seen in Table 6.1. The empty constituents are chosen by a verb (or possibly by other constituents) which carries information on which empty constituents it needs, and has a relation for the sense of the verb including these empty constituents.

	Empty	Contentive
Reflexive pronoun	perjure oneself	shave oneself
Particle	make $up$	climb $up/down$
Preposition	depend on something	put something on/under something

Table 6.1: Examples of empty and contentive constituents

The existing Matrix types do not provide good support for empty complements, so I had to incorporate a number of basic types for this into BiTSE. The existing Matrix types for words with complements did not give the correct semantics to empty complements, which should receive no semantic bindings at all. Swedish verbs can have up to two empty complements, as in (12), which has an empty reflexive and particle, and besides that also contains an empty preposition, in the prepositional complement.

<sup>&</sup>lt;sup>2</sup>There is no general consensus on these terms, many other terms have been used such as *substantive* versus *functional* (Pollard & Sag, 1994), *content words* versus *function words* (Aitchison, 1999) and for prepositions: *type A, type AB* and *type B* (Tseng, 2000).

```
(12) a. Jag satte mig upp mot Bo. I defied Bo. I sat myself up against Bo.
```

Besides empty complements verbs can of course also have contentive complements. Thus new types for different combinations of empty and contentive complements in different order were needed, totally six new types, but more might be needed for a fuller coverage. Figure 6.6 shows the difference between the Matrix type for a verb with one contentive complement and the BiTSE type for a verb with one empty complement. They both have the same parent, **basic-two-arg**, but the type with the empty complement contains no semantic linking for it, which the type with the contentive argument does.

Figure 6.6: Base types for a verb with one empty complement and with one contentive complement

The Matrix contains base types for up to three arguments. But Swedish can have up to five arguments, as in (13), which has an empty particle and reflexive, and two contentive objects.

```
(13) a. Jag tar med mig det till Kim. I bring it to Kim.

I take with myself it to Kim.
```

I thus also had to include the base types basic-four-arg and basic-five-arg into BiTSE, which have the same function as the Matrix types for one to three arguments, i.e. to ensure correct basic mapping.

#### 6.2.5 Modifiers

As described in section 2.1 modifiers are constituents that can be appended to other constituents adding new information. There are many types of modifiers, in the core of

BiTSE adjectival modifiers of nouns and prepositional modifiers of noun and verb phrases are included.

The basic strategy of modification is that the modifier contains information about what kind of words or phrases they can modify. Unlike complements, which are specified in the valence, modifiers specify what they can modify in the HEAD, in a feature called MOD. The Matrix contains base types for the two basic types of modifiers, scopal and intersective. Intersective modifiers always have the same scope as the constituent it modifies and takes an event or an instance as its argument. Scopal modifiers outscopes the constituent it modifies. Both types of modification described in this section are intersective. Figure 6.7 shows the basic Matrix type for intersective modifiers. The LOCAL value *intersective-mod* is used in the rules for adding modifiers in order to assure that the correct rule is used for each type of modifier.

Figure 6.7: Matrix type for intersective modifiers

#### **Adjectives**

Adjectival modifiers of nouns are also often called adjectival attributes. In both Swedish and English adjectival modifiers of nouns are placed before the noun. Figure 6.8 shows the BiTSE type for adjective lexemes. This type specifies that adjectives modifies nouns, and that the adjective should precede the noun since the value of POSTHEAD is –. In Swedish adjectives also have to agree with the noun they modify for number and definiteness, as in (14) showing different inflections of Swedish "stor" ("big").

(14)	a. Ett <i>stort</i> hus	A big house
	b. En <i>stor</i> dörr	A big door
	c. Det stora huset	The big house
	d. Två <i>stora</i> hus	Two big houses

Thus Swedish adjectives are marked as uninflected in the lexicon, and there are a number of lexical rules that adds inflection to Swedish adjectives, and makes sure that the modified noun agrees with the inflected adjective. In English the adjective always appear in the base form. They are thus marked as inflected in the lexicon, which means that they are handled as words and do not have to go through any lexical rules.

Figure 6.8: BiTSE type for adjective lexemes

Swedish definite singular nouns which are modified takes an obligatory definite determiner, which is not possible for unmodified nouns which can only take a demonstrative determiner, see (15). The fact that definiteness is expressed in two ways, both with a determiner and a suffix is sometimes called "double definiteness" (Hellan & Beermann, 2005). This phenomena is in many respects similar to well discussed phenomena like double negation (e.g. de Swart, 2004) and negative concord (e.g. Przepiórkowski & Kupść, 1996), which is outside the scope of this work.

(15)	a.	Huset är fint	The house is pretty
	b.	Det huset är fint	That house is pretty
	c.	*Stora huset är fint	The big house is pretty
	d.	Det stora huset är fint	The big house is pretty

#### **Prepositions**

Currently only preposition phrases with a noun phrase complement are handled. Other possible complements for Swedish PPs are infinitival clauses and subordinate clauses (Jörgensen & Svensson, 1987), and for English past participle clauses and subordinate clauses. Preposition phrases can be used as verbal complements and modifiers.

It is important that the semantic content of the complement of the preposition is somehow accessible in the prepositional phrase. This is necessary in order to handle some phenomena like binding correctly. In (16a) the information about the noun "himself" is needed to verify that "himself" refers to the same person as the subject "he", and in (16b) that it does not refer to the subject. Binding is not yet included in BiTSE, but it is desirable to have correctly analysed preposition phrases anyway.

```
(16) a. He<sub>i</sub> believes in himself<sub>i</sub> b. *He<sub>i</sub> believes in him<sub>i</sub>
```

The way to achieve this depends on the semantic content of the preposition. If the preposition is empty I follow the suggestion by Pollard & Sag (1994). They claim that the syntactic head of a prepositional phrase is the preposition and the semantic head is the complement. Slightly adopted to MRS this practically means that prepositions have the same index as nouns, and co-index it with the index of the NP-complement in the types for prepositions. Figure 6.9 shows a partial base type for empty preposition lexemes, where the *index* value for the preposition, which contains the features *person*, *number* and *gender* and for pronouns *pronoun type*, is co-indexed with the index of the preposition phrase. Noun phrase complements of prepositions is always in the accusative case which is also constrained in the type.

Figure 6.9: Partial type for preposition lexemes

All prepositional modifiers in BiTSE are intersective and posthead, i.e. appear after the constituent they modify, although other types of prepositional modifiers could be identified in both Swedish and English, and would have to be added if BiTSE were to be extended. Prepositional modifiers consists of a prepositional phrase, and they can modify a noun or verb phrase. Like the type for adjective lexemes the type for preposition lexemes that can appear as modifiers inherits from the Matrix type for intersective modifiers. It further constrains the head type of the constituent it modifies to be either verb or noun.

It is often ambiguous which constituent a prepositional phrase modifies. One such case is the sentence "I hunt lions with a gun", in which case it is ambiguous if "gun" should attach to "hunt" meaning that the hunting took place using a gun, or to "lions" meaning that the lions had a gun. This is a syntactic ambiguity, since semantically it is not probable that lions have guns. Such a sentence gives two alternative analyses and MRSs. This ambiguity is the same in Swedish and English, so both MRSs gives the same English and Swedish sentences as a translation. Here it would thus be desirable to get only one analysis, since

the ambiguity is the same. Crysmann (2004) proposes a solution for this, and other issues with intersective modifiers in German, that he implemented a German HPSG. He extends MRS with a constraint that is added to hoos and allows underspecification of intersective modifier attachment.

## 6.2.6 Polar questions

Polar questions are questions that can be answered by yes or no, for instance "Can he sleep?". Swedish and English have different strategies for polar questions. In Swedish all polar questions are formed by inverting the subject and the main verb as in (17). In English questions involving auxiliaries are formed the same way, as in (18), but for ordinary words auxiliary "do" is needed for the question to be well formed, as in (19).

(17)	Jag sover. I sleep.	Sover jag? Sleep I?
(18)	I can sleep.	Can I sleep?
(19)	I sleep.	Do I sleep?

In order to handle polar questions the boolean features INVerted and AUXiliary were used for verbs. INV indicates if the verb has inverted word order, and AUX indicates if a verb is an auxiliary or not. A lexical rule to invert the order of the verb and subject was introduced, which moves the subject from SUBJ to COMPS. All elements in COMPS are constrained to appear after the verb, and thus the verb now precedes the former subject. This rule can be applied to all Swedish verbs and to English verbs with AUX +.

To be able to handle English questions with the auxiliary "do", a new lexical type, **do-support-lex**, for "do" is introduced, which has an empty semantic relation, since "do" does not contribute semantically in polar questions, but only syntactically by turning the clause into a question. It is marked as AUX + and thus handled in the same way as other English auxiliaries when forming questions. This type is further constrained so that it can only be used in contexts where it should be empty, such as negation, and not in contexts where it is contentive, such as reinforcement, as in "I do sleep".

I follow the idea that questions are open propositions suggested by Bender (2004) and Ginzburg & Sag (2000). This means that the relation for questions, question\_m\_rel, always takes a proposition as an argument. An example of a MRS for a question is shown in Figure 6.10. Note that it has two message relations, one outer, for the question, which has the proposition with "sleep" as its argument. To achieve this a new rule was introduced that adds a question relation to the clause, this rule is the same for all types of polar questions and is applied on complete verb phrases of finite tense with INV +.

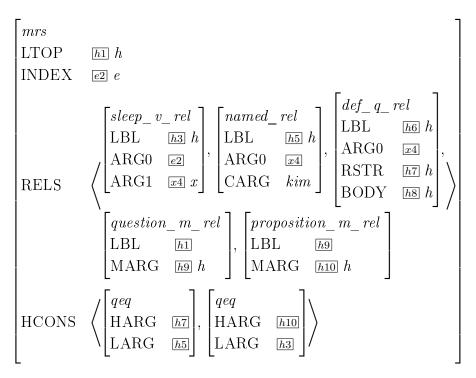


Figure 6.10: MRS for the question "Sover Kim?" ("Does Kim sleep?")

# 7 Solutions for divergence problems

This chapter will discuss the identified VFDs and both suggest solutions for them, and in the cases where solutions are implemented also describe the implementation. It has a section that describes each VFD category as of chapter 5. I have tried to find general solutions for grammatical phenomena, rather then just trying to find the simplest solution for each problem, which means that some grammatical phenomena that are not strictly necessary for VFDs are also discussed. I believe that this strategy makes it easier to add future extensions to the grammar.

Each section discusses some examples of each type, an overview of sentences translated by BiTSE can be found in Appendix B.

# 7.1 Structural divergences

Structural divergences occur when the same logical constituent in two languages have different syntactic categories. These cases can generally be handled by not assigning a semantic relation to purely grammatical constituents. This section describes the four groups of constituents involved in the structural divergences that have been identified: reflexives, empty prepositional complements, idiomatic particles and infinitive markers.

In all these cases one language have one more constituent than the other. The general solution is to treat one of these constituents as empty, i.e. carrying no semantics, and let the other carry all semantic information. Sigurd (1995) suggests a similar solution for particles, reflexives and prepositions in a Prolog definite clause grammar MT system. His general idea is to let all rules concerning verbs have optional slots for these three types of constituents, which let them co-occur with verbs, but do not attribute them any semantic content. These slots are specified in the lexicon for the verbs which takes any of them as complements and ignored else. I use the same basic idea, to treat these constituents as semantically empty and to let verbs specify which of these constituents it needs as complements. I however use the type hierarchy, to specify special types for each type of verb. Infinitive markers are also empty and verbs that take an infinitival complement has to state if it should be headed by an infinitive marker or not.

## 7.1.1 Reflexives

Reflexives are pronouns that refer back to another noun phrase referring to the same person, like "himself"/"sig". I have divided the use of reflexive pronouns into two groups, fake reflexives (Toivonen, 2002), which occur with certain verbs, which I will call pure reflexive verbs, as in (20a), and ordinary reflexives, which occur as objects to normal transitive verbs, as in (20b). A pure reflexive verb takes an obligatory reflexive pronoun as complement.

(20) a. I perjured myself.b. I shaved myself.I shaved him.

There are two basic strategies of what semantic analysis to use. Either the reflexive pronoun can be analysed as empty and the verb carry a semantic relation for both the verb and the reflexive pronoun, as suggested above, or both the verb and the reflexive pronoun could have a relation each. When a pure reflexive verb corresponds to a non-reflexive verb in the other language the first strategy would mean that it could be analysed as it normally would in that language. The second would mean that the verb in some way has to contribute a reflexive pronoun relation as well as its verb sense relation. Thus the first strategy seems more natural since it means that special treatment has to be taken in the language which has the reflexive, rather than on seemingly non-reflexive verbs in the other language, as for "bring" which is a translation of the Swedish reflexive "ta med sig", which would have to add a reflexive pronoun relation if the latter strategy were chosen.

Thus reflexive pronouns appearing as complements to pure reflexive verbs have been analysed as semantically empty, with the pure reflexive verb carrying the relation for the meaning of the verb plus the fake reflexive and selecting the correct reflexive pronoun. Figure 7.1 shows the type for intransitive pure reflexive verbs like "perjure oneself" or "uppföra sig" ("behave oneself"). The type inherits from two BiTSE types, one for verbs in general, and one for lexemes with a semantically empty second argument. The png (person, number and gender) value of the reflexive is co-indexed with that of the subject, to assure agreement.

A normal transitive verb on the other hand can have either a reflexive pronoun or a regular NP as its object. Here the reflexive pronoun is clearly contributing semantically since it contributes the information that the verb expresses a reflexive relation by refering to the subject and not to any other possible object. It is thus analysed as contentive. Reflexive pronouns thus have two entries in the lexicon, one empty and one contentive.

Another possibility that might seem tempting at first would have been to combine verbs and reflexive pronouns in the lexicon. There are however good reasons against this: the verb and the reflexive pronoun are discontinuous, i.e. other words can appear between

Figure 7.1: BiTSE type for pure reflexive verbs

them as in (21). In addition, the fact that the pronoun varies on person also makes it difficult to combine the words in the lexicon.

```
(21) Kom han inte upp sig? Did he not succeed? Came he not up himself?
```

Reflexive pronouns have to agree with their antecedents, generally with the subject. This is incorporated for pure reflexive verbs, as shown above, but not for normal transitive verbs. For English the agreement is not much of a problem in an MT application, since non-agreeing reflexives are not likely to occur. For Swedish it is problematic since reflexive pronouns and accusative personal pronouns are the same in first and second person, "mig" means both "me" and "myself". A pronoun object is reflexive if it agrees on person, number and gender with the subject. There is however no simple way to express this with the TFS version used, but binding theory would have had to be used, which is outside the scope of this study. It is thus unresolved if a Swedish pronoun is reflexive or personal. This means that the semantic analysis obtained from Swedish sentences with a pronoun object will be ambiguous and in English two sentences will be generated from it, one of which will be both ungrammatical and incorrect as a translation.

Another issue in connection to reflexives are, as Toivonen (2002) points out, the way that the word "själv" (self) can be used in connection with reflexives in Swedish. How this happens seems to be different for fake and ordinary reflexives. There are also other differences between the use of reflexive versus non-reflexive pronouns in English and Swedish. This has not been not investigated.

## 7.1.2 Empty prepositional complements

The structural divergence including prepositional complements only concerns prepositional complements with empty prepositions as in (22a), not prepositional complements with contentive prepositions, as in (22b).

- (22) a. I depend on him
  - b. I put it on the table

Verbs taking empty prepositional complements usually only accept them to be headed by one or possibly a limited set of prepositions; for instance, "depend" in (22a) always takes the preposition phrase headed by "on" as a complement, other prepositions, such as "under" or "at" are impossible. For this to be achieved the verb has to specify what the legal possible preposition(s) of its complement are. This selection has to be syntactic, since these prepositions do not have any semantic information that could be used for selection. To achieve this prepositions have a feature PFORM, as suggested by e.g. Tseng (2000). PFORM-values are organised in a hierarchy making it possible to choose either a specific preposition or a class of prepositions such as dative prepositions in Swedish.

To simplify lexical entries for verbs with empty prepositional complements verbs have a head feature GOVPR that is co-indexed with the PFORM of its prepositional complement. GOVPR is thus basically a pointer to the PFORM of the prepositional complement. The reason for introducing GOVPR, instead of using PFORM directly is that the prepositional complement can be in different positions of the comps-list, which means that the placement of the *pform* value in the lexical entries would have been different for different lexemes. Now the mapping between the prepositional complements and GOVPR is done in the lexical types, which simplifies the lexical entries, as Figure 7.2 shows. This depends on the assumption that no verb has more than one syntactically chosen prepositional complement, which is currently true in BiTSE. There are however verbs with more than one prepositional complement, as in (23), but at least "about" should probably be considered contentive. In case examples are found of verbs with more than one syntactically chosen prepositional complement, the current model could be extended by changing GOVPR from corresponding to a single PFORM, to a list of PFORM values corresponding to each preposition.

#### (23) talk to someone about something

Currently syntactic selection, by PFORM, is the only possible way for a verb to choose a prepositional complement. This is suitable for empty prepositions, but for contentive prepositions this is not as suitable, they should be chosen semantically (Tseng, 2000). These verbs can generally take more than one possible preposition, for instance all locational prepositions in (24). Tseng (2000) also points out that there is sometimes no

Figure 7.2: Actual lexical entry for a verb with an empty prepositional complement (top), the alternative entry without govpr (bottom)

clear line between semantic and syntactic selection, and that some prepositions should be chosen by combining the selection methods.

(24) I put it on/under/behind/... the table.

## 7.1.3 Particles

Particles are "small" words that are closely connected to a verb. They often coincide with prepositions or adverbs, but unlike prepositions they are always stressed. Verb and particle combinations are often called phrasal verbs (Sroka, 1972). Especially in Swedish particles can be of many different parts of speech, as illustrated in Table 7.1. Particles also have several different functions. Toivonen (2003) divides the function of Swedish particles into two main groups, resultative (including directional) and aspectual. She further identifies idiomatic particles, which can be either resultative, aspectual or unclear, in which case she analysis the verb-particle combinations as complex predicates. The same particle can appear in all functions in some cases. In English particles are not as common as in Swedish and are usually adverbs or prepositions.

Toivonen further claims that even though different particle constructions have different semantics they are all alike syntactically. I do not agree on this since I think that most

Word class	Examples
none	gå an (be proper)
	go PART
preposition	stå i (be busy)
	stand in
adverb	åka hem (go home)
	go home
noun	äga rum (take place)
	own room
adjective	bryta löst (start)
	brake loose
present participle	göra gällande (claim)
	make valid
lexicalized group of words	ha på känn (sense)
	have on feel

Table 7.1: Examples of Swedish particles of different word classes

idiomatic particles are part of the verb frame for the verb whose semantics includes the particle. Resultative and aspectual particles on the other hand are generally modifiers.

In Swedish particles are always placed before the objects, in English they can generally be placed either before or after the direct object. This can give rise to a syntactic VFD.

## Idiomatic particles

Idiomatic particle-verb combinations are cases where the semantics of the verb plus particle cannot be derived from the semantics of the verb and the particle respectively. There are also cases where the particle+verb combination is semi idiomatic, the total semantics have something to do with the respective semantic components, but only in a very limited sense, an example of this is "throw up" meaning "vomit".

In some cases the idiomatic uses are the same in Swedish and English, as in "throw up" versus "kasta upp". In these cases nothing specific needs to be done to achieve a translation of the "vomit"-sense as opposed to the "heave something to a higher physical level"-sense. In most cases though, the idiomatic combinations are different in the two languages, and a common semantic representation for a plain verb and a phrasal verb has to be found.

To achieve this I use the same basic technique as for prepositional complements, the idiomatic particles are treated as empty constituents, and the verb carries information

about which particle it needs for a specific sense of the verb. As for prepositions, particles also carries the PFORM feature. Parallel to the GOVPR feature used for prepositional complements, a GOVPA feature is introduced that is linked to the PFORM of the particle. The particle does not carry a semantic relation, and the verb carries a relation that carries the semantic meaning of the verb+particle combination. See Figure 7.3 for a phrasal verb type and lexical entry.

```
Type for intransitive phrasal verbs
intrans-part-verb-lex := ord-verb-lex & intrans-empty2ndarg-lex-item &
 [ SYNSEM.LOCAL [ CAT [ HEAD.GOVPA #pform,
                  VAL [ SPR < >,
                        COMPS < #part & [ OPT - ] >,
                        SUBJ < #subj & [ LOCAL.CONT.HOOK.INDEX #xarg] >,
                        SPEC < > ] ],
                  CONT.HOOK.XARG #xarg],
   ARG-ST < #subj & [ LOCAL.CAT.HEAD noun &
                                      [ CASE nom ] ],
            #part & [ LOCAL.CAT.HEAD part &
                                      [ PFORM #pform ] ] > ].
Lexical entry for "qå av" ("break")
gå-av := intrans-part-verb-lex &
  [ STEM < "gå" >,
    SYNSEM [ LANG sw,
             LKEYS.KEYREL.PRED "break_v_rel",
             LOCAL.CAT.HEAD [ CONJUG r-ir,
                              GOVPA av-pf ] ].
```

Figure 7.3: BiTSE type for intransitive phrasal verbs

It is often the case that a verb can combine with different particles, giving the verb different meanings. In this case the verb will have a different lexical entry for each verb+particle combination. If the verb can appear without a particle it also has a lexical entry for its particle-less sense. Table 7.2 shows how the meaning of Swedish "gå" can differ with different particles. Some of the verb+particle combinations in the table have more than one possible meaning, and would have to be resolved for ambiguity in some further way.

## 7.1.4 Infinitive markers

Verbs in the infinitive verbform can appear with or without the infinitive marker, "to" in English and "att" in Swedish. Verbs that take infinitival complements ususally specifies

Verb	Particle	English translation	Proposed relation
Gå	=	walk	walk_rel
Gå	an	suffice	suffice_rel
Gå	bort	die	die_rel
Gå	om	be repeated	repeat_v_rel

Table 7.2: The meaning of "gå" ("go") with different particles

whether the complement should have an infinitive marker or not. When an infinitive marker is needed in one language, but not in the other, as in (25), it leads to a divergence.

A common suggestion of how to treat the infinitive marker "to" in English is to analyse it as an empty auxiliary verb (Sag et al., 2003; Pollard & Sag, 1994). Sag et al. (2003) constrain it to have verbform (or FORM in their terminology) base, which means that it can not be inflected. They further introduce a new boolean feature INF, which is + for "to" and - for all other verbs. The auxiliary "to" is then constrained to take an infinitival complement. Other verbs that take an infinitival complement can then distinguish between one headed by "to", which is INF +, and one without marker which is INF -. Since "to" is empty it does not contribute to the MRS, and divergent sentences will get the same MRS, regardless of the presence or absence of an infinitive marker. Since Swedish infinitive markers have a similar distribution this analysis could be expected to work well for Swedish as well.

# 7.2 Conflational divergences

This section first discusses conflational divergences in general, and suggests and discusses some possible general ways to handle them. Then it discusses understood optional complements, of which some cases are included in BiTSE, in some more detail.

Conflational divergences occurs when an argument that is explicit in one language is implicit, or conflated, in the other language. Such as "begå mened" - "perjure oneself", "svara" - "give an answer".

As for any divergence there is always a choice of which language is seen as divergent from the "norm" or interlingua. For conflational VFDs this means that there is a choice of handling the divergence in the language where the conflated argument is present or where it is missing. This choice also means that the concept will be represented either with a single relation, in case the language where it is conflated is seen as the norm, and with two relations, one for the verb, and one for the argument, if the non-conflated language is seen as the norm.

If conflational divergences are handled in the language where it is missing, that would mean that the implicit argument should be part of the MRS for the phrase in question. This could be achieved by letting the verb have an extra relation for the missing argument, that is properly linked to the verb's relation. There is currently no type for a sign with two relations in the Matrix so this would have to be added either in BiTSE or possibly in a future release of the Matrix if this approach were chosen. Another possible way to handle this would be to have a feature representing the missing argument. There would also be the need for a rule adding the relation, and the correct mapping for it. The feature representing the missing argument could either choose it syntactically, with a feature like PFORM for prepositions, or semantically, by specifying the relation it wants in some way.

If conflational divergences instead are handled in the language where the argument is present a construction that is somewhat similar to the handling of reflexives and idiomatic particles could be used. This would mean that the verb carries the relation for the whole concept in question, and information about which word it has to combine with to achieve this concept. Then at least one rule would have to be added to combine the conflated verb with its complement. This means that the argument in question cannot contribute a relation, which in turn means that it either has to be empty, or the rule that adds it to the verb has to remove its relation. If it is empty it would be possible to use the same rule that combines reflexives and idiomatic particles with verbs, but it would also mean that all words that can be conflated in any other language in the system has to have two entries in the lexicon, one empty, and one contentive, which leads to a lot of redundant information having to be stored. The other strategy, to remove the relation, goes against one of the basic assumptions of HPSG.

The first approach seems to be less problematic to incorporate, and there are also examples that seems to suggest that the argument is implicit in the language where it is missing for at least some conflational divergences. (26) shows some variations of the phrase "I blow my nose" in Swedish. In the basic variant "nose" is conflated into the Swedish pure reflexive verb "snyta", and the word "nose" can not be added to this without it sounding un-idiomatic. However, (26c) shows that the conflated argument can be present if it has a modifier. This suggest that there should be a nose-relation in Swedish as well as in English.

(26) a. Jag snyter mig I blow my nose

I blow.nose me

b. ?Jag snyter min näsa I blow my nose

I blow.nose my nose

c. Jag snyter min röda näsa I blow my red nose

I blow.nose my red nose

There is also a possibility that different conflational arguments have to be treated differently, some with one relation, some with two. If one relation were chosen in some case the difficulties associated with that would somehow have to be solved, or possibly solved with another approach such as semantic transfer.

## 7.2.1 Understood optional complements

It is possible for verbs (and other constituents) to have optional arguments. One common example is an optional object such as "I eat", where the object, what is actually eaten, is left out. In this case the left out argument is unspecified; even though we have no idea what is being eaten, we know that it should be some kind of food. This type of optionality is the same for English and Swedish, the Swedish equivalent to "I eat" is "jag äter". There are also cases when the left out complement is not unspecified, but understood in some way. One of the most common examples of this appears in prodrop languages like Spanish where the subject can be left out, but is obvious from the inflection of the verb, as can be seen in (27). This phenomena does not occur in either Swedish or English.

One case where there is a divergence between Swedish and English is with respect to understood reflexive objects. In English some verbs that concerns caring for the whole body are implicitly reflexive when the object is unspecified (Levin, 1993). An example of this is "dress" which in its intransitive version, "I dress", is reflexive, the one getting dressed is obviously the subject, "I", which is then a conflated argument. If the object is not left out "dress" behaves like a normal transitive verb, and the object can refer to anyone. Swedish does not have this type of optionality so the proper translation for "I dress" is "jag klär på mig" ("I dress on myself").

To be able to handle the fact that the removal of optional complements should be able to result either in that role being left unspecified or that a semantic relation should be added, a new feature, OPTTYPE, is introduced on SYNSEMs. It describes what the synsem should be replaced with if it is removed as optional. The default value for it is *unspec*, which

means that removing the optional complement should result in it being left unspecified, as for "I eat". The unary rule that removes unspecified optional complements is then constrained to only work for complements with OPTTYPE unspec.

This clearly seem to be a divergence type which should be handled in the language where it is optional, i.e. where the argument can be conflated, because it is present there sometimes. Thus the relation for a reflexive pronoun agreeing with the subject should be added when the argument is missing. For verbs like "dress" where removal of the object should result in a reflexive relation being added, the object get OPTTYPE refl-opt. A unary phrasal rule that adds a relation for a reflexive pronoun when removing an optional complement is then added. This rule is constrained to work only for OPTTYPE refl-opt. It further controls that the reflexive pronoun agrees with the subject.

This solution can be used for any other type of understood optional complement that might be identified in the future. What is needed is to add a new value type for OPTTYPE for the new case, and a specific rule that inserts the particular semantics that should replace the optional complement. Every verb that has optional complements then has to set the OPTTYPE value for its complement to the correct value. This can be illustrated by a different case of understood optional complement, Swedish "hitta" ("find"). When a complement is present it behaves like "find" in English, but when it is missing it is equivalent to "find the way", see (28).

(28) a. Jag hittar boken
I find book.the
b. Jag hittar
I find
I find the way
I find

A possible way to solve this using the ideas above by introducing a new value for OPTTYPE way-opt, or similar, and creating a rule that adds a relation for "way" and a relation for a definite quantifier and links them correctly when removing the complement from "hitta". However, this leads to a specific rule having be created for each conflated argument, and some way of generalising the adding of the relation for the conflated argument would be preferable.

# 7.3 Head inversion divergences

Head inversion occurs when a main verb in one language corresponds to another constituent in the other language. Instances where it corresponded to adverbs and aspectual particles were found, and these two cases will be described in this chapter.

## 7.3.1 Adverb modifiers

In these head inversion divergences an adverb in one language corresponds to a main verb in the other. Usually this involves a scopal adverb and a raising verb, as in (29) where "brukar" is a raising verb and "usually" is a scopal adverb.

(29) Bob brukar sova Bob tends sleep

Bob usually sleeps

The standard Matrix types for this were incorporated into BiTSE so that I could see what the differences in the MRSs for such sentences would be. Figure 7.4 shows MRSs for the sentences in (29) for the raising verb and scopal adverb. As can be seen the MRSs are very similar, the only difference is in scope, "brukar" has the handle for "sleep" directly as an argument, while "usually" has it scoped, through a handle constraint. Even though these two MRSs are identical except for scope, generation does not give the other version. This could be solved by changing the type for either of these to fit with the other. Figure 7.5 shows the Matrix type for scopal modifiers, which is an ancestor of the type for scopal adverbs, and an alternative version giving scopal modifiers the same semantics as raising verbs. As can be seen the change is minor, only the handle constraint is removed. Using this type instead of the old one gives identical MRSs for the two sentences above, and is thus a possible solution to this divergence. It would also have been possible to instead change the type for raising verbs in a similar manner.

There is then a further choice between changing the definition generally for all scopal adverbs or raising verbs, or to have two versions of each constituent and let only adverbs/verbs that take part in divergences have the changed type. If all scopal adverbs or raising verbs are changed, the only thing needed for adverbial head inversion divergences is to make sure the relation names are the same for the adverb and verb, as they are in figure 7.4. If the general definition of scopal adverbs is changed it is necessary to investigate how that affects other phenomena in more complex sentences, which has not been done. It is possible that this change would lead to undesirable side effects, in which case other solutions had to be sought.

# 7.3.2 Aspectual particles

Aspectual particles are particles that change the aspect of the verb it modifies, as in "I drank *up* the beer", where the particle "up" expresses that the drinking, and the beer, is finished, i.e. it changes the event from untelic, with no specific end, to telic, with a clear end. Examples of Swedish sentences with aspectual particles, and an equivalent divergent English translation can be seen in (30). In (30a) and (30b) the English equivalent is a verb with a present participle complement, and in (30c) it has an adverbial modifier.

"Bob brukar sova"

$$\begin{bmatrix} mrs \\ LTOP & h1 h \\ INDEX & 22 e \end{bmatrix}$$

$$\begin{bmatrix} named\_rel \\ LBL & h3 h \\ ARG0 & 24 x \\ CARG & bob \end{bmatrix}, \begin{bmatrix} def\_q\_rel \\ LBL & 55 h \\ ARG0 & 24 \\ RSTR & 66 h \\ BODY & 17 h \end{bmatrix}, \begin{bmatrix} usually\_rel \\ LBL & 188 h \\ ARG0 & 22 \\ ARG1 & 190 h \end{bmatrix}, \begin{bmatrix} sleep\_v\_rel \\ LBL & 190 \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 190 \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 11 \\ MARG & 11 h \end{bmatrix}$$

$$\begin{bmatrix} mrs \\ LTOP & 16 h \\ INDEX & 22 e \end{bmatrix}$$

$$\begin{bmatrix} mrs \\ LTOP & 16 h \\ ARG0 & 24 x \\ CARG & bob \end{bmatrix}, \begin{bmatrix} def\_q\_rel \\ LBL & 165 h \\ ARG0 & 24 \\ RSTR & 166 h \\ BODY & 17 h \end{bmatrix}, \begin{bmatrix} usually\_rel \\ LBL & 188 h \\ ARG0 & 24 \\ RSTR & 166 h \\ BODY & 17 h \end{bmatrix}, \begin{bmatrix} usually\_rel \\ LBL & 188 h \\ ARG0 & 24 \\ RSTR & 166 h \\ BODY & 17 h \end{bmatrix}$$

$$\begin{bmatrix} sleep\_v\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}$$

$$\begin{bmatrix} sleep\_v\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}$$

$$\begin{bmatrix} sleep\_v\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}, \begin{bmatrix} proposition\_m\_rel \\ LBL & 1610 h \\ ARG0 & 211 e \\ ARG1 & 24 \end{bmatrix}$$

Figure 7.4: MRSs for sentences containing a raising verb (top) and a scopal adverb (bottom)

Figure 7.5: Matrix and alternative types for scopal modifiers

```
(30) a. Jag sprang på I kept running
I ran on
b. Jag åt upp I finished eating
I ate up
c. Jag hostade till I coughed once
I coughed to
```

Aspect is a complex issue. The basic meaning of each verb obviously affects the aspect of the clause it is part of, but other constituents also affects it, besides aspectual particles, e.g. prepositional modifiers. Thus aspect will not be handled in full here, but only discussed in connection with particles.

Toivonen (2003) has a theoretic solution for analysing aspectual particles, which is developed in the LFG framework, but that, as she points out, should be possible to carry over to other frameworks like HPSG quite easily. She divides aspect into three boolean features, telic, dynamic and durative. Each verb has values for each of the three features, with each feature set to either a default value or a final value. Default values can later be changed if the verb is combined with a constituent with another value for that feature. As an example activity verbs like "run" are specified to be dynamic and durative, but

telicity is unspecified with the default set to untelic. Later other constituents can change the value of the telicity, like the aspectual particles in (30).

Aspectual particles also have their aspect set, the particle "på" ("on") in (30a) above is telic –, dynamic + and durative +. The resulting aspect of combining a verb and a particle is the result of unifying the aspect of the particle and of the verb. If some non-default feature does not agree, the combination is not allowed, as "\*The man knew on", where "know" is dynamic –, and "on" is dynamic +, which makes them incompatible. If this analysis were incorporated into BiTSE it would mean that aspectual particles would not carry semantic relations, they would only change the aspect information of its head.

Another solution for aspectual particles are sketched by Villavicencio & Copestake (2002). They discuss this in connection to the treatment of particles in LinGO ERG. They discuss the example "he tore up the letter" which they give the semantics of (31), where the particle causes two relations to be added, an up-relation that modifies the event, and a fully\_affected-relation that affects the object. They treat aspectual particles as complements, not as modifiers as suggested here. First they have a lexical rule that adds an aspectual particle to the comps-list of verbs that can take one, and also adds the fully\_affected-relation. The up-relation is then directly contributed by the particle when the verb is combined with it. The big difference from Toivonen's analysis is that a relation is added that carries the information that something has ended, instead of changing the telicity to telic for the tearing event. The fact that up contributes an up-relation has in my opinion more to do with the fact that all particles contribute a relation in ERG, than with the fact that it actually contributes any kind of "up"-information to the clause, other than fully\_affected.

(31) 
$$tear(e1,x,y) \wedge he(x) \wedge letter(y) \wedge up(e1) \wedge fully affected(y)$$

Because aspect is such a complicated phenomenon it has not been included in BiTSE, but I think that both Toivonen's and Copestake & Villavicencio's strategies are useful, and would be possible to incorporate into BiTSE. In order to chose one over the other it would be useful to investigate other aspectual phenomena and see how they fit into the two strategies, to be able to achieve a uniform treatment of aspect if possible.

To solve the divergences caused by aspectual particles it is, besides choosing an appropriate analysis for aspectual particles, also necessary to find out if this analysis is compatible with the analysis of the divergent translation. Not until then can it be settled if it is possible to find an equivalent MRS, or if some kind of transfer would be more suitable.

# 7.4 Categorial divergences

Categorial divergences occur when the same semantic concept are expressed by constituents of different categories in different languages. This is a difficult problem, which I have no finished solution for.

In all the identified cases a copula verb is involved in one of the languages. The two used copula verbs are "vara" ("be") which expresses some kind of state, and "get" which expresses some kind of change of state. Copula verbs also contains tense information. Copulas are used much both in Swedish and English and to do a specific interpretation of them just to fit these divergences does neither seem plausible nor practical. The basic meaning of the copulas also seems to hold for all examples. Some kind of special care thus have to be taken either with the predicate of the copula verb, or with the corresponding constituent in the other language.

(32) a. Jag smutsade ner mig I got dirty
 I dirted down myself
 b. Jag gifte mig I got married
 I married myself

(32) shows two categorial divergences that contain the word "get", which indicates some sort of change of state, from clean to dirty and from single to married. In (32a) the Swedish verb "smutsa" contains the "dirty"-sense and some part of the "state change"-sense, with the resultative particle "ner" also contributing to the "change state"-sense. In (32b) the verb "gifta" contains the whole notion of "moving into marriage", which in itself is a change of state. The reflexive is empty and contributes nothing (see Section 7.1.3). In this particular case it might thus be possible to let "gifta" carry two semantic relations, one for the state change, and one for the marrying. As noted earlier there are no Matrix types with more than one relation, so this would mean an expansion of the basic set of types. In the other example, the situation is more difficult since the "state change"-sense is divided between two constituents. More research has to go into this, and other categorial divergences before finding a final solution.

This however seems to be a case where transfer could be useful. It would be possible to line up the divergent pairs and their differences in a transfer dictionary of some kind. This has been done by Abeillé (1990, in Dorr 1993) in a tree adjoining grammar, a paradigm quite different from HPSG, by aligning the trees of divergent pairs in a transfer lexicon. This is however, as Dorr points out burdensome with a large number of categorial divergences. This would probably be more suitable for the semantic transfer architecture described in Section 3.4.1. In writing transfer rules for MRSs it is possible to utilize the fact that relations can be organised in the type hierarchy, in order to write base types of transfer rules for a class of divergence types.

# 7.5 Syntactic divergences

The identified cases of syntactic divergences are both concerned with word order. In this section I will discuss the ditransitive alternation, which concerns word order in both languages, as opposed to English particle placement, that only concerns English, and could be dealt with language internally. All divergences concerning word order can be handled in a manner similar to that of those caused by ditransitive alternations, handling it either language internally or in the same way for both languages.

## 7.5.1 Ditransitive alternations

In both Swedish and English there are two basic valence patterns for ditransitive verbs, simple and prepositional. In the simple pattern both objects are noun phrases and the indirect object precedes the direct object, as in (33a). In the prepositional pattern the direct object is a noun phrase and the indirect object is a preposition phrase, as in (33b).

- (33) a. I gave him a book.
  - b. I delivered a package to her.

The preposition in the prepositional pattern varies with the verb in English. This pattern has sometimes been subdivided based on the preposition, Levin (1993) calls the alternation between the simple and prepositional patterns dative if the preposition is "to" and benefective if the preposition is "for". In the current work this difference will be handled by different subtypes of ditransitive verbs for each preposition. In Swedish the preposition also varies with the verb, but for many verbs there is a free choice between the prepositions "till" and "åt", with no apparent difference in meaning, as in (34)

(34) a. Jag gav den till honom I gave it to him
I gave it to him
b. Jag gav den åt honom I gave it to him
I gave it PREP him

Individual verbs can appear either in only one of the patterns or in both. Table 7.3 shows examples of all these types of verbs, with the English grammaticality judgements taken from Levin (1993). The two patterns are considered semantically equivalent and they give rise to the same MRS.

The distribution of the legal patterns for each verb are different for Swedish and English. (35) - (37) shows some examples of equivalent verbs with different distributions, with the English grammaticality judgements taken from Levin (1993). Thus this gives rise

	simple pattern	prepositional pattern
simple	I issued her a permit.	*I issued a permit to her.
prepositional	*I delivered him a book.	I delivered a book to him.
both	I gave him a book.	I gave a book to him.

Table 7.3: Ditransitive verbs with different valence patterns

to a divergence in the cases where the legal patterns are different for an equivalent verb in the two languages. But since the same divergence exist within each language this is categorized as a syntactic divergence. The solution for it is automatically achieved by mapping the two patterns to the same MRS.

- (36) \*I recommended him I recommended the shop
  Jag rekommenderade honom Jag rekommenderade butiken
  the shop. to him.
  butiken. för honom.
- (37) I promised him two I promised two dollars to Jag lovade honom två \*Jag lovade två dollar åt dollars. dollar. honom.

To solve this the grammar includes lexical types for both patterns. The lexical type for the prepositional pattern has a GOVPR value that takes the *pform* of the preposition used. For the Swedish verbs that can take either "till" or "åt" there is a supertype for the *pform* value for these prepositions. There is also a lexical rule to go from the prepositional pattern to the simple pattern. To assure that this rule is only applied to the correct verbs a boolean feature SHIFTS is introduced that allows or blocks the application of the valence changing rule. The different types of verbs are handled as:

<sup>&</sup>quot;simple pattern"-verbs have the simple pattern in the lexicon.

<sup>&</sup>quot;prepositional pattern"-verbs have the prepositional pattern in the lexicon, GOVPR constrains possible prepositions, and SHIFTS –

<sup>&</sup>quot;both pattern"-verbs have the prepositional pattern in the lexicon, GOVPR constrains possible prepositions, and SHIFTS +

# 8 Discussion

In this chapter the findings of the thesis, and issues in connection with that are discussed. It will also discuss the contributions of the thesis and propose directions of future research.

# 8.1 Multilingual grammars

Grammars with more than one language are not common. The author is only aware of a small grammar for the mainland Scandinavian languages (Søgaard & Haugereid, 2005). Thus it was interesting to see how well it would work to have Swedish and English in one grammar. Though these languages are also closely related they are not as close as the Scandinavian languages.

Having two languages in one grammar worked well, and big parts, especially of the core grammar could be the same for both languages. Necessarily the lexicons and inflectional rules had to be distinct, even though there are a common lexicon of proper nouns. The main type hierarchy were divided into three files, one common, and one separate for each language. Table 8.1 shows the size of each file. As can be seen the common file is by far largest both when considering number of types and number of lines. The Swedish file contains more than double the amount of types than the English, but this is to a large extent due to the big number of Swedish simple types for declination and conjugation which are not needed in English. Swedish also has a larger number of inflectional rule types since there are more types of inflection than in English, such as inflection of adjectives and for definite form of nouns. There are also many types for verbs with particle complements that are currently only used for Swedish, but that maybe could be used for English as well, only allowing a freer order of the complements.

Table 8.1: Size of the different language parts of BiTSE

	Number of lines	Number of types	Types for conjugation
			and declination
Common	1363	188	_
Swedish	373	76	27
English	403	32	_

The common part contains many types for lexemes. A majority of the lexeme types are shared, although they are further specialised by each language in several cases. All major phrase types are common. This is possible since most of the information are in the lexemes, and that they instead are language dependent in many cases. The common file further contains a large number of specifications of subtypes of **sort**, for instance **verb form**, **tense** and **gender**. It also contains the base types for empty complements and types for handling language.

The number of language specific verb lexemes is higher than for other word classes, mostly because verbs are the focus of the study, and thus are more specialised than other word classes. It could be expected that if the grammar were to grow, the percentage of types that are language specific might increase.

Since such a big part of the grammar could be shared this grammar design clearly is useful since it reduces the redundancy of entering the same information in one English and one Swedish grammar. Even though the grammar contains two languages the modularity of the grammar files makes it possible to exclude the parts that are only relevant for one language. In a way the common types could be seen as an extension to the Matrix for just two languages, just like the work of Søgaard & Haugereid (2005) was a step in the process of creating a Matrix for the Scandinavian languages.

# 8.2 MRS as interlingua

I hypothesized that there is a border where the semantic interlingua approach is not useful anymore and where semantic transfer is needed. I have not been able to identify a clear border. What I have done is to identify a number of cases where the interlingua approach works very well. For other cases I have proposed tentative solutions using interlingua. For other cases it is more unclear if the interlingua approach will work, it might turn out to be cumbersome to implement, for instance because it requires a large number of specific grammar rules to be written, something that would not be desirable.

Table 8.2 shows the identified divergences, their current status and their judged suitability for the interlingua approach. As can be seen most of the structural and syntactic divergences have solutions and are implemented in BiTSE. These divergences were judged as suitable for an interlingua approach and the implementation of them was quite straight forward.

For the other classes of divergences no finished solution was implemented in BiTSE. Tentative solutions were however discussed for many cases. Of these tentative solutions I was not able to find a clear border where interlingua could not be used, but there are cases where it seems more feasible than others. As an example I have implemented a solution for a subset of conflational divergences, understood reflexive pronouns, which works well. For

Table 8.2: VFDs, their current coverage and	a judgement of how	suitable they	are to be handled
by an interlingua approach			

Main divergence	Specific type	Tentative solution	Implemented solution	Suitable for interlingua
		301461011	501411011	micimigaa
Structural	particles	yes	yes	yes
	empty prepositions	yes	yes	yes
	reflexives	yes	yes	yes
	infinitive markers	yes	no	yes
Conflational	generally	sketch	no	possibly
	understood	yes	some	at least for
	optional			some cases
	complement			
Head inversion	adverbial modifiers	sketch	partly	possibly
	aspectual particles	sketch	no	possibly
Categorial	generally	no	_	?
Syntactic	word order	sketch	no	yes
	ditransitive	yes	yes	yes
	alternations			

other cases of understood complements it is in principle possible to implement a similar solution, but in its simplest version it leads to one new rule for every new divergence with a new conflated argument. This is not desirable, since adding new words of an existing class of words to the lexicon of a grammar should preferably not include changing the grammar, just using its types. It might however be possible to find a way to implement more general rules for conflated complements, that would simplify adding new words involved in conflation to the lexicon. This is also a case where it should be quite straight forward to write transfer rules. An interlingual solution should preferably not be much more complicated than competing strategies. More research has to go into this and other issues before deciding if these divergences are suitable to handle by an interlingua system.

The interlingua approach also depends on the fact that relations that are considered equivalent have the same name in both languages. In BiTSE this is achieved by using English relation names in Swedish. This becomes problematic when ambiguity is considered. Ambiguity was not part of this study, but it is easy to foresee problems with the current relation naming strategy when considering ambiguity. Consider the two English words "roof" and "ceiling" which both would be translated by Swedish "tak". Here it would be impossible to choose either English word as the relation name for "tak", and some type of disambiguation would have to take place on MRSs before generation. It

might to some extent be possible to utilize the type hierarchy for this, by creating an inheritance hierarchy of relations.

Although semantic interlingua does not seem to be able to cover all divergences that exist when translating between Swedish and English in any reasonable way I think it can be of use in a semantic transfer system, by minimising the amount of transfer that has to be done. In the standard DELPH-IN semantic transfer strategy some interlingual elements are already used, such as negation-relations. I have shown that it is possible to increase the number of interlingual elements used in translation between Swedish and English.

## 8.3 BiTSE and the Matrix

It is beneficial to the Matrix project that Matrix-based grammars are developed for new languages. The work with BiTSE thus had the potential of showing how well Swedish would fit the Matrix. As could be expected it worked fine since Swedish is closely related to Norwegian and English, for which Matrix-based grammars already exists. There were, however, one case where appropriate base types were missed. This was in connection with empty constituents, particularly empty complements, for which there were no types that mapped them correctly. Also there were types allowing up to three complements, and it turned out that Swedish needed up to five complements. Types for covering this were suggested in Section 6.2.4.

In some of the tentative suggestion for VFDs I have suggested lexemes with two relations, instead of the normal one (or zero). If these suggestion were to be implemented base types for these lexemes would be needed either as part of the Matrix or of BiTSE.

## 8.4 Contributions

The main contributions of this thesis are:

- An inventory of English and Swedish verb frames and verb frame divergences
- BiTSE, a bilingual grammar for Swedish and English covering the core of the languages and some verb frame divergences, with a common semantic representation
- Showing that it is possible to use MRS as an interlingua in machine translation in many cases

## 8.5 Future work

As this thesis mainly consisted of explorative work there are many possible research directions that could be based on it.

It would be interesting to extend the divergence classifications for Swedish and English for other types of divergences. It would also be interesting to do a divergence study towards a translation corpus, where one could see what divergences actually exists, and also how common they are. In this type of study both VFDs and other divergences can be studied.

Another interesting issue would be to extend BiTSE further. This could be done with different foci:

- extend it as a bilingual grammar keeping the focus on Swedish-English divergences by
  - extending the coverage of VFDs, both by the suggestions in this thesis, and further than that
  - incorporating other types of divergences than those concerning verb frames
- extend only the Swedish part in the ultimate goal of creating a broad coverage HPSG for Swedish, since there is currently a lack of Swedish publicly available broad coverage computational grammars
- extend it with one or several other languages, which would give a multilingual grammar and MT system, and show how well the interlingua strategy works in a more general setting
- extend it for other applications than MT where semantic information is needed such as question-answering systems

There are also many issues within the machine translation field that were not addressed in this study, but would be interesting to explore in the future. One such issue is ambiguity as discussed above. Another is lexical acquisition. It is not desirable to add every word to the lexicon by hand, so finding ways to map words to the correct type, with the correct relation name is a challenge. Yet other issues are robustness and efficiency.

## 8.6 Conclusion

In this study a classification of Swedish-English verb frame divergences have been made. It has also shown that many of the issues concerning the identified VFDs that are handled

by semantic transfer in the general DELPH-IN MT design, can naturally be handled by an interlingual design, minimising the need of transfer. It has resulted in BiTSE, a bilingual grammar of Swedish and English, covering the core of the two languages and a subset of the identified VFDs. In this grammar more than half the types are common for the two languages, reducing the redundancy of having two separate grammars.

# References

- Aitchison, Jean (1999) Linguistics. London: Hodder & Stoughton, 5th ed.
- Arnold, Doug; Lorna Balkan; Siety Meijer; R. Lee Humphreys & Louisa Sadler (1994)

  Machine Translation An Introductory Guide. London: Blackwells-NCC.
- Barnett, Jim; Inderjeet Mani; Paul Martin & Elaine Rich (1991) "Reversible Machine Translation: What to do when the Languages don't Line Up". In *Proceedings of the Workshop on Reversible Grammars in Natural Language Processing, ACL-91*, pp. 61–70, University of California, Berkely, US.
- Bender, Emily M. (2004) "Linguistics 471: Grammar Engineering". http://courses.washington.edu/ling471/, last visited Mar 28, 2006.
- Bender, Emily M. (2005) "Linguistics 567: Grammar Engineering". http://courses.washington.edu/ling567/, last visited Mar 28, 2006.
- Bender, Emily M.; Dan Flickinger; Frederik Fouvry & Melanie Siegel, editors (2003) Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development, ESSLLI 2003, Vienna, Austria.
- Bender, Emily M.; Dan Flickinger & Stephan Oepen (2002) "The Grammar Matrix: An Open Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars". In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th Conference on Computational Linguistics*, pp. 8–14, Taipei, Taiwan.
- Bender, Emily M.; Dan Flickinger; Stephan Oepen & Scott Drellishak (2005) "LinGO Grammar Matrix". http://www.delph-in.net/matrix/, last visited Mar 28, 2006.
- Bond, Francis; Stephan Oepen; Melanie Siegel; Ann Copestake & Dan Flickinger (2005) "Open Source Machine Translation with DELPH-IN". In *Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit*, pp. 15–22, Phuket, Thailand.

- Catford, J. C. (1965) "Translation Shifts". In Venuti, Lawrence, editor, *The Translation Studies Reader*, pp. 141–147, London: Routledge.
- Copestake, Ann (2001) Implementing Typed Feature Structure Grammars. Stanford: CSLI Publications.
- Copestake, Ann; Dan Flickinger; Rob Malouf; Susanne Riehemann & Ivan Sag (1995) "Translation using Minimal Recursion Semantics". In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, TMI-95, Leuven, Belgium.
- Copestake, Ann; Dan Flickinger; Ivan Sag & Carl Pollard (2003) "Minimal Recursion Semantics: An Introduction". Language and Computation, vol. 1(3):pp. 1–47.
- Crysmann, Berthold (2004) "Underspecification of Intersective Modifier Attachment: Some Arguments from German". In Müller, S., editor, *Proceedings of the 11th International Conference on HPSG*, Leuven, Belgium.
- Dorr, Bonnie J. (1993) Machine Translation: A View from the Lexicon. Cambridge: MIT Press.
- Dorr, Bonnie J. (1994) "Machine Translation Divergences: A Formal Description and Proposed Solution". *Computational Linguistics*, vol. 20(4):pp. 597–633.
- Flickinger, Dan (2000) "On Building a More Efficient Grammar by Exploiting Types". Natural Language Engineering (Special Issue on Efficient Processing with HPSG), vol. 6(1):pp. 15–28.
- Flickinger, Dan & Emily M. Bender (2003) "Compositional Semantics in a Multilingual Grammar Resource". In Bender et al. (2003), pp. 18–29.
- Flickinger, Dan; Emily M. Bender & Stephan Oepen (2003) "MRS in the LinGO Grammar Matrix: A Practical User's Guide". manuscript, http://faculty.washington.edu/ebender/papers/userguide.pdf, last visited Mar 28, 2006.
- Flickinger, Dan; Jan Tore Lønning; Helge Dyvik; Stephan Oepen & Francis Bond (2005) "SEM-I Rational MT. Enriching Deep Grammars with a Semantic Interface for Scalable Machine Translation". In *Proceedings of the 10th Machine Translation Summit*, pp. 165–172, Phuket, Thailand.
- Gambäck, Björn (1997) Processing Swedish Sentences: A Unification-Based Grammar and Some Applications. Ph.D. thesis, Swedish Institute of Computer Science, SICS, Kista, Sweden.

- Ginzburg, Jonathan & Ivan A. Sag (2000) Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives. Stanford: CSLI Publications.
- Hellan, Lars (2003) "The Norwegian Computational HPSG Grammar NorSource: An Introduction, as of December 2003". manuscript.
- Hellan, Lars & Dorothee Beermann (2005) "Syntactic and Semantic Constraints on Noun Phrases for a Deep Processing Grammar of Norwegian". In *Proceedings of the 15th NODALIDA conference*, Joensuu, Finland.
- Hogan, Christopher & Robert E. Frederking (1998) "An Evaluation of the Multi-Engine MT Architecture". In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pp. 113–123, Langhorne, US: Springer.
- Hornby, A. S.; E. V. Gatenby & H. Wakefield (1963) The Advanced Learner's Dictionary of Current English. London: Oxford UP.
- Hutchins, John (2003) "Machine Translation: General Overview". In Mitkov (2003), pp. 501–511.
- Isabelle, Pierre & Laurent Bourbeau (1985) "TAUM-AVIATION: Its Technical Features and Some Experimental Results". Computational Linguistics, vol. 11(1):pp. 18–27.
- Ishikawa, Masahiko & Ryoichi Sugimura (1992) "Natural Language Analysis Using a Network Model: Modification Deciding Network". In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, pp. 55–66, Montreal, Canada.
- Jörgensen, Nils & Jan Svensson (1987) Nusvensk grammatik. Malmö: Gleerups.
- Kamp, Hans (1981) "A Theory of Truth and Semantic Representation". In Groenendijk, J.; T.M.V. Janssen & M Stokhof, editors, Formal Methods in the Study of Language, vol. 136, pp. 277–322, Amsterdam: Dordrecht.
- Kay, Martin (1996) "Machine Translation: The Disappointing Past and Present". In Cole, Ronald A.; Joseph Mariani; Hans Uszkoreit; Annie Zaenen & Victor Zue, editors, Survey of the State of the Art in Human Language Technology, http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html, last visited Mar 28, 2006.
- Levin, Beth (1993) English Verb Classes and Alternations. Chicago: University of Chicago Press.

- Lønning, Jan Tore; Stephan Oepen; Dorothe Beermann; Lars Hellan; John Carroll; Helge Dyvik; Dan Flickinger; Janne Bondi Johannesen; Paul Meurer; Torbjörn Nordgård; Victoria Rosén & Erik Velldal (2004) "LOGON. A Norwegian MT Effort". In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*, Uppsala, Sweden.
- Merkel, Magnus (1999) Understanding and Enhancing Translation by Parallel Text Processing. Ph.D. thesis, Linköping University, Linköping, Sweden.
- Mitkov, Ruslan, editor (2003) The Oxford Handbook of Computational Linguistics. Oxford: Oxford UP.
- Oepen, Stephan; Helge Dyvik; Jan Tore Lønning; Erik Velldal; Dorothee Beermann; John Carroll; Dan Flickinger; Lars Hellan; Janne Bondi Johannessen; Paul Meurer; Torbjörn Nordgård & Victoria Rosén (2004) "Som å kapp-ete med trollet? Towards MRS-based Norwegian-English Machine Translation". In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 11–20, Baltimore, US.
- Pollard, Carl & Ivan Sag (1994) Head-Driven Phrase Structure Grammar. Chicago: University of Chicago Press.
- Przepiórkowski, Adam & Anna Kupść (1996) "Unbounded Negative Concord in Polish: A Lexicalist HPSG Approach". In Landsbergen, J.; J. Odijk; K. van Deemter & G. V. van Zanten, editors, Computational Linguistics in the Netherlands 1996: Papers from the Seventh CLIN Meeting, pp. 129–143, Eindhoven, the Netherlands.
- Quirk, Randolf; Sidney Greenbaum; Geoffrey Leech & Jan Svartvik (1985) A Comprehensive Grammar of the English Language. London: Longman.
- Sag, Ivan; Thomas Wasow & Emily M. Bender (2003) Syntactic Theory: A Formal Introduction. Stanford: CSLI Publications, 2nd ed.
- Sigurd, Bengt (1995) "Analysis of Particle Verbs for Automatic Translation". Nordic Journal of Linguistics, vol. 18:pp. 55–65.
- Søgaard, Anders & Petter Haugereid (2005) "The Noun Phrase in Mainland Scandinavian". Presented at the 3rd meeting of the Scandinavian Network of Grammar Engineering and Machine Translation, Gothenburg, Sweden.
- Somers, Harold (2003) "Machine Translation: Latest Developments". In Mitkov (2003), pp. 512–528.
- Sroka, Kazimerz A. (1972) The Syntax of English Phrasal Verbs. The Hague: Mouton.

- Svartvik, Jan & Olof Sager (1977) Engelsk universitetsgrammatik. Uppsala: Esselte.
- de Swart, Henriëtte (2004) "A Typology of Negation in a Constraint-Based Framework of Syntax and Semantics". In Müller, Stefan, editor, *Proceedings of the HPSG 2004 conference*, pp. 112–118, Leuven, Belgium.
- Sågvall Hein, Anna (2005) "Datorn behöver statistik och grammatik". Språkvård, vol. 1:pp. 23–30.
- Sågvall Hein, Anna; Per Weijnitz; Eva Forsbom; Jörg Tiedemann & Ebba Gustavii (2003) "Mats - A Glass Box Machine Translation System". In *Proceedings of the Ninth Machine Translation Summit*, pp. 491–493, New Orleans, US.
- Toivonen, Ida (2002) "The Directed Motion Construction in Swedish". *Journal of Linguistics*, vol. 38(02):pp. 313–345.
- Toivonen, Ida (2003) Non-Projecting Words: A Case Study of Swedish Particles. Dordrecht: Kluwer.
- Tseng, Jesse (2000) The Representation and Selection of Prepositions. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.
- Tseng, Jesse (2003) "LKB Grammar Development: French and Beyond". In Bender et al. (2003), pp. 91–97.
- Uszkoreit, Hans; Dan Flickinger; Walter Kasper & Ivan Sag (2000) "Deep Linguistic Processing with HPSG". In Wahlster, Wolfgang, editor, Verbmobil: Foundations of Speech-to-Speech Translation, pp. 216–237, Berlin: Springer.
- Villavicencio, Aline & Ann Copestake (2002) "Phrasal Verbs and the LinGO-ERG". LinGO Working Paper No. 2002-01.

# A Verb frames in English and Swedish

This appendix contains verb frame inventories of English in Table A.1 and Swedish in Table A.2.

Table A.1: Verb frames in English

Type	Example
Impersonals	
Regular	It rains
Intransitive	
Regular	Snore
Refl.	Cut oneself
Part.	Fall down
Refl. Part	Buy oneself free
PP-compl	Talk to someone
Refl. PP-compl	Find oneself in trouble
Part. PP-compl	Close in on someone
Obl. adv.	Live here
Pred.	Seem honest
Transitive	
Regular	Like someone
Part + obj	Throw something out
Obj + part	Throw out something
Obj + pred.	Get something dirty
Obj + Noun-part	Elect someone king
Obj + obl.adv	Put something down
Obj + past. part.	Get something cut
Ditransitive	
Simple	Give someone something
Prepositional	Give something to someone
Dative+ prep. acc.	Thank someone for something

Table A.1: cont.

Type	Example
Propositional	-
That-prop.	think that it works
WH-prop	Wonder why it works
Part + WH-prop	Find out why it works
WH-prop inf.	Learn what to say
To-inf.	Continue to work
Pres. part.	Stop talking
Part + pres. part	Keep on coming
Bare inf.	Can work
Trans. WH-prop	Tell someone what they did
Trans. regular	Tell someone that something happened
Trans + to-inf	Want someone to do something
Trans + bare inf.	Watch someone work
Trans + pres.part.	Dislike someone singing
Trans + past part.	Keep someone posted
Auxilliaries	
Tense	Have done something
Modal	Can do something
Copulas	
past part.	Is sold
NP	Is a car
Adj	Is red
PP	Is on the road
Adverb	Is outside
Poss. pronoun	Is mine

Table A.2: Verb frames in Swedish

Type	Example
Impersonals	
Regular	Det regnar
Reflexive	Det ordnar sig
Intransitive	
Regular	Snarka
Refl.	Harkla sig
Part.	Falla ned
Part. Refl.	Vika ut sig
Refl. Part	Göra sig till
PP-compl	Lita på någon
Refl. PP-compl	Gifta sig med någn
Part. PP-compl	Råka ut för något
Part.refl. PP-compl	Klä ut sig till något
Pred.	Verka trevlig
Pred. + part.	Se förskräcklig ut
Obl. adv.	Komma hit
Transitive	
Regular	Gilla någon
Refl	Närma sig någon
Part	Tycka om någon
Refl. + part.	Slå sig på något
Obj. + pred	Göra någon sjuk
Obl. adv. + Obj	Lägga ner något
Ditransitive	
Simple	Ge någon något
Prepositional	Ge något till någon
Dative+ prep. acc.	Tacka någon för något
Part. prepositional	Hyra ut något åt någon
Part + refl, prepositional	Ta med sig något till någon
Propositional	
Regular	Försöka göra något
That-prop.	Hoppas att något hänt
To-inf.	Fortsätta att göra något
Refl. to-inf	Förvänta sig att något händer
Part. to-inf	Tycka om att göra ngt
Refl part to-inf	Bestämma sig för att göra ngt

Table A.2: cont.

Type	Example
WH-prop	Undra vad någon gjort
Part WH-prop	Komma ihåg vad någon gjort
Refl WH-prop	Lära sig hur ngt fungerar
Refl. Part. WH-prop	Förhöra sig om vad någon gjort
Trans regular	Låta någon göra något
Trans. WH-prop	Visa ngn vad ngn ska göra
Trans prep. WH-prop	Upplysa någon om vad han gjort
Trans. Part WH-prop	Säga åt ngn vad ngn ska göra
Trans. to-inf	Tvinga ngn att göra ngt
Trans. prep. to-inf	Berätta för någon att något hänt
Trans. Part to-inf	Säga åt ngn att sluta
Auxilliaries	
Tense	Har gjort något
Modal	Kan göra något
Copulas	
VP, past participle	Är såld
NP	Är en bil
Adj	Är röd
PP	Är på vägen
Adverb	Är ute
Poss. pronoun	Är min

# B Sample sentences and the translations proposed by BiTSE

Table B.1 contains a sample of sentences accepted by BiTSE, and all the sentences that were generated from each sample sentence.

Table B.1: Sample sentences and the translations of them proposed by BiTSE. In the cases where BiTSE gives more than one analysis of a sentence, each analysis have a table entry marked with a number.

Input sentence	Generated sentences
I shave	Jag rakar mig
	I shave myself
	I shave
I greet him	Jag hälsar på honom
	I greet him
I visit Ann	Jag hälsar på Ann
	Jag besöker Ann
	I visit Ann
Jag hälsar på Bob (1)	Jag hälsar på Bob
	I greet Bob
Jag hälsar på Bob (2)	Jag hälsar på Bob
	Jag besöker Bob
	I visit Bob
Jag kommer mig upp	Jag kommer mig upp
	I succeed
Jag kom ihåg hunden	Jag kom ihåg hunden
	I remembered the dog
I insist	Jag står på mig
	I insist
Jag jagade stora lejon	Jag jagade stora lejon
	I hunted big lions
Do you sleep	Sover du
	Do you sleep

Table B.1: cont.

Input sentence	Generated sentences
Can you sleep	Kan du sova
	Can you sleep
Jag vet att han litar på mig	Jag vet att han litar på mig
	I know that he trusts me
Jag levererar en hund åt Kim	Jag levererar en hund åt Kim
	Jag levererar en hund till Kim
	I deliver a dog to Kim
I promised a dog to Bob	Jag lovade Bob en hund
	I promised a dog to Bob
	I promised a dog to Bob
Kim undrar om jag sover	Kim undrar om jag sover
	Kim wonders whether I sleep
Jag klär på mig	Jag klär på mig
	I dress
	I dress myself
Jag betrakatar hunden	Jag betraktar hunden
	I look at the dog
Jag hämnades på honom	Jag hämnades på honom
	I revenged myself on him
Jag berättar en historia för honom	Jag berättar en historia för honom
	I tell him a story
	I tell a story to him

# THE WAS UNIVERSITY LAND

## Avdelning, Institution

Division, Department

Datum Date

2006-06-07

## Linköpings universitet

Institutionen för datavetenskap 581 83 Linköping

ISRN  LIU-KOGVET-D06/08S  Serietitel och serienummer  Title of series, numbering	LIU-KOGVET-D06/08SE Serietitel och serienummer ISSN
	LIU-KOGVET-D06/08S Serietitel och serienummer

Titel
Title

 ${\bf Swedish\text{-}English\ Verb\ Frame\ Divergences\ in\ a\ Bilingual\ Head\text{-}driven\ Phrase}$ 

Structure Grammar for Machine Translation

Skillnader i verbramar mellan svenska och engelska i en tvåspråkig HPSG-

grammatik för maskinöversättning

#### Författare

Author

Sara Stymne

#### Sammanfattning

Abstract

In this thesis I have investigated verb frame divergences in a bilingual Head-driven Phrase Structure Grammar for machine translation. The purpose was threefold: (1) to describe and classify verb frame divergences (VFDs) between Swedish and English, (2) to practically implement a bilingual grammar that covered many of the identified VFDs and (3) to find out what cases of VFDs could be solved and implemented using a common semantic representation, or interlingua, for Swedish and English.

The implemented grammar, BiTSE, is a Head-driven Phrase Structure Grammar based on the LinGO Grammar Matrix, a language independent grammar base. BiTSE is a bilingual grammar containing both Swedish and English. The semantic representation used is Minimal Recursion Semantics (MRS). It is language independent, so generating from it gives all equivalent sentences in both Swedish and English. Both the core of the languages and a subset of the identified VFDs are successfully implemented in BiTSE. For other VFDs tentative solutions are discussed.

MRS have previously been proposed as suitable for semantic transfer machine translation. I have shown that VFDs can naturally be handled by an interlingual design in many cases, minimizing the need of transfer.

The main contributions of this thesis are: an inventory of English and Swedish verb frames and verb frame divergences; the bilingual grammar BiTSE and showing that it is possible in many cases to use MRS as an interlingua in machine translation.

**Nyckelord** Keywords verb frame divergences, machine translation, grammar engineering, Head-driven Phrase Structure Grammar, translation divergences, Minimal Recursion Semantics, Grammar Matrix, typed feature structures