Univerzita Karlova v Praze
Filozofická fakulta
Ústav germánských studií

Silvie Cinková

WORDS THAT MATTER

Towards a Swedish-Czech Colligational Lexicon of Basic Verbs

DŮLEŽITÁ SLOVA

Podklady ke kolokačnímu švédsko-českému slovníku základních sloves

Disertační práce

Program: Filologie
Obor: Germánské jazyky a literatury

Vedoucí práce: Doc. RNDr. Vladimír Petkevič, CSc.

Praha 2009

Prohlašuji, že jsem disertační práci vykonala samostatně s využitím uvedených pramenů a literatury.

Silvie Cinková

Contents

1 Introduction **15**
1.1 The Importance of Collocations 15
1.2 Motivation 16
1.3 Objectives 17

I Theoretical Background **19**

2 Key terms **21**
2.1 Words, Lexemes, and Lexical Items 21
2.2 Collocation 21

3 Grammaticalization **25**
3.1 The Notion of Grammaticalization 25
3.2 Emergent Grammar 27
3.3 The Most Grammaticalized Verb Categories 29
3.3.1 Voice and Tense 29
3.3.2 Mood 29
3.3.3 Number and Person 30
3.3.4 Valency 30
3.3.5 Aspect 30
3.4 Discovering Regularities in Verb Usage 32
3.5 Semantic Changes in Grammaticalizing Verbs 32
3.5.1 Metaphorical Extension from one Semantic Domain to Another 33
3.5.2 Context-induced Reinterpretation 34

4 Light Verb Constructions **37**
4.1 The Notion of Light Verb Construction (LVC) 37
4.2 LVCs as Collocations 40
4.3 Semantic Aspects of LVCs 41

4.4 LVCs and Event Structure 41
4.5 Productivity vs. Lexicalization in LVCs 42
4.6 Aspects of Valency in LVCs 43
4.7 Communicative Aspects of LVCs 48
4.8 Conclusions 50

5 The Transitivity Hypothesis 51
5.1 What is Transitivity and (why) does it Matter? 51
5.2 Grammatical Interference in Lexicalized Collocations? 55
5.3 Transitivity Indicators 60
5.3.1 Participants 61
5.3.2 Kinesis 61
5.3.3 Aspect 61
5.3.4 Punctuality 62
5.3.5 Volitionality 62
5.3.6 Affirmation 62
5.3.7 Mode 62
5.3.8 Agency 62
5.3.9 Affectedness of the Patient 62
5.3.10 Individuation of the Patient 63
5.4 **T**ransitivity as a Discourse Marker 63
5.5 Morphosyntactic Consequences of Transitivity Hypothesis 64
5.6 Relation between Aspect and Definiteness 66
5.7 Transitivity Hypothesis in Light Verb Constructions 67

II Methods and Approaches 69

6 Valency Theory in Functional Generative Description 71
6.1 Functional Generative Description 71
6.2 Tectogrammatical Representation 71
6.3 Valency 75
6.3.1 Basic Notions 75
6.3.2 Inner Participants and Free Modifications 76
6.3.3 Obligatoriness and Dialog Test 79
6.3.4 Shifting 80
6.3.5 Quasi-Valency Complements 82
6.3.6 Discussion 83
6.3.7 Noun Valency in FGD 86
6.3.8 FGD-valency for Learners of Basic Verbs 87

7 Corpus Pattern Analysis **89**
7.1 Theory of Norms and Exploitations (TNE) 89
7.2 Defining Lexical Sets 90
7.3 Applying CPA 92

8 Lexical Functions **93**
8.1 Brief Overview 93
8.2 Lexical Functions and FGD 93
8.3 Basic Lexical Functions in LVC Description 94
8.3.1 Oper_i 95
8.3.2 $\text{Labor}_{i,j}$ 96
8.3.3 Func_i 96
8.3.4 Copul 97
8.3.5 Cross-linguistic Comparison of the Predicate Noun *disposal* in Swedish vs. in Czech 97
8.4 Phasal Lexical Functions 98
8.5 Causative Lexical Functions 99

9 Examples of Valency Lexicons and Collocational Lexicons **101**
9.1 BBI 101
9.2 Combinatorial Explanatory Dictionaries 104
9.2.1 Definitions in CEDs 105
9.2.2 Government Pattern 105
9.2.3 Lexical Combinatorics 106
9.2.4 Monolingual and Multilingual CEDs 106
9.3 Dictionary of Czech Phraseology and Idiomatics 107
9.4 PropBank, NomBank 109
9.5 VALLEX, PDT-VALLEX 110
9.5.1 Differences between VALLEX and PDT-VALLEX 110
9.5.2 Structure of VALLEX 111
9.5.3 Other Language Versions 113
9.6 FrameNet 113
9.7 Svenskt Språkbruk 114
9.8 PAROLE-SIMPLE 115
9.9 VerbaLex 115

III Grammar and Corpus – the Case of Swedish **117**

10 The Role of Verbs in the Vocabulary **119**
10.1 Verb Statistics 119

10.2 The 20 Most Frequent Verbs in Swedish 120

11 *Komma att* **125**
11.1 Future 125
11.2 Accident, Coincidence 129
11.3 A Contrastive Corpus Analysis of the Non-Future Uses 131
11.3.1 Significant Collocates 131
11.3.2 Coincidence in Czech 131
11.3.3 *Kom att* in Result Clauses 138
11.3.4 Czech Equivalents of *komma att tänka* 142

12 *Hålla på …* **147**
12.1 Valency Patterns 147
12.2 *X håller på med Y* 147
12.3 *X håller på (med) att verb-a* 148
12.4 Progressivity 149
12.5 Progressivity Hides Constancy 150
12.6 Tendentiality 153

13 Pseudo-coordinations with *ligga, sitta, stå* **155**
13.1 The Profile of *X står och verb-ar* in PAROLE 157
13.2 The Profile of *X sitter och verb-ar* in PAROLE 162
13.3 The Profile of *X ligger och verb-ar* in PAROLE 163
13.4 Deictic Markers in Pseudo-coordinations 165

14 Pseudo-coordinations with *ta* **169**
14.1 The Profile of *X tar och verb-ar* in PAROLE 170
14.2 Czech Equivalents 171

IV Implementation of Swe-VALLEX/PNL **173**

15 Preparatory Work **175**
15.1 Organizing Lexical Sources and Tools 175
15.2 Preparing the Corpora 175
15.3 Word Sketch Engine and Collocation Analysis 176
15.4 Adjusting the Word Sketch Engine for PAROLE 177
15.4.1 Lemmatization 177
15.4.2 Lemmatization Results Related to Generating Word Sketches 183
15.4.3 Computing Grammatical Relations 183
15.4.4 Empirical Evaluation of Word Sketches 185
15.5 Determining Entry Candidates 187

16 Data Structure **193**
16.1 Usefulness of Word Sketches 193
16.2 Main Principles and Features 193
16.3 SweVallex 196
16.3.1 Macrostructure 196
16.3.2 Lemma 197
16.3.3 Patterns 198
16.3.4 Slot 199
16.3.5 Surface Form 199
16.3.6 FGD-Information 200
16.3.7 CPA-Information 201
16.4 Predicate Noun Lexicon 201
16.4.1 Macrostructure 201
16.4.2 Predicate Noun Lemma 201
16.4.3 Light Verb Unit 202
16.4.4 Noun Definiteness, Modifier Insertion 203
16.4.5 Slot 204
16.5 Linking 206

17 Discussion **211**
17.1 Increasing Recall with Targeted Corpus Queries 211
17.2 Frequency Counts 214
17.3 Irrealis, Negation, and Semantic Definiteness 214
17.4 Undetected Information 215
17.5 Different GUI – New Corpus Annotation 216
17.6 Parallel Data 217

18 Conclusion **223**

Bibliography **225**

V Appendices **243**

19 Data sample **245**

20 A Sample of the Lemmatizer Script **259**

21 Targeted Corpus Query Templates for PAROLE **263**

Abstract

Basic verbs, i.e. very common verbs that typically denote physical movements, locations, states or actions, undergo various semantic shifts and acquire different secondary uses. In extreme cases, the distribution of secondary uses grows so general that they are regarded as auxiliary verbs (*go* and *to be going to*), phase verbs *(turn, grow)*, etc. These uses are usually well-documented by grammars and language textbooks, and so are idiomatic expressions (phraseologisms) in dictionaries.

There is, however, a grey area in between, which is extremely difficult to learn for non-native speakers. This consists of secondary uses with limited collocability, in particular light verb constructions, and secondary meanings that only get activated under particular morphosyntactic conditions. The basic-verb secondary uses and constructions are usually semantically transparent, such that they do not pose understanding problems, but they are generally unpredictable and language-specific, such that they easily become an issue in non-native text production.

In this thesis, Swedish basic verbs are approached from the contrastive point of view of an advanced Czech learner of Swedish. A selection of Swedish constructions with basic verbs is explored. The observations result in a proposal for the structure of a machine-readable Swedish-Czech lexicon, which focuses on basic verbs and their constructions. The lexicon is anchored in the valency theory of the Functional Generative description, coupled with analysis of collocations according to the semantically motivated principles of Corpus Pattern Analysis, in order to achieve the necessary level of delicacy to make meaning distinctions correctly.

The lexicon consists of two parts: SweVallex, which is a lexicon of verb frames, and a Predicate Noun Lexicon, which captures predicate nouns (the nominal components of light verb constructions). These two parts are interlinked. The verb collocates of predicate nouns are sorted according to the Mel'čukian Lexical Functions. Features such as telicity, punctuality, and volitionality are described for each light verb construction, whenever possible. Special attention is paid to the morphosyntactic behavior of the respective predicate nouns (determiner use, and modifier insertion).

In order to facilitate the routine of building such a lexicon, the 20-million morphosyntactically annotated Swedish corpus PAROLE was lemmatized and loaded into the corpus GUI Bonito, which includes the Word Sketch Engine, a tool for automatic collocation analysis. Word Sketch Definitions for Swedish were created and loaded into the Word Sketch Engine. In addition to the PAROLE corpus, a two-million parallel Swedish-Czech corpus was used, which has been built within a different project.

Abstrakt

Základní slovesa (basic verbs), tj. frekventovaná významová slovesa, jež zpravidla popisují fyzický pohyb, umístění, stav, nebo děj, procházejí řadou sémantických posunů, díky kterým se používají k vyjádření druhotných, přenesených významů. V krajních případech se dané sloveso stává pomocným, způsobovým, nebo fázovým slovesem a přestávají pro ně platit kolokační omezení, jež se vztahují na sloveso užité v jeho primárním (tj. doslovném) významu. Tato užití sloves bývají většinou dobře dokumentována v gramatikách i učebnicích, stejně jako kvalitní slovníky podávají podrobnou informaci o užití těchto sloves v ustálených frazeologických spojeních.

Mezi plně gramatikalizovaným užitím na jedné straně a idiomatickým, frazeologickým užitím na druhé straně však existuje celá škála užití základních sloves v přenesených významech, jejíž zvládnutí je pro nerodilého mluvčího značně obtížné: užití v přeneseném významu, jež mají omezenou kolokabilitu. To jsou především verbonominální konstrukce někdy nazývané analytické predikáty (light verb constructions), ale také užití, která za určitých omezených morfosyntaktických podmínek (např. pouze v negaci) aktivují abstraktní sémantické rysy u jiných predikátů, např. zesilují význam, nebo implikují, že daný děj již trvá dlouho, a podobně. Tato druhotná užití významových sloves (ve švédštině) většinou nepůsobí (českým pokročilým studentům švédštiny) potíže při porozumění textu, neboť bývají sémanticky transparentní, avšak tím, že jsou specifická pro konkrétní jazyk a v zásadě neprediktabilní na základě znalosti jiného jazyka, působí problémy při produkci textu. Rodilí mluvčí sami je většinou nevnímají nebo jejich typické kontexty považují za frazeologismy, a proto jim ani při výuce cizích studentů nevěnují dostatečnou pozornost.

Předkládaná práce se zaměřuje na švédská základní slovesa z kontrastivního pohledu českého pokročilého studenta švédštiny. Pozorování vybraných slovesných konstrukcí zobecňuje do návrhu struktury elektronického švédsko-českého slovníku zaměřeného na pochopení a osvojení méně zřejmých, ale přesto frekventovaných konstrukcí se základními slovesy. Slovník je zakotven ve valenční teorii Funkčního Generativního Popisu, spojené s kolokační analýzou podle sémanticky motivovaných principů Analýzy Korpusových Vzorců (Corpus Pattern Analysis), jež umožňuje přesnější definici jednotlivých slovesných užití pomocí jejich typických kolokátů.

Slovník sestává ze dvou vzájemně propojených částí: slovníku sloves SweVallex (vytvořeného na základě českého Vallexu) a slovníku nominálních částí analytických predikátů (Predicate Noun Lexicon). Slovesné kolokáty jednotlivých nominálních komponent jsou roztříděny podle Mel´čukových Lexikálních Funkcí. Kromě toho je u každého analytického predikátu, pokud je to možné, uvedena informace o jeho telicitě, o tom, zda děj je okamžitý, nebo spíše durativní (punctuality), a o tom, jestli jeho subjekt za normálních okolností jedná vědomě a ze své vůle (volitionality). Zvláštní pozornost je věnována morfosyntaktickému chování nominálních komponent ve spojení s jednotlivými slovesnými kolokáty (užití členů, možnosti rozvití).

Studie opouští slovník ve chvíli, kdy je jeho struktura navržena a prověřena na příkladech. Pro umožnění kvalitní rutinní práce byly připraveny lingvistické zdroje. Veřejně dostupný švédský korpus PAROLE byl lemmatizován pomocí zvlášť vyvinutého pravidlového lemmatizátoru a vložen do korpusového rozhraní Bonito, v němž je zabudován nástroj na automatickou kolokační analýzu – Word Sketch Engine. Tento nástroj bylo zapotřebí adaptovat na švédštinu, což bylo také předmětem této dizertační práce. V rámci jiného projektu vzniká paralelní švédsko-český korpus, který má v současné době přibližně 2 miliony tokenů. Ten byl v nejvyšší možné míře využit v kontrastivním pozorování vybraných slovesných konstrukcí. Doplňkově byl využíván také Český národní korpus.

Acknowledgements

Any merit that this work may have is dedicated to Karel Oliva (1927–2005), a great Czech linguist and the author of the largest Polish-Czech dictionary. He was a dear friend and mentor to many young people, myself included.

I would also like to express my gratitude to:

Anna Braasch, Kristín Bjarnadóttir, Ulrich Heid, Ann Lindvall, Markéta Lopatková, and Zdeněk Starý, who let me acquire up-to-date publications and who spent hours of their valuable time to share their knowledge with me,

Petr Pajas, Pavel Rychlý, and Zdeněk Žabokrtský for advising me on all technical matters and granting me access to relevant tools,

Jan Pomikálek, without whom I would never have been able to make my lemmatising rules comprehensible to any computer,

Pavel Straňák and Ondřej Bojar, who helped me with typographical issues,

Jan Hajič and Eva Hajičová, who gave me the opportunity and the joy of inspiring work at UFAL,

František Čermák and Patrick Hanks for reading my thesis and adding valuable comments at various stages,

Vladimír Petkevič for becoming my supervisor, for leading and encouraging me through all the years,

Petra Tesařová, Michal Zikán, and Laura Janáčková for putting me back in the game

…and especially to my family, who have always kept me going and whom I owe so much.

This work was funded in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434, by the Czech Science Foundation (GA-405/06/0589), and by the Czech Ministry of Education (0021620823).

Chapter 1

Introduction

Structure, or regularity, comes out of discourse and is shaped by discourse as much as it shapes discourse in an on-going process.

Paul Hopper: Emergent Grammar

1.1 The Importance of Collocations

Every human language has a grey area where grammar and lexicon overlap. It resists attempts at systematic description in grammar textbooks as well as in lexicons for foreigners; yet understanding the nature of this grey area is crucial, not only for effective language learning, but also for many other applications. It is necessary to seek language usage that is not only correct but also idiomatic. To do this, it is necessary to understand collocations.

In language production, collocational preferences in semantically transparent constructions rather than idioms pose problems for foreign speakers, [149]. As Sinclair notes, foreign speakers quite often abuse dictionaries to look up and combine rare and outdated words to express ordinary, everyday concepts instead of using common collocational blocks familiar to native speakers. English phrasal verbs provide nice examples of this issue: a foreign learner often gets lost in the jungle of these utterly unintelligible combinations of highly polysemous verbs with one or two preposition-like attachments. Therefore, a foreign learner often prefers e.g. *vomit* to *throw up* and *reconcile* to *make up*, although a native speaker would only use *vomit* or *reconcile* in formal contexts.

Textbooks and dictionaries for foreigners have been randomly treating diffuse uses of common lexemes either as grammar issues in text-books, or as 'idioms' and 'figurative senses' in lexicons, while many regularities have simply remained unnoticed by native speakers. Corpus linguistics has shown through analysis of vast amounts of data that the way in which native speakers believe words behave is often different

from the way in which words actually behave. Manual excerpts of lexical evidence for dictionaries and grammars tend to highlight the unusual in the language, while missing important patterns (see [57]).

The motto of the following work is *grammaticalization*, understood as "movement towards structure" [66]. As Hopper puts it: "…the more useful a construction is, the more it will tend to become structuralized, in the sense of achieving cross-textual consistency, and serving as a basis for variation and extension." For the purpose of this study, 'grammaticalization' is understood in its broadest sense. It is basically used interchangeably with 'lexicalization'. Structures that can combine with virtually any collocates (very generalized structures) like the English *come to + INF* (*come to know* etc.) will rather be called 'grammaticalized', while structures with stricter collocational restrictions will rather be called 'lexicalized'. Nevertheless, this thesis makes no attempt to draw a line between 'grammaticalization' and 'lexicalization'[1].

1.2 Motivation

The initial impulse for this work came from a previous experience with lexicographical processing of German lexical verbs. German, being famous for its abundant nominalizations, makes prominent use of *Funktionsverbgefüge* (light verb constructions), a way of integrating a nominalization into a sentence by means of a semantically-depleted lexical verb, such as *zur Verfügung stellen*, *eine Frage stellen* etc. The ultimate German grammar for foreign learners [63] even introduces a group of *Funktionsverben* (light verbs).[2] Examining Swedish light verbs seemed to be an interesting task. However, currently available Swedish grammars [152], [112] do not provide any explicit list of light verbs, and light verb constructions are only captured by the lexicon [3], sometimes under noun, sometimes under verb lemmas. The initial attempts to identify a generally accepted list of Swedish light verbs resulted in the insight that verbs, rather then *being* light verbs, *become* light verbs by being used in collocation with a predication-containing noun. Many lexical verbs belonging to the basic vocabulary have a shifting potential for occurring in collocation with predication-containing nouns. Some enter such constructions frequently and productively, while others only occur in one or few lexicalized cases, such as *bjuda* in *bjuda motstånd*. Only few lexical verbs occur almost exclusively in light verb constructions (i.e. *genomföra*). Most verbs with this ability have quite concrete meanings, i.e. *komma, ställa, hålla*, or *få*. Hence the field of inquiry expanded to semantic and syntactic changes in such verbs when

[1]Cf. Section 3.2.

[2]The English term *light verb* used in the present thesis was coined by Otto Jespersen [74]. A more accurate German equivalent for this term is **verblasste Verben**. Hanks et al. [60] argue that the semantic criterion of 'lightness' is preferable, as an identifier for these interesting verbs, to the syntactic criterion implied by *Funktionsverbgefüge*.

they occur in combinations that reach beyond their most concrete meaning.

1.3 Objectives

Originally, this thesis sought only to explore ways of creating a systematic description of light verb constructions in a lexicon. However, the usage of a verb as a light verb is quite often only one instance of its grammaticalized uses. It is typically the basic spatial verbs and verbs denoting physical actions that are the most productive light verbs, and they typically exhibit a significant semantic complexity.

With the human lexicon user in mind, a thorough description of all types of grammaticalized uses with some common verbs seemed more sensible than focusing strictly on light-verb-like uses and defining Swedish light verbs as a group.

An experimental lexicon has been created, called **Swe-VALLEX/PNL**. The first part of the name refers to VALLEX, valency lexicon of Czech verbs [90] based on the valency theory of the Functional Generative Description (see Chapter 6). Swe-VALLEX is meant to be a Swedish counterpart of VALLEX. The abbreviation *PNL* stands for *Predicate Noun Lexicon*. The lexicon contains two lemma types: verb lemmas and noun lemmas. The noun entries present significant collocates of the verbs included, primarily predicate nouns in light verb constructions. Their types will be specified in more detail (see Section 16.4). The respective entry types are stored in two separate XML documents (Swe-VALLEX and PNL), which are interlinked.

To become eligible for lemmatizing in Swe-VALLEX/PNL, a verb must have been found in a large corpus to participate in at least one light verb construction. The selection of lemma candidates is described in Section 15.5.

These are the main features of the proposed lexicon:

- The lexicon is theory-bound and formalized in a way that makes it eligible for NLP tasks, e.g. as a valency lexicon for an FGD-based Swedish treebank.

- When enhanced with a browser, the lexicon is a user-friendly electronic lexical resource for human users.

- The description focuses on grammaticalizing verb-noun constructions. To increase the informative value of the lexicon, neither 'literal senses' nor idioms have been ignored, but they might be incomplete. This applies especially to idioms, evaluative expressions and proverbs. For a human user, the lexicon is by no means a substitute for large monolingual dictionaries, e.g. Svenskt språkbruk and Norstedts stora svenska ordbok, but it seeks to complement them by stressing the morphosyntactic information necessary for language production.

- The source language is Swedish, and the target language is Czech. Emphasis has been laid on the Swedish part. The Swedish part contains information on

spelling, valency and morphosyntactic features, as well as numerous authentic examples in order to provide all knowledge relevant for improving the user's foreign language production. The Czech part comprises only equivalents. No further description of the Czech equivalents is provided.

- The data structure makes it possible to view the entries sorted according to the Czech equivalents. It can thus be used as a – very simplified – Czech-Swedish lexicon of verbs.

- The lexicon is corpus-driven in the sense of Hanks [57], but the results of the corpus-driven analysis are compared to large Swedish monolingual dictionaries (cf. Section 15.4.4).

One more word should be said about the decision to make the lexicon both human- and machine-oriented: the examples of the Czech VALLEX, PropBank and many other resources show that a user-friendly interface makes even a very formalized data structure easily understood. Information overload with human users is not to be feared. The final layout of any lexicon can be adapted to any target user's needs, provided the relevant information is present and organized consistently enough to be retrieved – which has been the aim of this project from the very beginning, in full agreement with Bolshakov, Gelbukh and Haro [12]: "…the information should be accessible to both human users and other programs. A dictionary is so large piece of data, and its development is so expensive, that it is unacceptable to maintain, keep and use separate versions of dictionaries for humans and for the machine".

Part I

Theoretical Background

Chapter 2

Key terms

2.1 Words, Lexemes, and Lexical Items

This Chapter is dedicated to the keywords of the present study, which is all about words and the relations that exist among them. *Word* itself is a highly polysemous expression, and for the purpose of this thesis it does not appear useful agonizing too long over what a word is. Nevertheless, this text never operates with *word* in the sense of *a single token*, which otherwise could become a main source of misinterpretation. The number of tokens a 'word' comprises is a non-issue here. *Word* is understood as a distinct meaning unit constituted by a combination of one or more word stems and the appropriate morphosyntactic characteristics – like *lexeme* in the Czech linguistic tradition, or *lexical item* in the English linguistic discourse.

Therefore, *word* is used interchangeably with *lexeme*, which is the proper term in the Czech linguistic tradition, or *lexical item*. Mostly *lexeme* is being used, but *word* is preferred in passages that refer to Hanks.

2.2 Collocation

The most important keyword of this study is *collocation*. The degree of grammaticalization is estimated with respect to which *collocates* a target item combines with. Light Verb Constructions have been regarded as *collocations*.

The conception of *collocation* differs with each single author, being located somewhere on a scale between significant statistical co-occurrence of two or more words on one extreme and semantic non-compositionality on the other extreme. Evert and Krenn [40] give a summary of the two predominant views of collocations:

1. Collocations as recurrent combinations of words. This approach is concerned with co-occurrences that express semantic and conceptual relations, ignoring the syntactic relations between the collocates.

2. Collocations as pre-constructed syntactic units or lexically determined elements in syntactic constructions. Collocations are characterized by their semantic, syntactic, and distributional irregularity rather than by their distribution in corpora.

Some authors (mainly Firth [45] and Hoey [65]) even draw a distinction between *collocations* and *colligations*. The term *collocations* then focuses the semantic relations between lexical items, while *colligations* focuses on the interplay between lexical items and grammatical categories. Many of the phenomena discussed later in this study could be regarded as colligations, especially the morphosyntactic variations observed in Light Verb Constructions. Also Benson, Benson and Ilson [108] distinguish two types of collocations: grammatical collocations and lexical collocations. While lexical collocations are formed by content words, grammatical collocations comprise a content word and a function word governed by it, e.g. a verb and a preposition. *Grammatical collocation* as used by the authors of BBI corresponds to surface valency in this work.

[39] give a simple working definition of *collocation*: "Collocations are understood as **unpredictable combinations of words in a particular (morpho-)syntactic relation** (adjectives modifying nouns, direct objects of verbs, or English noun-noun compounds)." The 'colligation-like' flavor of this definition was also adopted for this study, while 'unpredictability' is regarded as a very weak criterion as it decreases, the less plausible it is for a verb to act as a non-light Verb (e.g. *carry out*). In accordance with the observations of Hanks et al. [60], however, the determination and description of collocations/colligations appears to rely on semantic as well as syntactic criteria, the significance of which, respectively, varies in individual cases.

Sinclair [149] introduces the very useful terms *node* and *collocate*. By *node* he understands the lexeme that has been sought in a corpus or otherwise focused. *Collocate* is the lexeme that the node collocates with. Sinclair further defines two possible relations the node and the collocate enter into, depending on their respective frequencies in the given corpus. *Downward collocation* is a collocation whose node is more frequent than the collocate, while *upward collocation* is a collocation whose node is less frequent than the collocate. Observing downward collocation is useful for exploring a lexeme's semantics while observing upward collocation is useful when exploring the syntactic patterns the node enters into.

The most practical delimitation of *collocation* is found in Čermák [23], in Czech. *Collocation* is regarded as a basic term, which is to be further specified if necessary. Čermák applies the basic criteria of **stableness** (part of the language system vs. part of text) and (semantic and/or syntactic) **regularity**. In terms of [23], this study uses *collocation* for the following collocation types[1]:

[1]The original Czech labels in brackets are complemented with the alphanumeric code introduced in Čermák's text.

- *terminological collocations – multiword terms* (termínové kolokace – víceslovné termíny: A1a)
- *proper noun multiword units* (propriální kolokace – víceslovná propria: A1b)
- *idiomatic collocations – idioms and phrasemes* (idiomatické kolokace – idiomy a frazémy: A2)
- *common collocations – gram-semantic combinations* (běžné kolokace – gram-sémantické kombinace: B3a)
- *common usage collocations* (běžné kolokace uzuální: C)

Čermák also considers only combinations of content words, though he does not raise any objections to extending the term to combinations of function words governed by the content words as valency complementations.

Chapter 3

Grammaticalization

3.1 The Notion of Grammaticalization

This thesis focuses on grammaticalized uses of *basic verbs*. *Basic verb* is not an established term. It is merely a label used for describing lexical verbs that are very common in everyday communication, typically spatial verbs (*to sit*), verbs of motion (*to go*) or verbs of physical contact/control (*to hold, to keep*). They are themselves stylistically neutral, though they naturally might change their stylistic value when employed in context. They belong to what Heine, Claudi and Hünnemeyer ([8], p. 33), call *basic vocabulary*, which they define as "lexemes that are less subject to replacement than others". e.g. the verb *to march* is less eligible to be classified as a basic vocabulary verb than *to go*, though they are to a significant extent synonyms. Basic vocabulary lexemes denote what Lakoff [82] calls *basic level categories* – simply entities, events and their relations identified and classified with just the level of granularity that is first and intuitively best perceived when acquiring common knowledge.

Basic verbs apparently denote basic level events (processes, transitions and states)…do they? This assumption fits well for sentences like

(1) *She was lying on the bed watching television.*

(2) *Where did Sue go?*

etc., but it appears odd with sentences like

(3) *As we go into the third round…*

(4) *Evans is lying in third place.*

or even

(5) *Everything's going to be all right*,

where *go* is very basic, but as a function word, not a content word.[1]

[1]The example sentences come from the Macmillan English Dictionary for Advanced Learners [38].

Sentences 1 and 2 are the most *cognitively salient* uses of the verbs *to lie* and *to go*, while 3, 4, and 5 are less *cognitively salient*, but they are perhaps equally or more *socially salient* uses of the example verbs. *Cognitive salience* and *social salience* are terms coined by Hanks [57]. *Cognitively salient* uses of lexemes are uses which speakers believe to be the most typical, 'normal' ones ("what we think words mean" [57]) whereas *socially salient* uses are uses of the same lexemes that were found in large corpora to be actually the most frequent ones ("the actual meanings that we use" [57]). Hanks draws the conclusion that *cognitive salience* and *social salience* are "independent variables or possibly even in an inverse relationship" as textual corpora show that the 'less typical' uses of the given lexemes are far more frequent than the *cognitively salient* ones.

Sinclair [149] speaks about *phrase-dependent* uses of lexemes. The most socially salient uses of basic lexemes are almost regularly the *phrase-dependent* ones. The phrase-independent uses (or *core meanings*, cf. [56] and *cognitively salient* uses [57]) can be described as the uses most tightly associated by most speakers. Core meanings are supposed to "have wider ranges of normal phraseology than derivative, pragmatic, metaphoric, and idiomatic senses" ([56], p. 58); i.e. the collocational potential of the core meanings is supposed to be higher than that of the other uses. If we compare the sentences 1 and 2 vs. 4 and 5, we observe that this is true to a certain extent: basically whoever can *lie* anywhere and any footed being is a prototypical *'goer'* but only a racer (a specially trained human or a representative of a few animal species like horses and greyhounds) can *lie* in (i.e. *occupy*) a rank position during a race. The collocational potential of *lie* in 1 or 2 is distinctly higher than that of 3 or 4. Implicitly, it is considered likely to occur more frequently in a large corpus. High frequency means high social salience, while low frequency means low social salience. To achieve high social salience, a lexeme must not be confined to one single semantic domain, which is exactly the case of a racer *lying* in a particular position.

On the other hand, the lexeme *to go* as used in 5 is evidently not confined to a single domain: virtually anybody and anything can *be going to* do anything, though expressing intention or likelihood is hardly an especially cognitively salient use of *to go*. It is anyway one of the most socially salient uses of *to go*, though it is intuitively a derived use. Thus there must be another force working against the decrease in the collocation potential. The force's name is *grammaticalization*, and it can be viewed as a range of semantic changes which a lexeme with a quite concrete meaning can undergo. The following sections discuss how introducing of *grammaticalization* affects the general conception of language, and they give a brief overview of the semantic changes symptomatic of *grammaticalization*.

3.2 Emergent Grammar

The leading idea behind the lexicographical description of grammaticalized uses of basic verbs is that of *emergent grammar* explicitly formulated by Paul Hopper [66] and further refined by others, especially Joan Bybee [17] and Bybee, Perkins and Pagliuca [75]. Hopper provides an alternative view of grammar as "a real-time, social phenomenon", which is "always in process but never arriving, and therefore emergent" and "not abstractly formulated and abstractly represented, but always anchored in the specific form of utterance". He challenges the *langue-parole* dualism postulated by de Saussure by questioning the notion of a pre-existing abstract system of language rules, considering grammar as a changeable set of the most useful communicational strategies instead: "Structure, then, in this view is not an overarching set of abstract principles, but more a question of a spreading systematicity from individual words, phrases, and small sets. [...] Grammar is now not to be seen as the only, or even the major, source of regularity, but instead grammar is what results when formulas are rearranged, or dismantled and re-assembled, in different ways.[2] [...] Its forms are not fixed templates but are negotiable in face-to-face interaction in ways that reflect the individual speaker's past experience of these forms, and their assessment of the present context, including especially their interlocutors, whose experiences and assessments may be quite different."[3]

Grammatical changes proceed very slowly. The mechanisms of change manifest themselves in small changes going on scattered on many fronts ([75], p. 24). With this in mind, Hopper's understanding of the linguist's task meets that of the lexicographer's: **"to study the whole range of repetition in discourse, and in doing so to seek out those regularities which promise interest as incipient sub-systems"**. Symptomatically, Hopper does not make any distinction between *grammaticalization* and *lexicalization*. Also [17] regards grammaticalization as a dynamic but very gradual process of **fusion of lexical items into morphemes**.

The approach working with *grammaticalization* is diachronic by nature. The diachronic approach is expected to formulate generalizations about language similarities more effectively when regarding them as *paths of development* than if synchronic states were compared ([75], p. 4). It works on the assumption that there is a set of cross-linguistically universal cognitive concepts which give rise to grammatical categories in the development of each particular language[4], which a cross-linguistic comparison of

[2]Similarly, Bybee in ([75], p. xxii): "We do not take the structuralist position [...]. Rather, we consider it more profitable to view languages as composed of substance – both semantic substance and phonetic substance. Structure, or system, the traditional focus of linguistic inquiry, is the product of, rather than the creator of, substance. Substance is potentially universal, but languages differ in as to how it is shaped [...]"

[3]Cf. [17], p. 212: "...highly frequent forms are learned by rote memorization, and stored and processed as unanalyzed units."

[4]Cf. [8] and partly [82], [83].

the *paths of development* can help trace back.

Bybee [17] lists three ways in which semantic elements may be combined:

1. *Lexical* expression: the semantic elements are expressed by a single monomorphemic lexical item. e.g. the lexical item *kill* combines the semantic elements of 'die' and 'cause'.

2. *Syntactic* expression: the semantic elements are rendered by separate and independent units; e.g. *come to know* for 'inchoative' and 'know', or *to be going to* for future.

3. *Inflectional* expression: the semantic elements are bound into a single word. The respective semantic elements can either take the form of affixes added to a stem *(walk-**ed**)* or just a change in the stem *(bring – br**ought**)*

All three forms of expression can be regarded as *grammatical morphemes* or *grams*, provided that they are "closed-class elements whose class membership is determined by some unique grammatical behavior, such as position of occurrence, co-occurence restrictions, or other distinctive interactions with other linguistic elements" ([75], p. 2). When looking for *grammatical meaning*, Bybee, Perkins and Pagliuca ([75], p. 48), actually pick up areas of the universal semantic space that are frequently grammaticalized across languages.

Inflectional expressions are thought to be the top end of the grammaticalization scale and lexical expressions at the bottom end. To become an inflectional morpheme, the semantic element in question must have a very general meaning, such that it can be sensibly combined with all lexical stems of the appropriate semantic and syntactic category, and it must be obligatory (*the generality principle* – [17]). Also, the more one semantic element affects or modifies the other, the more it tends to take the form of an inflectional morpheme (*the relevance principle* – [17]). Relevance depends on cognitive and cultural salience. The combination of highly relevant elements typically denotes concepts important for the language community. Although the existence of a set of universal concepts is assumed, their importance may vary across languages due to cultural differences. Also, there are undoubtedly some concepts that are community-specific.[5] This is one reason why morphological categories are language-specific.[6]

[5][17], p. 137: "The cross-linguistic comparison of the contents of morphological categories must allow for differences among languages, and it is our task as linguists to discover the systematicity in these differences."

[6]Cf. also [154], p. 27: "How much and what sort of information is expressed by morphology differs widely between languages. Information that is expressed by syntax in one language is expressed morphologically in another one. [...] Also, some type of information may be present in one language while missing in another one."

3.3 The Most Grammaticalized Verb Categories

Bybee [17] made a list of all grammatical categories appearing as inflectionally marked on verbs in 50 non-related languages. She treated the grammatical categories as semantic concepts (or "cognitively significant semantic domains" – [75], p. 3) rather than granting them a structural status, and she examined the way they manifested themselves in the respective languages of the entire language sample. The inflectional expression was considered to be the most grammaticalized. The categories that most often (i.e. in most languages) had inflectional expression were regarded as the most universal verbal grammatical categories. It was the categories *valence, voice, aspect, tense, mood, number, person, person (object)* and *gender* (in this order) that came out as the most universal inflectionally morphological categories of verbs.

It is very tempting to track Bybee's "cognitively significant semantic domains" in a contrastive examination of Swedish and Czech basic verbs and to state which expression they take – the inflectional, the syntactic or the lexical. Inflectionally morphological forms will not be of interest as they are clearly to be described by grammars. Many syntactic expressions are regularly described by grammars of both languages. So far, no breaking discoveries are expected, though the languages have been monitored with respect to contrasting morphosyntactic structures. Yet the main point of interest is lexical expressions that might have already have set off on the way to becoming a syntactic expression.

3.3.1 Voice and Tense

The comparison of both languages reveals that voice and tense regularly take inflectionally morphological and syntactic expressions in Swedish as well as in Czech. These categories are created on the base of well-known rules and therefore they are lexicographically irrelevant. However, their syntactic and lexical expressions will be recorded if found. e.g. light verb constructions that can be characterized by the Lexical Function Oper2 have passive meaning (e.g. the Swedsih *möta kritik* – lit. *to meet criticism* as well as the Czech *sklízet kritiku, čelit kritice* – lit. *to harvest/face criticism*: when X 'meets/harvests/faces criticism' from Y, it means that X is being criticised by Y.

3.3.2 Mood

Mood is expressed mainly syntactically in both languages (except imperative in both and conjunctive forms in Swedish, most of which are obsolete). Mood is rule-based, thus lexicographically irrelevant. Sometimes mood is affected by conversational implicatures in Swedish as well as in Czech, e.g. when using a yes-no question (often in future tense) as a polite imperative: *Will you please shut the window?* As far as it is not

a given lexeme that triggers a conversational implicature, the discourse-conditioned expression form of mood modification is naturally lexicographically irrelevant.

3.3.3 Number and Person

Number and subject person in verbs are expressed by inflectional morphology in Czech and syntactically in Swedish (by the obligatory subject). Neither language expresses object person.

3.3.4 Valency

Valency is expressed syntactically in both languages. Czech uses 7 direct cases, of which 2 cannot be prepositional (nominative, vocative) 4 can be prepositional (genitive, dative, accusative, instrumental) and one must be prepositional (locative), while Swedish only employs subject case, genitive and object case. The object case is used where Czech would use the dative, accusative, locative and instrumental cases. Objects that would typically correspond to Czech accusative and dative objects can be sometimes expressed by direct cases in Swedish, whereas complementations corresponding to Czech locative and instrumental have to employ prepositional cases. As the respective Swedish prepositions are not straightforward matches to the Czech, a cross-linguistic comparison of their uses is appropriate. Therefore valency is to be recorded (as it is anyway common with larger lexicons).

3.3.5 Aspect

While Czech employs many inflectionally morphological aspect pairs (e.g. the imperfective *střílet* and the perfective *střelit*), numerous derivationally morphological aspect pairs (e.g. the imperfective *střílet* and the perfective *vystřílet*) and some lexical aspect forms (e.g. the perfective *vzít* and the imperfective *brát*), Swedish does not employ inflectionally morphological aspect at all, making mainly use of lexical and syntactic means. The implication for a lexicographical description of Swedish verbs designed for speakers of Czech is that special attention must be paid to aspect.[7]

Bybee [17] emphasizes that aspect is a cross-linguistically valid grammatical concept (the third most frequent in the sample of languages).[8] Its commonest values are the *perfective-imperfective* opposition and the second commonest values are the *habitual-continuous* opposition.

The analysis of the 50-language sample showed that the values *perfective-imperfective* are far more grammaticalized (i.e. they take inflectional expression and are placed

[7] [17] does not discriminate between aspect and telicity.

[8] This appears surprising to Slavic speakers, as inflectional aspect is traditionally viewed as a category specific to Slavic languages!

closer to the stem) than the *habitual-continuous* values. (Out of the 50 languages in the sample, 14 had inflectional *perfectivity-imperfectivity* while only 7 had inflectional *habituality-continuity*. Bybee [17] explains this with the *generality criterion*. The value *imperfective* covers *habitual* as well as *continuous*. The meaning of *perfectivity-imperfectivity* has reached such a high degree of generality that it does not even have to affect the meaning of the stem. The values *perfective-imperfective* are therefore also used to mark the position of an utterance in the discourse.[9]

The *habitual-continuous* values, on the other hand, are more specific than the *perfective-imperfective*. Therefore they are less prone to total grammaticalization (i.e. inflectional forms). They are usually rendered as syntactic constructions (very common cross-linguistically). Unlike the inflectional *perfectivity-imperfectivity*, the inflectional *habituality-continuity* is not claimed to have any more generalized grammatical meanings in any of the 7 languages where it occurred.

Other aspectual values often found in the sample were *iterative* and *inceptive* (inchoative). While *habitual-continuous* are regarded as subparts of *imperfective*, *inceptive* and *iterative* are not integrated into *perfective*. They are more often (i.e. in more languages) rendered by derivational or even by syntactic means rather than by inflection. More to say, Bybee [17], p. 149, has noted that in some cases "inceptive meaning was coded grammatically by auxiliaries, some of which were very similar in their original meanings", and she ensures that "it occurs often enough to justify its consideration as a universal of grammatical meaning". Inceptive and iterative syntactic constructions are thus systematically captured in Swe-VALLEX/PNL.

Bybee ([17], p. 151) names a few more meanings listed in grammars as "aspects" that do not qualify. It is other modifications of an event, such as "to do something while moving", "to do something a little", etc., that do not affect temporal relations of the given event. However, they are still interesting for a lexicon whenever they are rendered as lexemes.[10]

This gives the following implication for the lexicographical description: aspect

[9] *discourse foregrounding*, see [67]. Thus the same event can be framed both as perfective and as imperfective, which [17], p. 142, illustrates with a Spanish sentence pair:

(1) *Llovió ayer.*"It rained yesterday." (perfective)

(2) *Llovía sin parar.* "It rained continuously". (imperfective)

The imperfective verb form in 2 suggests that it is describing a circumstance under which another event took place. It is *backgrounded*. The perfective verb form in 1 seems to be the actual event in the discourse. The meaning of *rain* does not change. The speaker selects the aspect value according to his or her own estimation of the role of the given event in the entire discourse.

[10] To name an example from Swedish, such a construction would *be hålla på* in combination with a telic verb, meaning " to be on the verge on finishing something when something happened", e.g. *Han höll på att tömma brunnen när de kom = He was just on the verge of having emptied the well when they arrived.*

must not be reduced to *perfectivity-imperfectivity*. Instead, as many semantic components of event structure must be observed as possible.

3.4 Discovering Regularities in Verb Usage

Adopting Hopper's view [66] that "…any decision to limit the domain of grammar to just those phenomena which are relatively fixed and stable seems arbitrary", an intuitive criterion for grammaticalization is being used here: to be able to grammaticalize, a sequence of morphemes[11] must have a meaning that is learnable as a whole ([17], p. 42).[12] A large corpus of Swedish has been searched for such verbal collocations to identify them as "recurrent strategies for building discourses" (cf. [66]), to itemize the most salient ones, and to sketch out their productive potential.

The candidates for grammaticalization have primarily been sought among verbs that "encode major orientation points in human experience" [75], p. 10, including verbs of motion and location (*stå, gå, sitta, falla, komma* etc.), verbs of physical control (*ta, få, hålla, sätta, ställa, lägga, fatta, fälla*), and even verbs of basic social interaction (*bjuda*).

The description proceeds from form to meaning by means of retrieval of suitable concordances from a corpus. It concentrates on forms rendering semantic features that have to do with event structure, i.e. valency and aspect in the most general sense.

3.5 Semantic Changes in Grammaticalizing Verbs

Accepting that grammaticalizing elements of language must conform to the *generality principle*, grammaticalizing verbs must have a fairly general meaning. The best cross-linguistic instance is the auxiliary verb *to be*, followed by *to become* and *to have*. Nevertheless, there are many more commonly occurring verbs that take part in grammatical structures, e.g. *to go* in the English future *to be going to* and *to come* in the inceptive construction *come to + infinitive*. How do they qualify for auxiliaries? They certainly lose some semantic components during the process of grammaticalization and become more dependent on the context ([75], p. 6). The mechanisms of generalization are thoroughly explained in [8] and discussed in [75]. As the most used terms suggest, many authorities regard generalization, which lies behind or accompanies grammaticalization, as loss of certain semantic components, compared to the core meaning of the original lexeme: *semantic bleaching* (coined by Givón) and *weakening of semantic content* (Bybee, Perkins and Pagliuca). Yet Heine, Claudi and Hünnemeyer [8] argue

[11] In Bybee's terms even entire words are considered as morphemes.

[12] Then it can be accepted as belonging into a single word by a generation of language learners and it can even result in a total fusion of the morphemes.

that generalization is not always a reduction of meaning (p. 40f.). They name examples of negation of the core meaning and examples of addition of further semantic components not present in the core meaning.

When the original meaning of a grammatical lexeme has completely disappeared, the motivation and course of the change cannot be traced back any more; e.g. in the French negative particle *pas* (= "*step*").

The following example from Papago illustrates how a new semantic component can be added to a lexeme: the Papago word for *eye* is also used as the preposition *towards*. The *towards* reading is derived from *eye*. This appears strange, since body parts are normally conceived as static objects. How could the directionality component have arisen? It was due to the association of the eye as a body part with its function: eyesight. Eyesight is in the naive world understanding a trajectory leading from the eye to the viewed object. Hence the expression of 'towards' in Papago is not motivated by the eye itself, but by the way it works.

Generalization typically occurs in the following types of semantic changes:[13]

- *Metaphorical extension from one semantic domain to another*[14]
- *Context-induced reinterpretation*[15].

3.5.1 Metaphorical Extension from one Semantic Domain to Another

According to Heine et al. [8], the opposition between grammatical and lexical meanings of a lexeme is often identified with the *concrete – abstract* opposition. The simplest abstraction is making an object non-referential. In terms of logic, the intensional content of the concept shrinks while its extensional content increases. Then it can be easily exploited as a metaphorical vehicle for associating a given semantic feature with another concept by highlighting the semantic component they share. The semantic component can naturally belong to a stereotypical picture (or *Idealized Cognitive Model*, cf. Lakoff – [82]) rather than being an objectively existent feature; e.g. the utterance *John is a donkey* exploits the non-referential (categorial) *donkey* for transferring the semantic feature "simple-minded", typically associated with donkeys, towards *John*. The same happens to body-part terms when they denote spatial entities. When a concept like *back* is used in the sense of a part of an inanimate object or even the location behind an object, it highlights the semantic feature "location relative to a defined reference point" they share.

[13] according to [8] and partly [75]
[14] Cf. [8].
[15] *inference* or *conventionalization of implicature* in [75]

Metaphorical abstraction relates concepts across semantic domains.[16]
Metaphorical abstraction is the way humans conceptualize the non-concrete aspects of the world. It is the naive picture of the world, in which it does not matter what the world actually is like, but what humans believe it is like. The naive view of the world is anthropocentric. Thus the closest and most discrete objects are parts of the human body and objects that can be physically manipulated. They help to 'manipulate' the less distinct entities in discourse by acting as metaphorical vehicles [82]. This type of metaphor has to be held apart from the traditional view of metaphor as a poetic figure, which is ignored here.

Metaphorical abstraction has linguistic consequences of two types:

1. *structure-preserving abstraction*
2. *structure-changing abstraction*

In the former, the topic (the target structure) has not undergone any linguistic transformation (e.g. part-of-speech change). This transformation would approximately include the lexical and some syntactic expression forms of morphemes according to Bybee [17]. The latter triggers categorial changes (e.g. a noun becoming a postposition), and probably to a significant extent it would correspond to derivational/inflectional and to some syntactic expression forms of morphemes.

Heine et al. assume that the semantic domains make a hierarchy of metaphorical abstraction, through which source structures develop into target structures:

PERSON-OBJECT-ACTIVITY-SPACE-TIME-QUALITY[17]

Just to illustrate one pair, the SPACE-to-QUALITY transfer means that structures suggesting that an object is located at a place or aims in a direction regularly express that the object finds itself in a certain state or a certain situation:

(6) *The country is **sliding into** a depression.*

(7) *Belinda fell completely in love with her daughter: 'I felt **high** for about four days, not thinking about anything but caring for her.'*

etc.

3.5.2 Context-induced Reinterpretation

Heine et al. ([8], p. 70), note that metaphorical transfer, the cognition part, appears rather discrete. However, they propose that the transitions from one semantic domain into another create a continuum of linguistic expressions and call this continuous grammaticalizing process *context-induced reinterpretation*. It is explained on the verb *to go* in the following sentences:

[16] though metaphorical transfers also occur within a single semantic domain

[17] [8], p. 48.

(8) *Henry is going to town.*

(9) *Are you going to the library?*

(10) *No, I am going to eat.*

(11) *I am going do to the very best to make you happy.*

(12) *The rain is going to come.*[18]

Examples 8, 9, and 10 illustrate a SPACE-TIME metaphorical transfer. In Example 8, the verb *to go* has a clearly spatial meaning, whereas in 11 and 12 it has a clearly temporal meaning. Yet Sentences 9 and 10 are ambiguous, depending very much on the context. The sentences can be interpreted in the following way:

(13) *Henry is going to town.* SPACE

(14) *Are you going to the library?* SPACE

(15) *No, I am going to eat.* (as answer to 9) INTENTION (+ relics of spatial meaning are still present)

(16) *I am going do to the very best to make you happy.* INTENTION

(17) *The rain is going to come.* PREDICTION

Both 11 and 12 have temporal meaning, but they differ in the desire of the respective subjects to pursue the event as *rain*, let alone the empty *it*, cannot have a will or desire, while a human can.

To explain the semantic continuum, [8] introduce three idealized stages of semantic shifts:

Stage I: A lingustic form F acquires a side-meaning B in addition to its core meaning A when employed in a certain context. At this stage, the utterance can be ambiguous as long as the context (both intra- and extralinguistic) does not eliminate the ambiguity, and it can be misunderstood by the recipient. (This would apply for 10.)

Stage II: The form F can be used in contexts where only the meaning B can be employed. (This would apply for 10.)

Stage III: The meaning B becomes conventionalized and cognitively salient enough to be conceived as a second meaning of the form F, which becomes polysemous. (This applies for11 and 12). However, the meanings A and B are conceptually linked as the transition was continuous (p. 72).

Heine et al. [8] later revised their A to B model, introducing the terms *focal sense* and *non-focal sense*. In this revised model, A and B at Stage III would be focal senses. At Stage I, B would only be a non-focal sense. It would be only an exploitation of the meaning A. The meaning A is supposed to have a set of conversational implicatures in

[18]Quoted from Heine et al., p. 70. According to an English native speaker's view, 12 sounds unidiomatic and should be rephrased as *It is going to rain*.

addition to its core, partial pragmatic meanings which are triggered by various contexts. When a non-focal meaning B becomes highlighted as particularly suitable for expressing a given communicational purpose, it becomes more frequent and gradually gains its own set of conversational implicatures. Then it grows to a new focal meaning B. B, undergoing grammaticalization, then generalizes even to contexts where formerly only A was accepted.

The revised model of *context-induced reinterpretation* implies the following: when determining the meaning of a grammatical entity, not only the focal meanings have to be observed, but also the conceptually prior non-focal meanings and recurring 'later' meanings likely to develop into new focal meanings must be recorded. Sentences 11 and 12 show the completed development from a volitional to a predictional future. When the structure *be going to* is used with an agentive subject, it typically has the meaning of INTENTION: *I am going to draw this ... so that he can have a full picture.*[19] As a result of the PERSON-OBJECT metaphorical transfer[20], the volitional future construction has been exploited in order to create a new convention, which marks future in events with non-human and non-agentive subjects. The evident conversational implicature is that non-human and non-agentive subjects do not activate the *will* feature in the future since they cannot pursue any will on their own: *It is going to be hot today* (PREDICTION). However, due to the generalization of the new interpretation, the PREDICTION-meaning is extendable back to sentences with agentive and human subjects: *We are going to have a new mum.* Here the structure *to be going to* is ambiguous since without the context or knowledge of the situation it is impossible to tell whether the speakers (potentially volitional) are planning to have a new mum or whether they are rather assuming that this happens, no matter their will.

The context-induced reinterpretation appears to be the most interesting semantic change for a lexicographer seeking out "regularities which promise interest as incipient sub-systems" [66]. It has also been described in other words by Hanks (*exploitations of norms* in [57]) as the result of a long-termed lexicographical work with authentic language data. A few cases have been discussed in Section III.

[19] [8], p. 171ff.

[20] The trasformation of volition into prediction can be seen as the transformation of *X wants* into *X wants to happen* = *X will happen*.

Chapter 4

Light Verb Constructions

4.1 The Notion of Light Verb Construction (LVC)

Kärt barn har många namn.
(A dear child has many names.)
A Swedish proverb

As already mentioned in Section 1.2, the initial impulse for the work on Swedish basic verbs were the German *Funktionsverbgefüge*. The ultimate German grammar textbook for foreign learners [63] contained a list of them. There seemed to be no reason to doubt that Swedish textbooks for foreigners would treat this issue with the same matter-of-factness, especially since the two languages are quite closely related. However, Swedish textbooks for foreigners did not pay much attention to them. This seemed astonishing, as *Funktionsverbgefüge* are very common in Swedish, too. A deeper look into the issue of *Funktionsverbgefüge* alias *support verb constructions* alias *funktionsverbumskonstruktioner* (and many other aliases in different languages) revealed that the innocent, no questions provoking list in a German textbook must have been based on a longlasting and multi-faceted academical discourse, which was not limited to German linguistics and which turned out to be very lively in the Scandinavian linguistics as well[1], only it has not penetrated the students' textbooks yet. This was the actual entrée into the fascinating world on the edge between lexicon and grammar.

Lexical verbs which lose their concrete meaning when combined with abstract nouns and nominalizations and which occur in such combinations very productively, appear to be very common in modern European languages, but also beyond Europe, as already noted by R. Jakobson (for reference see Jelínek [72], p. 50). They were even observed in South-Asian languages [16], which are linguistically as well as culturally very distant from the European languages.

[1]Cf., among others, [6], [10], [11], [14], [15], [34], [36], [141].

German linguistics has studied *Funktionsverbgefüge* and *Funktionsverben* (also under different terms) intensively since the term was coined by von Polenz [130]. Interest in this issue rose especially with the onset of generative and transformational grammar (among others in Rothkegel's studies on fixed syntagms [140]). To be mentioned are also at least Persson's studies on causativity [126], [127], as well as the research in German as a foreign language [63] and [51].

In the English linguistics, a common term for this phenomenon is *light verbs* [74] and *light verb constructions* (alongside *support verbs, support verb constructions, expanded predicates, verbo-nominal phrases, delexical verbs, stretched verbs*, among others).

The terms, especially the German terms as *Funktionsverben, Nominalisierungsverben, verblasste Verben, Streckformen*, etc., cannot be used interchangeably. Some authors using the respective variants were observing only the combinations of a verb and its direct object, others only the combinations of a verb and its prepositional object. For a summarizing comparison of the light-verb related terms in German and English see e.g. [60].

Butt in [16] claims that although light verbs potentially are a universal linguistic phenomenon, they have different structural features in the respective languages[2]. Hence all syntactic tests for defining light verbs and light verb constructions are language-specific (p. 24 in the web-released manuscript of [16]). E.g., in the Germanic languages, the following criteria are commonly named:

- Light verb construction with the predicate noun in the position of the direct object cannot be passivized.
- The predicate noun cannot be replaced with an anaphoric expression.
- There should be at least an option of the predicate noun to occur without determiner (a criterion applied to Swedish, see [34]).

Few verbs are light under all circumstances: those ones that combine only with nominalizations or event nouns, such as *perform, carry out*[3]. The syntactic behaviour of the word combination is an important clue in all verbs that can either act as lexical verbs **or** as light verbs according to their context. However, the syntactic criteria do not apply 100%, and their strict application results in a long list of light-verb-like constructions, defined exclusively by their lacking syntactic salience.

Hanks et al. point out in [60] that "lightness is a matter of degree", and that "some uses [of verbs that can act as light verbs, S.C.] are lighter than others" (p. 441). They emphasize the collocational and semantic criteria for deciding whether a verb

[2]E.g. in Butt's example from Urdu, a light verb construction even requires a second lexical verb attached to the light verb in the verb-noun structure.

[3]In this context it is to be added that evaluative expressions that are neither nominalizations nor event nouns act as such in light verb constructions; e.g. *He committed something horrible.*

use is light or not: "The problem lies in the expectation that necessary and sufficient conditions can be established for delicate grammar categories, as opposed to characterizations of typical features. Light verbs typically focus attention on an event or process, and events and processes are very often expressed in nouns that are nominalizations (i.e. cognates of verbs) – but the focus is still on the event, even when the direct object is a word that denotes a physical entity" (p. 443). They introduce the notion of *semantic lightness* in their analysis of the verb – direct object combinations, and there is no apparent reason not to relate this term also to verb – prepositional object combinations, which their paper does not address.

Butt [16] draws an interesting conclusion from diachronic English studies, which supports favouring semantic and collocation criteria over syntactic – although in their function similar to auxiliary verbs, light verbs, unlike auxiliaries, do not underlie the grammaticalization process in the development of a given language: "Light verbs straddle the divide between the functional and lexical in that they are essentially lexical elements but do not predicate like main verbs" (p. 4 and 13 in the web-released manuscript of [16]).

After a careful consideration, the terms *light verbs* and *light verb constructions (LVCs)* have been adopted for this thesis, as well as the semantic and collocational view of these constructions rather than only the strictly syntactic ones. The nominal part of these constructions is in this thesis called the *predicate noun* or *predicate noun phrase* when the modifiers of the given predicate nouns are explicitly included. The terms *light verb* and *light verb construction* are used as opposed to *support verbs* and *support verb constructions*, with the German equivalent *verblasste Verben* rather than *Funktionsverben, Nominalisierungsverben*, etc. Hence the term *light verb constructions* as used in this thesis includes all the following phenomena:

- constructions with the predicate noun in the position of a direct object
- constructions with the predicate noun in the position of a prepositional object
- constructions with the predicate noun in the position of subject (captured only occasionally in PNL (see Section 16.4))
- verbs that combine only with event nouns
- verbs that combine with event nouns as well as with nouns denoting entities, when the entire construction focuses on an event which gives rise to the entity rather than on the entity itself (e.g. *take a photograph*, cf. [60]).

4.2 LVCs as Collocations

LVCs can be regarded as a type of collocation. Malmgren ([97], p. 12)[4] describes a number of candidate LVCs, calling them a kind of "prototypical collocations" that consist of a semantically impoverished verb and an abstract noun. The abstract noun keeps its meaning, hence it is considered to be the more stable member of the collocation – the collocational base (or *node*, see [149]). Its verbal collocate is generally unpredictable.

Inspired by Mel'čuk's Meaning-Text-Theory (see Chapter 8), Malmgren finds and associates Swedish verbal collocates with nouns by means of the lexical function Oper. Fontenelle ([47], p. 142) also claims that "Support Verbs roughly correspond to the type of lexical relation that can be encoded through the Oper Lexical Function used by Mel'čuk".

The understanding of nouns as collocational bases in verb + abstract noun constructions is clearly shared by Čermák, (e.g. [22]): "Abstract nouns seem to follow a few general patterns in their behaviour, which seem to be more structured, allowing for much less freedom than concrete nouns. The patterns the abstract nouns enter are determined by their function and meaning".[5]

While Helbig and Buscha were seeking to identify a distinct class of "Funktionsverben", and Baron and Herslund [6], Rothkegel [140], and Persson [126], [127] were trying to define light verb constructions by the semantic relation between the noun phrase and the verb, Fontenelle, Malmgren, and Čermák focused on the noun, in full accordance with the pregnantly formulated observation of Hanks [57]: "…it seems almost as if all the other parts of speech (verbs and function words) are little more than repetitive glue holding the names in place".

Even in the cross-linguistic perspective, it is usually the noun that is the common denominator for the equivalent light verb constructions: "The verb […], although often the only one that is correct and idiomatic, can seem totally arbitrary. In another language – mutatis mutandis – totally different verbs often occur which work as place holders; that is why prototypical collocations often cause translation problems" ([97], p. 11, and cf. [143]).[6] Malmgren further notes that "sometimes, though far from every time, one can anticipate a sort of metaphorics" in the choice of the verb. According to Malmgren, the eventual metaphors can be traced back and explained ex post facto, but they are definitely not predictable within any one given language, let alone cross-linguistically.

[4]Malmgren's starting point is the system-oriented understanding of collocations coined especially by German linguists as Hausmann and Heid ([62] p. 302) rather than the original English contextualist approach to collocations.

[5]Though Čermák explicitly avoids the term 'collocation', using the expression 'stable combinations' instead, among which "some are undoubtedly more frequent than others".

[6]The quotations of Malmgren, Ekberg and Dura were translated from Swedish by S.C.

4.3 Semantic Aspects of LVCs

From the semantic point of view, the noun seems to be part of a complex predicate rather than the object (or subject) of the verb, despite what the surface syntax suggests (cf. [143], p. 93, and [11], pp. 53, 145). As already stated by many authors (e.g. [63]), light verbs are in fact lexical verbs that have to some extent lost their lexical meaning, in order to provide the predicate nouns with verbal morphological categories (which is the feature that makes them resemble a verb class according to [63] – *Funktionsverben*, and [72] – *operational verbs* (*operační slovesa,* p. 40)).

Many students of this topic have observed that verbs, when occurring in an LVC, start to carry more abstract semantic features. Rothkegel [140] considers the semantic bleaching[7] of the verb to be the antipode of verbal polysemy. She shows that the meaning of a given lexical verb in LVCs neither matches any of its meanings outside LVCs, nor does it create new meanings when associated with the respective noun phrases, which implies that instead of just being deprived of a part of its original meaning, the lexical verb acquires an additional, more abstract meaning that is reserved for the verb's occurrence in LVCs.

Butt ([16], p. 18 of the web-released manuscript) proposes that light verbs are characterized precisely by the ability to express general features, as described by Rothkegel [140]. However, Butt is explicit in that she does not regard light verb uses as semantic derivations of the primary meanings of the verbs, but contrary to that, she assumes that "the lexical specification of a handful of verbs (somewhere between 5 and 20) cross-linguistically allows for a use as *either* a main verb *or* a light verb. Some common examples crosslinguistically are the verbs for *come, go, take, give, hit, throw, rise, fall*, and *do/make*. [...] Their lexical semantic specifications are so general that they can be used in multitude of contexts, that is, they 'fit' many constellations."

4.4 LVCs and Event Structure

LVCs are often referred to as a means of modifying the event structure of a locution, especially in languages such as Swedish, which do not (regularly) indicate aspect by morphological means (i.e. by stem vowel alternations or affixes). In such languages the aspect remains underspecified, unless lexical markers (e.g. temporal adverbs) are employed in the utterance. A kind of event structure opposition is assumed between an LVC and its corresponding synthetic predicate (when there is one). Butt ([16], p. 18) of the web-released manuscript, in accordance with many other authors, emphasizes that "light verbs modulate or structure a given event predication and do so in a manner similar to that of modifiers with respect to semantic notions such as

[7] She quotes other authors' terms, such as 'das Verblassen der Merkmale bei den Verben", "Bedeutungsentleerung", "depletion of the designatum".

benefaction, suddenness, etc.[8] [...] The light verbs also tend to add further information about the aktionsart of the complex predication. In particular, there is often a telic/boundedness or a causation component." In this respect they have similar function as verbal prefixes or particles ([16], p. 16).

LVCs are built as compositional events or constructions consisting of a 'verbal' and a 'nominal' subevent. Yet the 'verbal' event does actually never 'take place' due to the semantic depletion in light verbs (cf. [44]). The given light verb only passes some semantic features on to the 'nominal' event. Durative events are by definition atelic (e.g. *to have problems*), with the reservation that multiple telic 'nominal' events combined with a durative atelic light verb express iterativity, e.g. *to suffer from attacks*.

LVCs denoting transitions (i.e. changes of state) are generally regarded as telic (cf. [134]), no matter what telicity value the given light verb would have if used as a lexical verb outside the LVC. Bjerre [10] puts it this way: "LVCs denoting transitions are invariably achievements[9], either inchoatives or causatives [...], the SV [i.e. *support verb*, S.C.] always denotes an underspecified subevent$_1$. [...] Not surprising *terminative* is the negative counterpart of *inchoative*."

Bjerre's examples make it more clear: "*Situationen kom ud af kontrol – [The situation came out of control]* denotes a situation in which the resultant state is the negative of that in *Situationen kom under kontrol [The situation came under control]*. [...] This may be paraphrased: (subevent$_1$:) The situation was under control when something happened as a result of which (subevent$_2$:) the situation was out of (= not under) control". Bjerre notes that light verbs denoting transitions are either achievement verbs with inherently underspecified subevent$_1$ (*come, bring* etc.), or they are verbs of motion or location which lose their specific relation when used as light verbs.

4.5 Productivity vs. Lexicalization in LVCs

Whereas traditional views emphasize that it is mostly the lexicalized units that tend to show specific syntax behaviour and, therefore, LVCs are to be considered as more or less lexicalized phrases, Ekberg [36] and Dura [34], as well as Persson [127], concentrate on the apparent productivity of LVCs and the regular production patterns they form. Ekberg notes that many lexicalized phrases "have an almost completely or at least partly predictable meaning and new ones can be formed according to productive rules within the grammar" ([36], p. 32), while Dura goes even further, adding that "even the newly-formed phrases show the same syntactic restrictions as the lexicalized ones" and interpreting this phenomenon as evidence that "these restrictions

[8] Cf. also [143].

[9] Transitions are further divided into two subtypes. In *achievements* the subevent$_1$ is underspecified, unlike in *accomplishments,* e.g. *Carl built a house* (accomplishment) × *The expedition reached the top of a mountain* (achievement). See [10].

indicate that something is meant as a lexicalization rather than that they are the result of lexicalization" ([34], pp. 1–3). She considers article-less verb-noun combinations to be evidence that there is "a kind of word combination that is not controlled by the regular syntax but aims at lexical composition" and that it is thus "possible to form new phrases which can act as lexical units. The ordinary syntax is oriented at combining lexical units with obligatory grammatical categories, but there even seems to be another syntax, a syntax which allows language users to build larger conceptual units without involving the grammatical categories". Dura and Ekberg approach the issue from the semantic side, though they seek to draw syntactic conclusions. The syntactic criteria are eventually more important for Dura and Ekberg than they are for Hanks et al.

4.6 Aspects of Valency in LVCs

The central issue of valency in LVCs is whose valency frame is the predominant one: the one of the light verb, or the one of the predicate noun? Or possibly a combination of both? When drawing an LVC as a dependency tree, which complementations will depend on the verb and which will depend on the noun?

Baron and Herslund ([6], p. 106–111) observe simplex transitive verbs, their nominalizations, and LVCs that contain the given nominalization. They draw the conclusion that nominalizations (i.e. event nouns occurring independently of LVCs as well as their compounds) inherit the valency frame from the entire LVC rather than from the simplex verbs they are derived from. The following example sketches a scale along which the prepositional selection is ordered:

(18) *Terroristerne truer ambassadøren.*
The terrorists threaten the ambassador.

(19) *Terroristerne fremsætter trusler mod ambassadøren.*
lit. *The terrorists make threats against the ambassador.*

(20) *terroristernes trusler mod ambassadøren*
the terrorist's threats against the ambassador

(21) *terroristtrusler mod ambassadøren*
terrorist threats against the ambassador

Example 18 is a sentence whose predicate is a simplex verb. Example 19 is a sentence that employs an LVC whose predicate noun is the nominalization of the verb that occurs in 18. Example 20 is a noun phrase that consists of the nominalization plus the other members of 18. Example 21 is the most 'nouny' one: a noun phrase in which one of the members of the first sentence has become part of a compound.

Baron and Herslund show by means of the selection of the same preposition in the nominalization as in the LVC that the valency pattern of the nominalization is

identical with the one of the corresponding predicate noun but not with the one of the verb. They consider LVCs to be "transitional forms between clauses with simplex verbs and complex nominals". Nøhr Pedersen ([125], p. 210) even claims and shows on examples that light verbs have no valency of their own but they inherit the valency of their predicate noun.

FGD, whose valency theory is the basis for the valency description of LVCs in Swe-VALLEX/PNL, shares the opinion of Čermák [21] and Macháčková [95], who observe the valency potential in the light verb as well as in the predicate noun. Čermák shows different potentials of 'abstract nouns' to acquire valency complementations in LVCs, observing also the valency complementations of the light verb. Macháčková claims that: "When a noun enters an AP[10] it can keep its valency pattern (*mít, chovat úctu ke komu (place trust in someone)*), or – if the light verb has its own valency pattern – it 'obeys' the verb." It is the case with verbs that acquire three valency complementations as *dát (give), poskytovat (provide), vzdát (pay [tribute]), věnovat (grant), projevit (show), vyslovit (convey, express)*. Despite the pattern *důvěra ke komu (trust in someone)* it says *projevit, vyslovit důvěru komu (place one's trust in someone/something)*; similarly, *péče o Jana (someone's care of John)* but *poskytnout péči Janovi (provide care to John)* because of the verbal valency pattern *poskytnout komu co (provide something to someone)*[11].

Hence the surface realization of the valency complementations as well as their organization depend mainly on the light verb; if the verb does not have any valency pattern other than the one with the abstract noun, all other valency complementations are governed by the noun, e.g. *mít zalíbení v kom, čem (have a delight in someone/something)*[12].

However, there is an exception, according to [95]: if a valency complement of the predicate noun takes the form of a spatial adverbial, the given adverbial is usually governed by the verb; e.g. in *budit Janův obdiv (arouse John's admiration)*, *Janův (John's)* is governed by the noun *obdiv (admiration)*, whereas in *budit obdiv u Jana (arouse admiration in John)* the prepositional phrase *u Jana (in John)* is governed by the light verb *budit*[13]. Sometimes the prepositional phrase could be easily associated with the noun, since it can complement the noun even outside an LVC; e.g. *oprava na vodovodu (repair to the water-main) – provést opravu na vodovodu (perform a repair to the water-main)*. The reason is that the prepositional phrase is not an obligatory complementation of the verb, but it is an alternative way of saying *oprava vodovodu (repair of the water-main)*, where *vodovodu (of the water-main)* is clearly governed by

[10]AP = 'analytical predicate' = LVC (S.C.)

[11]This example works much better in Czech than it does in the English translations.

[12][95], p. 136, translated by SC

[13]This construction is ambiguous in Czech. It can mean *arouse John's admiration for X* as well as *arouse someone's admiration for John*, and it is not clear whether Macháčková means both the interpretation or which of the two possible.

the noun *oprava (repair)*.

Kolářová-Řezníčková has performed a detailed analysis of Czech deverbal nouns within and outside LVCs for the FGD-based annotation of the Prague Dependency Treebank [139], [26], and especially [80]. For an introduction into the FGD valency theory see Chapter 6. In short, in the FGD valency theory, the frame-evoking word has valency frames. Each frame is defined by the number and type of the inner participants and obligatory free modifications (for terminology issues see Chapter 6) .

Kolářová-Řezníčková's conclusions have been adopted by the PDT- tectogrammatical annotation manual [105] and also by Swe-VALLEX/PNL. The essential points on the issue of valency in LVCs within the FGD framework are given below[14].

- An LVC takes two entries in the valency lexicon: the verb has its own frame in an entry and the noun has its own frame in an entry.
- When a verb is used as a light verb, the light-verb usage has its own valency frame. The predicate noun acquires the functor CPHR (Compound Phraseme). All nouns marked with CPHR in the data have their entries in the valency lexicon.
- The CPHR-frame of a verb can have three different forms compared to its non-CPHR use(s):
  - The CPHR-frame can keep the non-CPHR valency.

    (22) *Poskytují jim potravu.*PAT
    They give them food.

    (23) *Poskytují jim pomoc.*CPHR
    They give them help.

    (24) *dostat od někoho.*ORIG *úkol.*CPHR
    get a task from someone.

    Some verbs have obligatory free modifications in non-CPHR uses.

    (25) *klást něco.*PAT *někam.*DIR3
    put something somewhere

    The CPHR-use, however, employs the same form (preposition + case) to render the Addressee (ADDR) rather than direction. The semantic change is to be respected and the functor DIR3 is to be replaced with ADDR, as in Examples 26 and 27.

[14]Starting here, the terminology of FGD will be used consequently for valency issues.

(26) *klást něco*.PAT *na podnos*.DIR3
put something on the tray

versus

(27) *klást na někoho*.ADDR *nároky*.CPHR
lit. *put requirements on somebody*

– The CPHR-frame can acquire a new complementation, as in Examples 28 and 29:

(28) *Udělal tuto část*.PAT *diplomové práce*
He made this part of the master's thesis.

versus

(29) *Udělal na mě*.ADDR *dojem*.CPHR
He made an impression on me.

– The CPHR-frame can lose a valency complementation, as in Examples 30 and 31:

(30) *Podal kolegovi*.ADDR *šroubovák*.PAT
He handed a screwer to his colleague.

versus

(31) *Podává špičkové výkony*.CPHR.
lit. *He hands a top-quality performance.*

- When annotating the data, it is the annotator's decision whether a complementation which has occurred in the given sentence should be attached to the verb or to the noun. Nevertheless, there are some basic rules to be followed.

The predicate noun takes its own complementations just as the light verb does. Presumably, "a noun occurring in a complex predicate as its nominal component may borrow a form for the expression of its valency modification which is used to express a referentially identical valency modification of the verbal component (the noun itself would not require such a (form of) modification; in the case of deverbal nouns this form of modification is not present even with the base verb). The given valency modification may then also be expressed in this borrowed form when the noun occurs outside the complex predicate (within the complex predicate the given valency modification is interpreted as dependent on the verbal component)" [105].

When deciding whether a valency complementation belongs to the verb or to the noun, three cases are to be distinguished:

1. The deep frame slot filler (the functor) and the corresponding surface frame slot filler (the surface syntax form) occur either only within the verb or only within the noun frame.

2. The deep frame slot filler (the functor) occurs in both the verb and the noun frame, but its surface frame slot fillers differ for the noun and the verb frame. The surface form associates the deep frame slot filler either with the verb or with the noun.

3. The deep frame slot filler (the functor) and its corresponding surface frame slot filler occur in both the verb and the noun frame, and it can belong in the noun frame as well as in the verb frame and the surface form does not help decide it, e.g. *Petr dostal od šéfa rozkaz přijít včas*, which can be interpreted as *Petr dostal od šéfa rozkaz od šéfa*.

The last case (called 'dual function of a valency modification of the complex predicate') is resolved by means of an annotation convention: valency modifications with dual function are represented as primarily dependent on the node for the verbal component of the complex predicate ([105], Section 9.3.4.1). The complementation is preferably attached to the verb. The frame of the predicate noun is completed by generating new nodes according to the valency lexicon. The node that corresponds to the complementation with dual function acquires the substitutional tectogrammatical lemma QCor (Quasi-Control). Quasi-Control is a subtype of verbal control, which in turn is a type of grammatical co-reference.

Fig. 4.1 shows the tectogrammatical tree structure of an LVC with a valency modification with dual function.

As Swe-VALLEX/PNL is a lexicon and coreference is only indicated in the data in the FGD-based annotation, it does not indicate the coreferential relations between the light verb and the predicate noun within an LVC. It describes the complete light-verb

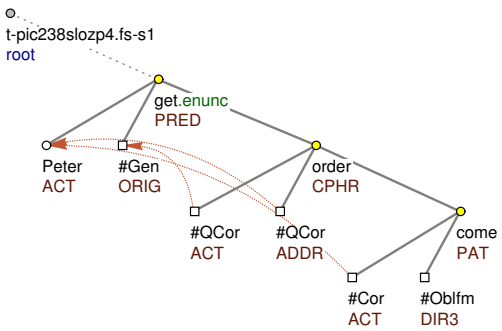


Figure 4.1: *Peter got the order to come.*

frame (in Swe-VALLEX) and the complete predicate-noun frame (in PNL). Preferences of surface realization of the members of the noun frame are included in PNL (see Section 16.4).

4.7 Communicative Aspects of LVCs

Since von Polenz [130], authors have observed numerous semantic features in LVCs that distinguish them from the corresponding simplex verbs and justified their use as not being only an evidence of officialese mannerism. Jelínek [72] points out cases in which using an LVC is advantageous for communication, and so do Helbig and Buscha [63]. Hanks et al. [60] also show that a given LVC usually cannot be exchanged for its corresponding simplex verb, since their meaning is different, and this fact has nothing to do with stylistics.

Jelínek notes that by making use of a nominalization the speaker gains a noun, with which he can refer back in text to a verb:

(32) *Pružina, na které je zavěšeno závaží, se prodlouží. Stejného prodloužení lze dosáhnout, táhneme-li za pružinu rukou.*
The spring, on which the weight is hung, grows longer. The same prolongation can be achieved when we pull the string by hand.

Nominalizing is an efficient way of creating new terms from events. Adding an attribute to the nominalization creates a more specific term, e.g. *vybíjení nabitých izolovaných vodičů (discharging of the charged isolated wires)*.

By expressing the event with a nominalization, the speaker gains the possibility to compress an adverbial into an adjectival or a nominal attribute, cf.

(33) *Vítaným oživením byly ukázky skladeb pro staré nástroje.*[15]
The samples of compositions for old instruments were a welcomed animation.

Sometimes, there is no option other than nominalization, since the modifier cannot be expressed as an adverb:

(34) *M. Bacháček z Nauměřic konal s Keplerem v jednoduché hvězdárně astronomická pozorování.*
M. Bacháček from Nauměřice was performing astronomical observations with Kepler in a simple observatory.

The same sentence shows that the transitive event of observing does not have to express its object when nominalized, whereas when using a simplex lexical verb, the object of the observation must be rendered by the surface syntax. The sentence is unacceptable with object omission, although the adverb *astronomically* would fill the gap from the semantic perspective:

[15] All the Czech examples are taken from [72].

(35) **M. Bacháček z Nauměřic s Keplerem v jednoduché hvězdárně (astronomicky) pozorovali.*[16].
**M. Bacháček from Nauměřice astronomically observed with Kepler in a simple observatory.*

Helbig and Buscha [63] give many examples of LVCs that 'passivize' an event. This can be of use when the agent of the event is unknown, irrelevant, or should not be mentioned:

(36) *Das Buch hat allgemeine Anerkennung gefunden.*
lit. The book found general acknowledgement.
The book was well received.

Sometimes the agent is even present on the surface but 'camouflaged' in a prepositional phrase:

(37) *The problem of biting in companion parrots is a serious one.*

Light verb constructions also facilitate event coordination. Helbig and Busha [63] illustrate it on Examples 38 and 39. The first sentence is rather overloaded, which is particularly evident due to the German word order rules (separable prefixes are placed at the end of the predicate field in finite main clauses), while the other one, where three predicate nouns are coordinated within a light verb construction, is normally acceptable, or at least much better than the previous one.

(38) *Wir klagen ihn an, werfen ihm etwas vor und beschweren uns über ihn.*
We are prosecuting him, accusing him of something and complaining about him.

(39) *Wir erheben Anklage, Vorwürfe und Beschwerde gegen ihn.*
lit. *We are raising a prosecution, accusations, and complaints against him.*

Especially languages with fixed word order make use of LVCs when moving the focus to the right (also [63]):

(40) *Er führte bei den Verhandlungen der Kommission* <u>*Protokoll*</u>. (focus: *Protokoll*)
lit. *He kept at the negotiations of the commission* <u>*a record*</u>.

(41) *Er protokollierte bei* <u>*den Verhandlungen der Kommission*</u>. (focus: *Verhandlungen der Kommission*)
lit. *He recorded at the* <u>*negotiations of the commission*</u>.

Light verb constructions were initially condemned as officialese. Jelínek, Macháčková, and Čermák, as well as Helbig and Buscha, emphasize the significance of LVCs for

[16]This particular example sentence is made-up. All the correct examples in this section have been quoted from [72].

intellectual texts (though Macháčková gives a nice bunch of very expressive and colloquial ones in Czech). Jelínek observes that some LVCs have developed a touch of importancy or officiality of the given event itself (e.g. *navštívit– visit someone* vs. *vykonat návštěvu – pay a visit*). Yet due to the numerous other semantic and grammatical functions of LVCs it would be incorrect to limit their characteristics to making a given text sound more official (cf. also [60]).

4.8 Conclusions

This chapter is a summary of semantic, syntactic, and stylistic observations concerning light verb constructions. The plenitude of opinions raised in connection with verb-noun collocations makes it clear that, whichever criterion is given priority, light verb constructions remain a heterogeneous group with a continuum ranging from typical and non-typical members.

Their internal syntactic relations remain rather unclear. Butt ([16], p. 3) claims that in light verb constructions "the predication is primary and hence monoclausal, the grammatical functional structure is that of a simple predicate: there is only a single subject and no embedding (no control raising)"[17]. The FGD-based annotation indicates a special type of control (quasi-control) between the light verb and the predicate noun. Kolářová-Řezníčková has performed an analysis of the valency behaviour of verbs and nouns within the light-verb frames as well as an analysis of deverbal noun valency outside light verb constructions to find out more or less that the decision whether a modification is governed by the light verb or by the predicate noun is to be taken for each single light verb construction separately.

The light-verb constructions specific quasicontrol indicates that the light verb and the noun share the participants wherever it is not absolutely evident from the surface realization that the given participant belongs only to one of them, which enables the option of capturing them both in the light verb entries in Swe-Vallex and in the predicate noun entries in PNL. The lexically-centered approach formulated by [60] was adopted for the selection of entry candidates.

[17]It is, however, not clear whether this is a universal claim, since she mainly argues with a number of exotic languages.

Chapter 5

The Transitivity Hypothesis

5.1 What is Transitivity and (why) does it Matter?

This chapter attempts to apply a very abstract linguistic hypothesis (the Transitivity Hypothesis) to observations of Swedish daily language use, in order to trace the possible impact of this hypothesis on basic verbs and their noun collocates in light verb constructions. Admittedly, it may turn out that there is no impact at all. Even if there were some, proving this and exploring this impact in a proper way would go beyond the scope of this thesis and mainly beyond the linguistic competence of an occasional, non-native speaker of Swedish.

However, the hypothesis is so fascinating in its extraordinariness, that it invites to a closer investigation and comparison with living language data. Why not take its existence into account when anyway performing a detailed analysis of selected morphosyntactic units and why not provide sorted lexical evidence for further interpretation – the more so since a thorough corpus-based study on the impact of the Transitivity Hypothesis on Swedish has already been published [86], see Section 5.6. Its statistical interpretation was not directly a bullet-proof corroboration of the Transitivity Hypothesis, but, in the context of the entire study, it evidently inspired and encouraged further investigation.

According to the **Transitivity Hypothesis**, formulated by Hopper and Thompson [67], one of the primary cross-linguistically effective goals of certain morphosyntactic categories is to structure the discourse. Information structuring is very important in a discourse. For listeners and readers it is crucial to be able to distinguish between background information and the content being conveyed as substantial at the moment of speaking and writing. A cooperative speaker makes this distinction clear by employing – probably unconsciously – a number of morphosyntactic markers of **discourse backgrounding** vs. **discourse foregrounding** – in other words: the interplay of certain morphosyntactic markers can decide whether an event will be perceived as foregrounded or as backgrounded.

Foregrounded information has typically the following features:

- telic event
- punctual
- volitional
- agent inherently volitional (animate)
- affirmative (not negated)
- real (not interrogative or conditional)
- with a patient
- patient individuated

Backgrounded information has typically the following features:

- atelic event
- durative process or state
- non-volitional
- agent inherently non-volitional
- negated
- unreal (conditional or interrogative)
- lacking patient
- patient not individuated
- no agent (natural processes)

The more significant the effect of the Agent on the Patient, the more foregrounded such an event tends to be. Finally, the more **individuated** the patient (i.e. count, definite, referential, animate, and preferably a proper name), the more likely the given event would be foregrounded. All the mentioned cognitive features have to do with the speaker's assessment of how much the world has changed with the event described. The more substantial the change, the more foregrounding it deserves.

The explanation is, according to Hopper and Thompson, surprisingly straightforward: what attracts attention and deserves a verbal comment is usually a change in the non-linguistic reality, rather than a status quo – at least in narrative texts. Hopper

and Thompson [67] group all the cognitive features just mentioned above (see also Figure 5.1) into one semantic concept, which they call **Transitivity** (with a capital T). Transitivity in the conception of Hopper and Thompson is not confined to verbs that require a direct object, but it is rather a scale-like, continuous semantic phenomenon. It expresses the extent to which a Patient was affected by an Agent in an event. Transitivity can be higher or lower, depending on the values of the features mentioned. Even events with only one or no agent have their degree of Transitivity depending on the feature values that are not related to the number of participants of the given event.

Hopper and Thompson ([67], p. 254) generalize their observations by the claim that the component features of Transitivity "CO-VARY extensively and systematically [...] Whenever an obligatory pairing of two Transitivity features occurs in the morphosyntax or semantics of a clause, THE PAIRED FEATURES ARE ALWAYS ON THE SAME SIDE OF THE HIGH-LOW TRANSITIVITY SCALE". Both [67] and [86] suggest that there is a correlation between verb aspect and the individuation of the noun in the direct object position (cf. [86] and Section 5.6 on Greek compared to Swedish and Polish).

Example sentences 42 – 45 show alternations in the Transitivity features Telicity, Punctuality, and Individuation of the Patient.

(42) *Peter wrote the names down.*

(43) *Peter was writing the names down.*

(44) *Peter was writing names down.*

(45) *Peter was writing.*

Example 42 describes a telic, punctual action affecting a patient that has a clearly limited shape – i.e., the Patient is individuated. It is the most Transitive interpretation of the event in 42 – 45. Example 43 is not punctual, but it is telic. Therefore it is less Transitive than 42. The event described by 44 is lower in Transitivity than 42: it is not punctual, and the Patient is not individuated (we do not know whether Peter is supposed to write down a specific number of names or whether he is supposed to keep on writing names down for an indefinite period of time). The last example (45) is the least Transitive of all – it is not punctual, and the Patient is generalized (even if it is evident from the context).

According to the Transitivity Hypothesis, if the verb-related features Telicity and Punctuality correlate with the noun-related features that together represent Individuation, as suggested by [67] and studied on corpus data by [86], they should, wherever co-occurring, have their values on the same side of the High – Low Transitivity scale: **"If two clauses (a) and (b) in a language differ in that (a) is higher in Transitivity according to any of the features [...], then, if a concomitant grammatical or semantic difference appears elsewhere in the clause, that difference will also show (a) to be higher in Transitivity"** ([67], p. 255). Hence, foregrounded

sentences are expected to combine punctual/telic verbs with individuated patients, whereas backgrounded sentences are expected to combine processual/atelic verbs with non-individuated patients. Indeed, a sentence such as 46 sounds odd, at any rate compared to 47, since *write* in the simple past tense evokes a more telic and/or punctual reading than when used as a past participle, and the absence of a Patient appears inappropriate in 46, but not in 47.

(46) *Peter wrote.*

(47) *Peter was writing.*

A problem with the Transitivity Hypothesis arises when we compare sentences that differ in more than one Transitivity feature value. Do all the cognitive features have the same weight? Is, e.g., a punctual transition with an agentive Agent and a volitional action like *Peter stood up* less transitive than the non-volitional *Peter forgot his bag at school* just because it cannot get any points for the features associated with the direct objects, while *forget* does?

Do the values of the Transitivity-affecting morphosyntactic features change according to which aspect of the event the speaker decides to focus on, or should 42 – 45 sooner be conceived as different event types? In other words, are the selected combinations of morphosyntactic values just an indicator of the given Transitivity degree that is inherent in the particular event, or do the morphosyntactic features, perhaps along with some lexical and contextual features, provide the sentence with its particular degree of Transitivity? Is it cognition that shapes the grammar here, or vice versa?

In addition, the Transitivity Hypothesis, as formulated above in bold, implies that languages can differ as to which feature values should be counted as salient and which are only 'concomitant' when expressed by a morphosyntactic category (and therefore probably should not affect the score).

Lindvall [86] showed that the assumed correlation between the verb event structure and the noun definiteness applies for a language in which both are rendered by morphological categories (Greek), but the comparison of Polish and Swedish parallel texts (see Section 5.6 for details) concluded with a supposition rather than with a truly revolutionary finding in respect of the correlation between the grammatically rendered verb aspect in Polish and the grammatically rendered noun definiteness in Swedish.

However, according to Lindvall, the statistics would have been much more in favour of the Transitivity Hypothesis if the rating had been expanded beyond the purely morphosyntactic markers towards semantic hints in the context. Therefore it is still not certain whether the correlation between event structure and noun definiteness is also present in languages that do not render one of the supposedly paired categories by obligatory morphological categories and whether such pairing can be assumed to be cross-linguistically universal or not.

Because the answers to these questions are obscured or absent, the idea of Hopper and Thompson is in danger of remaining just a sophisticated mental experiment without much practical use for linguistic analysis. On the other hand, the observations made on a parallel corpus of three concrete languages [86] are at any rate very interesting. Assuming that the Transitivity Hypothesis is effective in 'ordinary' verb + direct object phrases, could we also trace it in light verb constructions – at least in those formulated as verb + direct object? Are perhaps light verb constructions, or even certain morphosyntactic variants of light verb constructions in particular, preferred when expressing a particular type of event (with respect to the cognitive features relevant for Transitivity)?

Corpus research on light verb constructions shows that the noun phrases in (Swedish) light verb constructions often have certain morphosyntactic restrictions regarding article use, number, the option of modifier insertion, etc., but at the same time they are often very productive and far less 'frozen' than true idioms regarding their morphosyntactic variability. Unlike idioms, light verb constructions are collocation clusters positioned somewhere between lexicon and grammar. For instance, *låda* in the idiom *hålla låda (keep talking, typically expressing criticism, without letting the others have their say)* cannot be used with an article. On the other hand, *rekord* in *sätta rekord (set a record)* seems to have an articleless default, which can be modified with determiners and attributes according to the rules of regular grammar.

Intuitively, one would say that this is an indication of a lower degree of lexicalization than observed in idioms; but what about if the grammar comes in between for a particular cognitive reason and causes the cluster to be marked in some way, while the grammar-ignoring (article-less) realization is unmarked? Could the morphosyntactic restrictions or preferences of the noun phrases observed in light verb constructions possibly have something to do with the Transitivity Hypothesis rather than simply with 'incomplete lexicalization'?

A small case study follows in Section 5.2. Section 5.3 gives a more detailed account of the background of the Transitivity Hypothesis, as coined by Hopper and Thompson, and Section 5.6 gives an overview of the existing study of Transitivity in discourse [86]. The practical consequences of the Transitivity Hypothesis for a lexicon of predicate nouns are explained in Section 16.4.

5.2 Grammatical Interference in Lexicalized Collocations?

When the morphosyntactic behaviour of a multi-word cluster systematically deviates from the regular grammar rules, it is traditionally regarded as intensively lexicalized, i.e. several words are thought of as growing together into one single semantic unit. Moreover, Dura [34] suggests that the cause – consequence relation also works the

other way round: collocations that **are meant** by the speakers to be perceived as a single semantic unit are deliberately taken out of the regular language system.

Many authors since the onset of corpus linguistics have observed that the regular language use to a significant extent consists of prefabricated blocks. Needless to say, this phenomenon goes far beyond idioms and terminology. For instance, Wray [164] builds her hypotheses on formulaic sequences around the premise that "although we have tremendous capacity for grammatical processing, this is not our only, nor even our preferred, way of coping with language input and output. [...] much of our entirely regular input and output is not processed analytically, even though it could be" (p. 10).

Many basic verb collocations, especially light verb constructions, appear to be such formulaic clusters (see Chapter 4 for a more detailed discussion). Not surprisingly, an important part of the lexicographical description of basic verb collocations (though this particular context almost requires the term *colligations* – [46]) is the systematic identification of their possible morphosyntactic preferences. The morphosyntactic preferences are therefore explicitly recorded within the entry structure of the lexicon of predicate nouns presented in this thesis (PNL, see 16.4).

Collocations that sometimes behave according to grammar rules and sometimes do not would normally be regarded as somewhere about half-way to the ultimate lexicalization; i.e., they would be expected to exhibit only irregular behaviour in the future development of the given language.[1] However, the underlying assumption in this thesis is that morphosyntactic realizations of semantically transparent collocations in text do not just vary in the extent to which they comply with the rules of grammar in terms of 'right' versus 'wrong', but that different grammatical realizations of collocations can have different semantic/pragmatic implicatures in the particular context according to the speaker's preference. The default behaviour of lexicalized semantically transparent collocations may often be irregular (e.g. zero article, no modifiers allowed, etc.), but the corpus evidence suggests that there is not necessarily a clear ban on a step back to the regular grammar when the morphosyntactic features help reflect the communicational intentions of the speaker in a particular discourse situation.

In other words, the assumption is that regular morphosyntactic behaviour is reintroduced when the speaker explicitly wants to add the semantic features triggered by regular morphosyntactic behaviour, but he is by no means obliged to do it. The presence or absence of semantic differences between two or more alternative morphosyn-

[1]This is of course not the case of idiomatic expressions, whose idiomatic meaning is inseparable from their morphosyntactic realization; e.g. *abandon ship*. Example 1 implies that the ship is thought to be sinking, whereas Example 2 lacks this implicature:

(1) *Abandon ship!*

(2) *They abandoned the ship in a bay near Hong Kong.*

tactic structures is very much context-dependent, and the semantic oppositions can be obscured by the fact that they happen to be irrelevant in a particular context. That implies that the alternative expression forms will not always be mutually exclusive, but that the speakers only have the option to select the non-default pattern when they feel a particular reason for doing that.

To mention a Swedish example, the light verb construction *sätta rekord (set a record)* is normally used without an article, even when *rekord* is modified by one or more adjectives (adjective modifiers usually require the use of an article in Swedish):

(48) *Mustafa Mohammed satte personligt rekord.*
Mustafa Mohammed set a personal record.

(49) *Stefan Holm klarade 2,37 i Globen och satte nytt personligt rekord.*
Stefan Holm made 2.37 in Globen and set a new personal record.

The collocation *sätta rekord (set a record)* appears to be a very lexicalized one, judging from the predominating zero article. The large Swedish corpus Konkordanser showed that the absolute majority of the occurrences of *sätta rekord* had no article preceding *rekord*. The Konkordanser subcorpora yielded 223 occurrences of the forms *sätta, sätter, satte* and *satt*, respectively, with *rekord* following within the same sentence[2]. The noun *rekord* occurred with the indefinite article only 17 times. The percentual rates were the following:

- 2 % in the infinitive
- 0 % in the present tense
- 11 % in the simple past tense
- 9 % in the perfect tense

The definite singular form *rekordet* and the definite plural form *rekorden* occurred 11 times and once in collocation with *sätta*, respectively.

The 29 hits with (any) article represented 12% of the total of 235 hits.

The most frequent case (indefinite article) does not seem to be affected by tense. A closer analysis of the broader contexts showed at least one situation in which the insertion of the indefinite article may be triggered by the context (approx. 1/3 of the hits with the indefinite article) – it is when the discipline in which the record was set is specified later in the text (selection):

[2]Unfortunately, in Konkordanser, modern Swedish texts (newspapers and fiction) are split into 14 subcorpora, and the interface does not allow multiple selection. None of the subcorpora in Konkordanser is either tagged or lemmatized, and the interface does not support CQL. Simple Boolean queries or wildcard searches can be performed, but they cannot be combined, which significantly limits the searching power.

(50) *Svensson har satt ett oslagbart svensk rekord som sportjournalist: under cirka 49 år hade han fast jobb på samma redaktion i samma tidning, Arbetet i Malmö.*
Svensson has set an unbeatable Swedish record as a sports journalist: for approximately 49 years he had had a regular job at the same publishing office, at the same newspaper, Arbetet i Malmö.

(51) *Förre RIK-aren Peter Gentzel har satt ett nytt rekord i tyska Bundsliga. Den svenske landslagsmålvakten har på 34 omgångar tagit hela 53 straffar för Nordhorn.*
Former RIK-player Peter Gentzel has set a new record in the German Bundesliga. The goalkeeper of the Swedish national team has got 53 yellow and red cards for Nordhorn in 34 rounds.

(52) *Massorna, som köade i en halvmil för att slutligen komma till Hyde Park, satte ett nytt rekord i levande opinionsbildning.*
The crowds that were queuing for a half mile in order to finally get into Hyde Park set a new record in live opinion making.

(53) *Anette var andra halvlekens gigant och satte då ett personligt rekord. – Har aldrig gjort åtta mål i en och samma halvlek i elitserien.*
Anette was the giant of the second half and it was then that she set a personal record. – I have never shot eight goals in a single half in the elite series.

In other two cases (one with an indefinite pronoun) the sentence describes an unreal or non-specific condition (Cf. 5.3.7):

(54) *Han säger att visst, landslaget skulle väl vara kul och visst sätta ett svenskt rekord skulle väl också vara kul, men det är saker han inte går och tänker på.*
He says that yes, the national team would obviously be cool and obviously it would also be cool to set a Swedish record, but that is stuff he doesn't go thinking about.

(55) *Om jag sätter något rekord så kommer det snart någon och slår det.*
Even if I set a record, someone else will soon come and break it.

Also setting two entities in contrast normally requires an article, as can be seen in Example 56:

(56) *Hägerstenskillen […] satte ett personligt rekord och tangerade ett: Han presterade 60 kilo i stöt (tangerat pers.) och 47,5 kilo i ryck (personligt med 2,5 kilo).*
The guy from Hägersten […] set a personal record and attacked another one: He lifted 60 kg ……

In addition, the discipline in Example 56 was specified later.

Example 57 originates from a context where records were expected in several different disciplines. A certain swimming discipline was the first discipline in the entire

competition where it happened: a European record was set. In this particular context, the European record, in the context of one single discipline a unique uncountable entity, is regarded as countable and a member of a set.

(57) *Engelsmannen Adrian Morrhouse blev den första att sätta ett Europarekord i Strasbourg.*
The Englishman Adrian Morrhouse was the first one to set a European record in Strasbourg.

In all the other 10 hits except one, the noun *rekord* with the indefinite article was modified by one or two adjectives. All of the adjectives denoted restrictive attributes. The use of a restrictive attribute implies that that particular record was one of a set, which is normally a good reason for employing an article. Nevertheless, the zero-article is strongly preferred in this context and with the modifiers *svensk (Swedish), personlig (personal), ny (new)*, even when they concatenate. No differences in the broader context were observed that would explain why the article was used. Only a selection is presented here.

(58) *Även om serien inte var perfekt satte han ett nytt prydligt personligt och svenskt rekord med 387,60 poäng.*
Even though the series was not perfect he set a new nice personal and Swedish record by 387,60 points.

(59) *Orbit Air vann både försök och final i fjol och satte ett nytt svenskt rekord.*
Orbit Air won both the trial and the final last year and set a new Swedish record.

The definite article (found 12 times) was consequently used when referring back to one particular record mentioned before – either to the same entity (the same discipline, the same year, the same person), or to a contrasting entity. Only a selection is presented.

(60) *Hennes svenska rekord på 1.500 meter på 4.09,0 är internationellt gångbart och den tiden är ingen yttersta gräns för Gunilla. Det finns mer att ge. – När jag satte det rekordet var jag inte ens trött efter loppet. Det kändes som att dansa fram.*
Her Swedish record in the 1 500 meters at 4.09,0 is internationally accepted and this time is not the ultimate limit for Gunilla. There is more to give. – When I set that record I was not tired at all after the run. It felt like dancing.

(61) *När Bartova satte det kortlivade rekordet i Prag snodde hon det från just Flosadottir som tog sig över 4,42 …*
When Bartova set the short-lived record in Prague, she had just stolen it from Flossadottir, who got over 4,42..

(62) *Det svenska skattesystemet sätter det ena otroliga rekordet efter det andra.*
The Swedish tax system sets one incredible record after another.

It is interesting to investigate to which extent the regular grammar continues to affect multi-word clusters that already have reached the stage of lexicalization, which in principle allows them to ignore grammar. This kind of research suggests the cases in which speakers may deliberately decide **to exploit** grammar in pursuit of a particular communicative goal, since they are not forced to respect grammar for its own sake. Investigating grammar in positions where the default is not to use it at all can reveal a lot about the semantic potential of our traditional grammar categories in general.

The analysis of the noun phrase in the light verb construction *sätta rekord* shows the way the morphosyntactic behaviour of predicate nouns is investigated in PNL. PNL systematically captures the alternative morphosyntactic patterns when any are found. However, it does not explain the motivation for their use. Listing the morphosyntactic options is a way of gathering evidence for further research into how discourse and pragmatics affect the collocational patterns of basic verbs, especially light verb constructions; e.g. whether the article use or a modifier in the noun phrase affects the event structure. Proving and formulating the rules of this assumed 'discourse grammar' goes beyond the scope of this thesis. The following sections provide a brief overview of the Transitivity Hypothesis, which gave rise to the considerations discussed above.

5.3 Transitivity Indicators

The Transitivity Hypothesis claims that many common grammatical features in correlation determine the Transitivity[3] of each utterance in a given text. Transitivity is regarded as a central semantic concept consisting of several cognitive parameters (see Section 5.1). Each cognitive parameter of Transitivity 'involves a different facet of the effectiveness or intensity with which the action is transferred from one participant to another" ([67], p. 252). The respective cognitive parameters were obtained by observing the semantics of grammatical categories across a range of unrelated languages. The sets of cognitive parameters relevant to Transitivity and expressed by morphosyntactic means vary from language to language.

The notion of *Transitivity* extends the traditional perception of *transitivity* from the effect that an agent of a transitive verb has on the direct verb object to the effect an event has in general. This means that Transitivity (unlike *transitivity*) is not confined to verbs having direct objects. Hopper and Thompson note that "morphosyntactic markings tend to be sensitive to Transitivity as a whole, rather than to the actual presence or absence of a second participant." Fig. 5.1[4] shows the cognitive parameters

[3]When relating to [67], the term *Transitivity* is capitalized as in the original paper.
[4]The figure originates from [67].

	HIGH	LOW
A. PARTICIPANTS	2 or more participants, A and O.[1]	1 participant
B. KINESIS	action	non-action
C. ASPECT	telic	atelic
D. PUNCTUALITY	punctual	non-punctual
E. VOLITIONALITY	volitional	non-volitional
F. AFFIRMATION	affirmative	negative
G. MODE	realis	irrealis
H. AGENCY	A high in potency	A low in potency
I. AFFECTEDNESS OF O	O totally affected	O not affected
J. INDIVIDUATION OF O	O highly individuated	O non-individuated

Figure 5.1: The Cognitive Components of Transitivity

that determine the **degree** to which an utterance is Transitive. A short explanation loosely summarized from [67] follows.

The table has three columns. The first column lists the parameters. The second column lists their possible values that would imply **high Transitivity**. The third column lists the values that are characteristic of **low Transitivity**. *A* means Agent and *O* stands for Object (patient). The more of these parameters have the value 'high' in an utterance, the more Transitive the utterance is.

5.3.1 Participants

No transfer of any effect at all can take place unless at least two participants are involved. However, even utterances with one participant can be highly Transitive due to 'high' values in the other parameters!

5.3.2 Kinesis

Only actions (transitions and processes) can be transferred from one participant to another. Non-actions as states cannot.

5.3.3 Aspect

Here Hopper and Thompson use *telicity* and *aspect* interchangeably (despite a discussion on p. 270f.). This thesis has adopted the conception of Nakhimovski (cf. [109]): Verbs can be inherently telic by the character of the event they denote, or the entire utterance can be made telic by the context (e.g. by a suitable adverbial, a direct object etc.), or the verb itself again can be made perfective by the use of the perfective aspect (provided the given language expresses aspect by morphological, syntactic or affixation

means). Telic as well as perfective events are more Transitive than atelic/imperfective events.

5.3.4 Punctuality

Transitions (punctual changes of state) and punctual processes (e.g. *give someone a kick*) count as much more Transitive than events that are inherently ongoing.

5.3.5 Volitionality

Transitivity is high when the agent (A) acts purposefully (cf. *I wrote your name* vs. *I forgot your name*).

5.3.6 Affirmation

Affirmative utterances have high Transitivity, negative utterances have low Transitivity.

5.3.7 Mode

The values 'realis' and 'irrealis' denote the difference between an event that takes/took place and one that does not/has not. An event that either did not occur, or which is presented as occurring in a non-real (contingent) world, has a lower impact on the real world than one whose occurrence is actually asserted as corresponding directly with a real event.

5.3.8 Agency

This parameter captures the difference between agents that normally have their own will and are expected to pursue it and those that do not. Agentive agents are typically humans, non-agentive agents are typically nature forces, abstracts and ianimate items. Agentive agents increase the Transitivity of the given utterance, non-agentive agents decrease it.

5.3.9 Affectedness of the Patient

The degree to which an event is transferred to a patient is a function of how completely the patient is affected by the event. (Cf. *She drank up the milk* vs. *She drank some of the milk*.) High affectedness contributes to high Transitivity and vice versa.

5.3.10 Individuation of the Patient

The parameter Individuation captures the distinctness of the patient from the agent as well as the distinctness from its own background. The table in Fig. 5.2, quoted from ([67], p. 253), summarizes the features that decide whether an entity is individuated or not. High individuation contributes to high Transitivity, low individuation contributes to low Transitivity.

5.4 Transitivity as a Discourse Marker

Hopper and Thompson showed by data comparison that "language universally possesses morphosyntactic structures which reflect the degree of Transitivity of a clause. The pervasiveness of these devices and their similarity across languages seem to demand an explanation in a higher-level, functional framework", a "general pragmatic function" that has the power to create such a "linguistic universal". This pragmatic function of Transitivity is probably motivated by the requirement on the "language users to design their utterances in accordance with their own communicative goals and with their perception of their listeners' needs" ([67], p. 280).

One of the essential needs of both the speaker and the listener is a distinction between which utterances in a discourse that are the most relevant ones **(foregrounded)** and those that are just complementary **(backgrounded)**. **Foregrounded** utterances together "comprise the skeleton of the text, forming its basic structure" (p. 281). They are usually ordered in a temporal sequence (at least in narration). A change in the order of any two of them signals a change in the order of real-world events. **Backgrounded** utterances, on the other hand, add details, other circumstances, and comments to foregrounded events, but they could be left out of the text without affecting the overall meaning. They are not ordered with respect to each other, and may even be movable with respect to the foregrounded portions of text.

Hopper and Thompson believe that it is just Transitivity that is the linguistic universal that keeps apart the foregrounded and the backgrounded utterances in discourse. They claim (and prove on a small corpus) that the higher the Transitivity of a

INDIVIDUATED	NON-INDIVIDUATED
proper	common
human, animate	inanimate
concrete	abstract
singular	plural
count	mass
referential, definite	non-referential

Figure 5.2: Individuation of the Patient

given utterance is, the more likely it is that the utterance is one of the foregrounded ones. This claim is expected to operate also the other way round: utterances intuitively perceived as foregrounded are expected to exhibit morphosyntactic features (of the set available for the given language) that reflect high-Transitivity values rather than those that reflect low Transitivity.

5.5 Morphosyntactic Consequences of Transitivity Hypothesis

Hopper and Thompson name many examples from unrelated languages in which morphosyntactic features reflect the cognitive parameters listed in Figure 5.1. Very often two or more morphosyntactic categories correlate to reflect one or more cognitive parameters of Transitivity. The most illustrative example is the correlation between a verb and its object, which is a very common language phenomenon. The basic pre-knowledge of this issue says that a transitive verb is a verb that requires a direct object. However, Hopper and Thompson present many examples of languages whose morphosyntax is so sensitive to 'low' values of Transitivity parameters, that even an object that would be perceived as direct object in languages less sensitive in this respect would be explicitly marked as non-object. That a direct object is not regarded as a full-value object is in these languages indicated e.g. by assigning the non-object a case that is reserved for arguments of intransitive verbs or by merging the participant with the verb. It is typically non-referential objects that are regarded as such inferior objects, and they trigger morphosyntactic features that are used in combination with intransitive verbs; e.g. a different object case required solely by intransitive verbs, word order typical of intransitive predicates, or merging with the verb.

The observations made by Hopper and Thompson show very clearly that the correlation between the morphosyntactic features of a transitive verb and its direct object is a common feature among many languages. However, the examples collected by Hopper and Thompson are taken almost entirely from exotic languages. To make the issue more understandable, the example presented below is from German.

The German perfect tense uses two auxiliaries: *haben (have)* and *sein (be)*. Inherently transitive verbs always build the perfect tense with *haben*. Intransitive verbs use *haben* or *sein*, mostly according to context, but not interchangeably. The rule of thumb is to use *sein* whenever the verb allows no passive and whenever it cannot have any transitive reading. e.g. *fahren* (*go by a vehicle* or *[about some vehicles] go*) is normally intransitive:

(63) *Er ist zu schnell gefahren.*
He went/drove too fast.

It can nevertheless also have a transitive reading:

(64) *Er hat diesen Wagen nur ein Jahr gefahren.*
He had been driving this car just for a year.

Now, German productively combines the verb *fahren* with many different objects that specify the vehicle which was moved. Although German uses articles (definite, indefinite) to indicate noun definiteness, these vehicles are usually not introduced by any article, and the perfect tense is built only with *sein*; i.e. as **if *fahren* in the sense *drive something* were intransitive**. More to say, before the German spelling reform 2006 the spelling varied for different vehicles. For instance, *Auto fahren (drive a car)* has always consisted of two tokens, while *Rad fahren (go by bicycle)* could also be spelled as *radfahren*. (The current norm prefers separate spelling and noun capitalization in all vehicles.) The German speakers evidently conceived a vehicle as a sort of separable prefix of the verb *fahren* rather than as a regular direct object – and some vehicles more than others, to top it. German has an entire system of separable and inseparable prefixes, which made it quite easy for the predominating cognitive conception to develop a syntactic reflection by analogy. There are many more German verb-noun clusters in which the noun is somewhere between a direct object and a separable prefix. Before the recent 'spelling reform', direct objects could even develop into unseparable prefixes (e.g. *hohnlächeln (to smile mischievously)*).

Referentiality is just one facet of the complex phenomenon of noun definiteness. Many languages conceive definiteness as referentiality rather than an issue of prior familiarity ([67], p. 288), others single out animate entities (e.g. Czech animate masculine nouns have a specific declension pattern), while the rules for article use in Germanic languages involve a mixture of prior familiarity, implicatures from common knowledge, and referentiality. Swedish identifies animacy by pronominal anaphora, having two sets of personal singular pronouns for the animate and the inanimate entities, respectively. Whichever of these aspects of definiteness is positively present in the utterance in question, it increases its Transitivity.

Hopper and Thompson generalized their observations into the conclusion that the component features of Transitivity "CO-VARY extensively and systematically [...] whenever an obligatory pairing of two Transitivity features occurs in the morphosyntax or semantics of a clause, THE PAIRED FEATURES ARE ALWAYS ON THE SAME SIDE OF THE HIGH-LOW TRANSITIVITY SCALE" (p. 254), and they formulated the **Transitivity Hypothesis**: "**If two clauses (a) and (b) in a language differ in that (a) is higher in Transitivity according to any of the features A-J, then, if a concomitant grammatical or semantic difference appears elsewhere in the clause, that difference will also show (a) to be higher in Transitivity"** (p. 255).

5.6 Relation between Aspect and Definiteness

Lindvall [86] and [87] has performed a comprehensive parallel-corpora based comparison of Greek, Polish, and Swedish to look into verbal boundedness (the term she uses for telicity and perfectivity) and object definiteness as two major interacting components of Transitivity in many languages, proposed by [67]. Presumably, her inferences regarding Polish will also apply to Czech, as Czech and Polish are closely related languages. She showed by an analysis of Greek (a language employing both morphological aspect and noun definiteness) that utterances with high Transitivity tend to have perfective verb forms along with definite objects, while utterances with low Transitivity tend to have imperfective verb forms and indefinite objects. Then she compared translations between Swedish (a noun-definiteness language) and Polish (an aspectual language) in both directions. It became evident that in utterances with high Transitivity, Polish translations from Swedish tended to have perfective verb forms and Swedish translations from Polish tended to have definite noun forms, while low Transitivity utterances tended to have imperfective verb forms (Polish) and indefinite noun forms (Swedish). The observed noun definiteness was not confined to the article use, but resulted from the semantics of the entire noun phrase, which, on the other hand, was very often reflected by morphosyntax. The examples below illustrate the correlation between verb aspect and noun definiteness on pairs of Swedish and Polish sentences [86]:

Polish originals – Swedish translations

(65) *Wkładała* (Imperfective Past) *złoty łańczuszek z ziarenkami.*
lit. *She used to put on* (Imperfective Past) *golden chain with links.*

(66) *Hon satte på sig en gulkedja* (Indefinite Article) *med länkar.*
lit. *She put on a golden chain* (Indefinite Article) *with links.*

(67) *I opowiedzał* (Perfective Past) *historyjkę o myśliwych.*
lit. *And he told* (Perfective Past) *story about the hunters.*
Och så berättade han historien (Definite Article) *om jägarna.*
And he told the story (Definite Article) *about the hunters.*

Swedish originals – Polish translations

(68) *Halmhatten kastade en mörk skugga* (Indefinite Article)*över hans solbrända ansikte.*
The straw hat was casting a dark shadow (Indefinite Article) *over his sunburnt*

face.
Słomkowy kapelusz rzucał (Imperfective Past) *cień na jego opaloną twarz.*
The straw hat was casting (Imperfective Past) *a dark shadow over his sunburnt face.*

(69) *och såg katten* (Definite Article) *slinka in genom hålet i dörren*
and saw the cat (Definite Article) *slip through a hole in the door. i zobaczył* (Perfective Past) *kota wślizgującego się przez szparę w drzwiach.*
and saw (Perfective Past) *the cat slip through a hole in the door.*

Lindvall noted that in her data (parallel sentences with direct object predicates), the combination perfect tense+definite noun occurred in 57% of the sentences, which might not look very convincing, given that without any correlation expected it would have been around 51%. However, the ratio changed significantly when Lindvall went through the data and considered the **semantic** definiteness of the object (i.e. not only the article use, but also the use of personal vs. indefinite pronouns, count vs. uncount nouns, etc.). Statistically, even the original result based only on article use had only 0,001 (χ^2) coincidence rate. The most significant group in the data were nevertheless sentences that combined the imperfective predicate and a zero-article object.

5.7 Transitivity Hypothesis in Light Verb Constructions

Lindvall's observation that the most regular combination of verb aspect and noun definiteness was *imperfective verb* + *zero article* gave rise to the question how the Transitivity Hypothesis would apply to light verb constructions, which are an important issue in the description of basic verb collocations.

Light verb constructions typically use the zero article in predicate nouns, both when in the position of the direct object and in the position of indirect object. Alternatively, there is usually at least an option of using the zero article among other article options. Dura [34] even goes so far as to claim that Swedish light verb constructions are defined by the ability to employ zero-article in the predicate nouns.

This can be related to the Transitivity Hypothesis: the predicate noun in a light verb construction typically denotes an event or a state; or, even if it does not denote an event or a state on its own, in the collocation with a light verb it focuses on an event or a state (Cf. *take a photograph* discussed in [60]). The zero article often indicates non-referentiality. Event nouns are non-referential by default. Therefore it is hardly acceptable to replace a predicate noun with an anaphor (in the same position):

(70) *Varje dag körde hon in till stan och fattade* ***beslut*** *som rörde de mest utsatta och svaga.*

Every day she drove downtown and took decisions that affected the most exposed and weak ones.

(71) *Nu var det tid för beslut*[5]. **Varje dag körde hon in till stan och fattade* ***dem***[= beslut] *som rörde de mest utsatta och svaga.*
*Now was the time for decisions. *Every day she drove downtown and took* ***them/the ones/such ones*** *that affected the most exposed and weak ones.*

Many verbs, when functioning as light verbs, require the predicate noun in the position of a direct object, and they employ the zero-article, sometimes even as the preferred or only option. This seems logical, since the direct objects of light verbs are semantically not really direct objects but rather parts of the predicate at all. The reason may be their inherent non-referentiality (Cf. Section 5.5 above). Yet many of them are by default telic and combine with volitional and agentive subjects: *take, give*, etc., which suggests that the entire light verb construction is meant to be telic. This is interesting, since it would imply that light verb constructions systematically stride against the Transitivity Hypothesis by combining high Transitivity features with low Transitivity features (Cf. also Section 5.5).

It would be exciting to know how many prototypical light verb constructions comply with the pattern that Lindvall found in her data and how many do not, and, also, their distribution in a balanced corpus. Lindvall's data seems to comprise only narrative fiction, which can generally be expected to contain fewer light verb constructions than e.g. newspaper texts. Would her figures have looked different if she had taken e.g. parallel translations of texts about law, business, or industry?

These answers cannot be obtained from a monolingual corpus such as PAROLE or Konkordanser. They provide no parallel texts in a language that expresses verb aspect (or event structure) as a morphosyntactic category, such that the possible cross-linguistic interplay between the Swedish noun definiteness in direct objects and the verb aspect in the other language as in Lindvall's multilingual parallel corpus. The Swedish-Czech Section of the parallel corpus Intercorp is too small and contains also mostly narrative fiction (Cf. Section 11.1 and Chapter 15). Despite this limitation, the PNL predicate noun lexicon proposed in this thesis (see 16.4) gathers at least the monolingual information from PAROLE as a starting point for further research in discourse-based grammar.

[5]a hypothetical preceding sentence

Part II

Methods and Approaches

Chapter 6

Valency Theory in Functional Generative Description

6.1 Functional Generative Description

The Functional Generative Description (FGD) is a stratificational formal language description based on the Prague functional and structural linguistic tradition. FGD started to develop in the 1960's [146], and it is characterized by the following features:

- inclusion of an underlying syntactic layer (tectogrammatics) into the description of language
- use of dependency syntax
- specification of a formal account of the information structure (topic–focus articulation) of the sentence and its integration into the description

For further reference see e.g. [148], [147], and [54].

6.2 Tectogrammatical Representation

The unique contribution of FGD is the so-called tectogrammatical representation (TR). The tectogrammatical representation, as defined by the Functional Generative Description, is being implemented in a family of tectogrammatically annotated treebanks[1].

The tectogrammatical representation annotation is always built on top of a morphological and a surface-syntax annotation. These annotation layers are separate, but

[1]Czech, English, and Arabic at the moment, see [54], [30], and [150].

interlinked by references going always from a higher layer to a lower layer. Therefore information is neither being duplicated nor gets lost across the respective layers. Fig. 6.1 (taken from [166]) shows the tectogrammatical, surface-syntax, and morphological annotation layers of the Czech sentence *Byl by šel do lesa. (He would have gone to the forest)*. The bottom layer is the linear text layer without any markup except tokenization and added identification codes in the respective tokens.

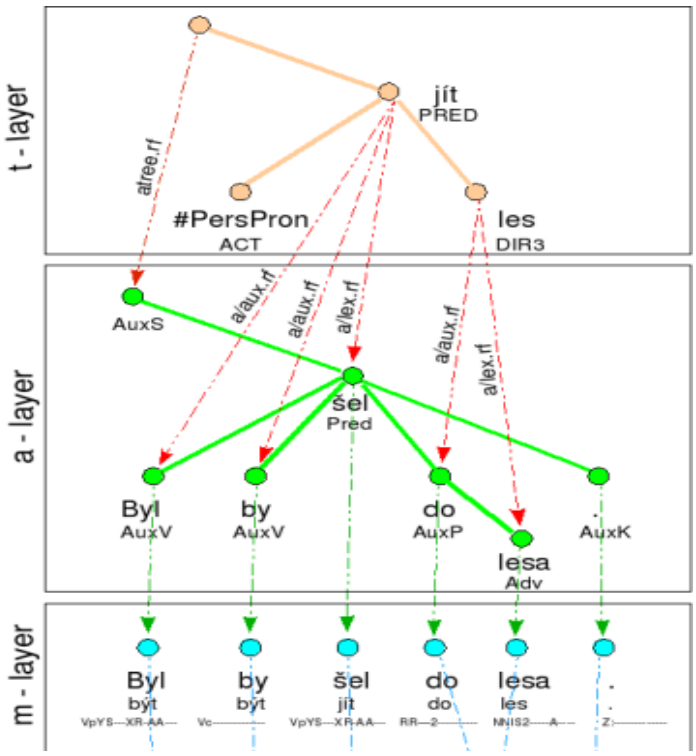


Figure 6.1: The system of annotation layers with references.

Being conceived as an underlying syntactic representation, the TR captures the linguistic meaning of the sentence, which is its basic description unit. The tectogrammatical representation of a sentence is a representation of its grammatical structure going up to a level of abstraction that is delicate enough to capture the essential features of its meaning. In the TR annotation, each sentence is represented as a projective dependency tree. The attribute values include references to the analytical (surface-syntax) layer. Only content words are represented by tectogrammatical nodes. Function words are represented as attribute values. Each tectogrammatical node (t-node) has a semantic label ("functor"), which renders the semantic relation of the given node to its parent node, and each word form represented by a t-node has a **t-lemma**, which is usually the dictionary form of the word. The TR annotation captures the following

aspects of text:

- syntactic and semantic dependencies
- syntactic/lexical derivation (t-lemmas)
- fine-grained morphological information (grammatemes)
- coordination, apposition, parenthesis
- valency
- information structure (topic-focus articulation)
- grammatical and contextual coreference
- ellipsis restoration.

Fig. 6.2 shows a short English sentence from a tectogrammatically annotated Wall Street Journal document. It is a projective dependency tree consisting of edges and labeled nodes. The topmost node of each tectogrammatical tree is the technical root node, which carries only the identification code of the given sentence in the corpus. The entire sentence is linguistically governed by the predicate.

Each node has (in the present visualization) three descriptions. The top line (black) text in each node is its tectogrammatical lemma (t-lemma). The tectogrammatical nodes of 'real' words, i.e. of words that are actually found on the surface, have **t-lemmas**. The t-lemmas are usually like the basic forms (infinitives, nominative singulars, etc.). However, not all tectogrammatical nodes render words that actually occur on the surface. Some tectogrammatical nodes are generated and have no direct correspondence in the surface shape of the sentence. Such nodes have **t-lemma substitutes** instead of t-lemmas. Here it is the node with the t-lemma substitute #Gen and the functor ADDR, which means the Addresse of what was being said.

The second line states the semantic label of the given node. The labels (functors) are printed in capitals. They describe the semantic relation of the given node to its parent node; e.g., *spokewoman* is the Actor of the verb *say* in *spokewoman said.*

The words printed on the third line of each node are (in the present visualization) printed in green and orange. These words represent the values of the the references to the surface-syntax annotation layer (analytical layer), which itself is linked to the morphologically tagged text (morphological layer). In the non-visualized data, this attribute value is not a word but the id-code of the given token in the surface realization of the text. The green/darker words *(said, spokewoman, Lorillard, is, this, old, story)* are the forms of content words as they are realized in the surface shape of the sentence. The orange/light words *(A, an)* represent links to the forms of function words (auxiliary words).

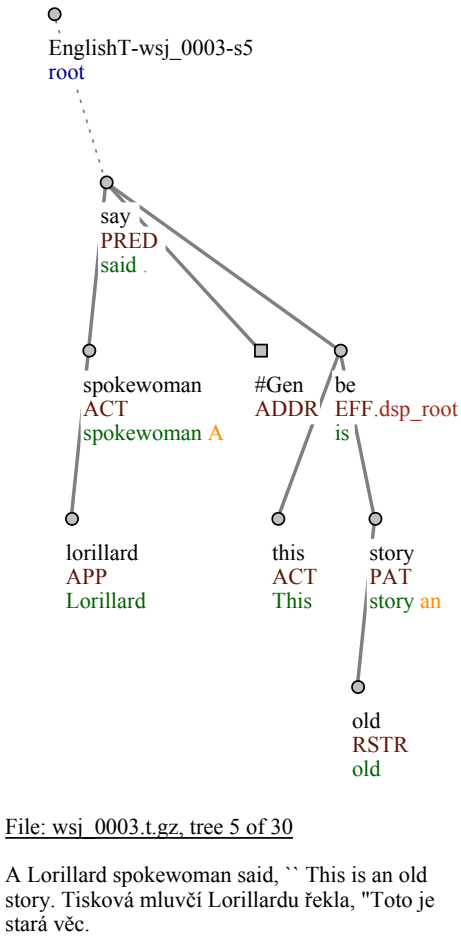


Figure 6.2: Tectogrammatical Representation – an example sentence from the Prague English Dependency Treebank 1.0

Artificially generated nodes, such as the #Gen. ADDR governed by *say*, are usually inserted to complete a valency frame. According to the valency lexicon (see more in the next section), the verb *say* has three obligatory arguments (participants): the Actor (the one who says), the Addressee (to whom), and the Effect (what). The Addressee is not known from the text, but it is generally known that things being said are said to someone. Therefore the Addressee is marked as being there, but 'generalized'. This valency frame also contains a Patient (about what), which is semantically optional, and therefore it is not inserted into the data when not present on the surface. For more detail about the valency principles in FGD see Section 6.3.

6.3 Valency

6.3.1 Basic Notions

The previous Section treated the basic features of the data annotation in FGD. However, the respective treebanks consist of two major parts: not only the data, but also a valency lexicon. In the data, each occurrence of a word that has a valency must be linked to the appropriate valency frame in the lexicon. This Section describes the main principles according to which the FGD-based valency lexicons are created.

The valency theory of FGD starts with the exploration of the valency of verbs, which are regarded as the pivot of the sentence in dependency syntax. The introductory article, which was published in parts and is founding the FGD-based valency theory [116], treated exclusively verbs. Recent research in this field [118], [89], but also [104] and many others in other theoretical frameworks suggests that nouns, adjectives, and adverbs require separate treatment, although many principles can be adopted from the valency of verbs. The FGD-formulation of non-verb valency is still going on, with some preliminary statements [129], [119], [81], and mainly [106] (p. 125–150), which are being verified and refined during the data annotation. The current data-linked Czech valency lexicon PDT-VALLEX [53] comprises some nouns, adjectives, and adverbs.

The original valency theory of FGD [116] analyzes the relations between the verb and its complements. It is mainly inspired by Tesnière's conception of dependency syntax, but it also reflects Fillmore's frame semantics. (A more detailed discussion of the semantic labeling of verb arguments is in [118].) The key issues in FGD valency theory are:

- distinguishing 'inner participants' from free adverbials,
- degrees of obligatoriness (obligatory and non-deletable, obligatory and deletable, optional but determined by the given verb frame, or free modifications,
- determining the label set: how many types of complements should be labeled and with which labels.

Stress is laid on the difference "between the obligatoriness of a participant at the semantic level and the necessity of the presence of an element that realized this participant at the level of surface syntax" ([116], p. 23). An element that is obligatory at the semantic level can be either obligatory at the surface level (e.g. *X meets Y* vs. **X meets.*), or it can be potential (*Don't disturb him, he is reading [something].*), or it is even prohibited to occur at the surface level at all: *He already speaks!* (a toddler can already speak).

6.3.2 Inner Participants and Free Modifications

FGD uses a set of semantic labels (functors) to describe the semantic relation of a given node to its parent node. The set of functors for verb complements distinguishes between **inner participants** and **free modifications**. The free modifications are semantically homogeneous (e.g. denoting temporal relations), they can be repeated (e.g. a parent word can have several modifications of the same sort), they do not occur specifically with just one class of verbs (nouns, adjectives) but can be added almost everywhere, and they are mostly optional. The inner participants, however, cannot be repeated (e.g. a verb cannot have two Actors)[2].

There are several classes of functors:

1. Functors for inner participants
2. Functors for free modifications
3. Functors for effective roots of independent clauses
4. Functors for rhematizers, sentential, linking and modal adverbials
5. Functors for complex lexical units and foreign-language expressions
6. Functors expressing the relations between members of paratactic structures.

The functor definitions as well as examples of their use can be found in [28], such that they will not be described here in detail. Not all classes are relevant for the valency theory. The valency-relevant classes are Functors for inner participants and Functors for free modifications, which appear in the valency lexicons, whereas functors of the other classes do not, except the functors CPHR and DPHR, which denote the nominal components of compound predicates (especially light verb constructions – see Chapter 4) and parts of frozen idioms.

There are six functors for inner participants:

- ACT (Actor)
- PAT (Patient)
- ADDR (Addressee)
- ORIG (Origin)
- EFF (Effect)

[2]Coordinated participants as *Peter and David* are just one Actor in the annotation. Coordination is not repetition.

- MAT (Material) – noun specific!

The set of functors that characterize the free modification is listed below.

1. **Temporal functors**
    - TFHL – for how long?
    - TFRWH – from when?
    - THL – how long?
    - THO – how often?
    - TOWH – to when? (set to which time?)
    - TPAR – in parallel with what?
    - TSIN – since when?
    - TTILL – till when?
    - TWHEN – when?
2. **Locative and directional functors**
    - DIR1 – where from?
    - DIR2 – which way? (*across, through, along*, etc.)
    - DIR3 – where to? which direction?
    - LOC – where? in which location?
3. **Functors for causal relations**
    - AIM – purpose, aim
    - CAUS – cause
    - CNCS – concession
    - COND – condition
    - INTT – intention
4. **Functors for expressing manner and its specific variants**
    - ACMP – accompaniment
    - CPR – comparison
    - CRIT – criterion
    - DIFF – difference

- EXT – extent
- MANN – manner
- MEANS – means, instrument
- REG – regard
- RESL – result
- RESTR – exception, restriction
- BEN – benefactor
- CONTRD – contradiction (*While*...)
- HER – heritage
- SUBS – substitution

Virtually all these functors can be used with nouns. However, there are a few noun-specific functors:

- APP – adjunct referring to the person or thing something or someone belongs to
- AUTH – author, creator, originator of artifacts
- ID – effective root of an identifying expression represented by an identification structure
- MAT – adnominal inner participant referring to the content (material etc.) of something
- RSTR – restrictive attribute and not unambiguously non-restrictive attribute
- DESCR – non-restrictive attribute in postposition

For determining the boundary between the inner participants and the free modifications at the tectogrammatical level it is important to decide for each type of participant whether it can depend on any verb, or whether it combines only with a certain group of verbs, and whether the given type of participant can depend more than once on a single verb token. The inner participants are very often represented by structural (i.e. preposition-less) cases, which is, however, far more evident in Czech than in English or Swedish.

The Actor (Agent) has a special position among the inner participants, as it appears with almost every verb, except for a few verbs denoting natural events such as rain. The group of verbs that the Actor combines with are therefore all non-impersonal verbs. Actor, as well as the other inner participants, cannot be repeated within one single frame.

The free modifications are typically adverbials. Adverbials are mostly possible to combine with all verbs, at least from the syntactic point of view. The cognitive and semantic appropriateness was left aside. Adverbials can, unlike inner participants, repeat within one single frame.

6.3.3 Obligatoriness and Dialog Test

The semantic obligatoriness of a modification is stated by means of the **dialog test**. The dialog test consists of hypothetical questions and answers about the given complements of a verb. Their assessment as appropriate or natural in a regular discourse helps to decide whether or not they are semantically obligatory in the frame. The assumption is that obligatory modifications, when omitted, must be known both to the speaker and to the listener, or at least that the speaker assumes that it is known to the listener. The dialog test is used to determine whether a modification, when not expressed, is regarded by the speaker as known or not.

When the omitted modification is known to the speaker as well as to the listener, a wh-question about that particular modification from the listener must appear to be out of place in the dialog. E.g. if the speaker says: *They have just arrived*, he believes that the destination of the arrivers is known to the listener. If the listener suddenly asked: "Where did they arrive"? , it would be considered an odd question in this dialog. However, if the speaker was able to give a proper answer, it would only mean that he made a wrong assumption about the listener's knowledge or that the information exchange had failed at some point. The speaker answering "I don't know" would be totally out of place at any rate. This outcome of the dialog test states that the information about destination location is semantically obligatory.

Of course, a context that would make all questions appropriate can be invented for almost any case. To make the dialog test work, simple, unmarked everyday situations must be chosen, preferably without consideration of the state-of-the-art telecommunication options such as videocalls, etc. For instance, the obligatoriness of the location with *arrive* can be compromised by assuming the following scenario: Officer Alex knows that his fellow soldiers were sent to a secret destination, but he does not know exactly where, and he just received a message saying that the group has just reached their destination. Officer Bill does not know that Alex does not know the destination, therefore he asks. Alex replies: "I don't know", which does not make Bill wonder, because secret missions are nothing uncommon in the reality he lives in. This would be a 'false negative' result of the dialog test for an obligatory modification.

The current FGD-based Czech valency lexicon distinguishes four types of complementations:

1. **obligatory inner participants**

    - cannot repeat

- are semantically obligatory

2. **optional inner participants**
   - cannot repeat
   - are not semantically obligatory
3. **obligatory free modifications**
   - can repeat
   - are semantically obligatory
4. **optional free modifications**
   - can repeat
   - are not semantically obligatory.

Fig. 6.5 shows a Czech verb entry.

6.3.4 Shifting

A necessary pre-knowledge about the inner participants is that they are deep-syntactic roles, which are not affected by the surface syntax. Therefore *Actor* is not identical with *subject* and *Patient* is not identical with *object*. In Example 72, Peter remains Actor and Mary remains Patient, even if the sentence is passivized as in Example 73:

(72) *Peter is beating Mary.*

(73) *Mary is being beaten by Peter.*

The five labeled inner participants correspond to coarse cognitive roles. The Actor (ACT) is – with a few exceptions – the subject of the active form of the verb. It does not occur with all verbs (e.g. verbs denoting natural events such as raining, dawning etc. do not express the subject in many languages, or use just a formal, dummy subject). Unlike Fillmore's conception, the Actor is not further distinguished as agent, bearer, instrument, etc. In this respect, Panevová's conception corresponds to that of Tesnière, in which the assignment of the label to an inner participant depends on the number of participants of the verb. Actor is simply the first inner participant.

The second participant is called Patient (PAT), and it is usually the direct ('accusative') object. It is also the complement *about what* in verbs of saying. When a verb has an affected as well as an effected object, Patient is always the affected one. However, when a verb has only two inner participants, the second is Patient, even if it is clearly an effective object. Examples: *tell (someone) about something, boil water, look for spectacles, achieve success, forget your name, ask your opinion*.

When a verb has more than two valency slots for inner participants, the labels of the third and higher inner participant are assigned according to its semantics. A verb that has at least three participants can have an Addressee as the third inner participant. Addressee (ADDR) is typically a 'dative' or indirect object. Examples: *tell someone, bring to someone, pay someone, ask someone, teach someone*.

Another inner participant that can occur as the third and higher participant is Origin. Origin (ORIG) is a participant that is rarely obligatory, and it must be understood as a specific type that denotes either the source in a transfer of something from one person to another person *(hand something over from someone to someone else, know something from someone)*, or the source material in a transition *(build from stone, something grew out of something).*

The last inner participant to choose from in verbs with more than 2 slots is Effect (EFF), which denotes result, effected object, or, when the given verb has one object that is primarily animate and one that is primarily inanimate, it is the inanimate one.

When labeling the complements of a given verb with more than two slots, labels are supposed to be assigned to the participants according to their cognitive roles. However, when a verb does not have five slots open for complements in its frame (which is the case in most verbs), then the **shifting** principle is applied. Shifting can be understood as "one aspect of the relationship between linguistic meaning and cognitive content. It can be said that the 'unshifted' units correspond rather to the cognitive or ontological content, while individual languages 'shift' them according to relevant conditions; i.e. every language classifies them with regard to its structure, so that at the level of linguistic meaning there appear 'shifted' (already classified) participants" ([117], part I, p. 29).

In sum: due to shifting, if a verb has only one inner participant, the participant is always labeled as ACT, no matter whether it is agentive or not; e.g.: *a book*.ACT *appeared*. When the verb has only two inner participants, the second participant is always PAT, although semantically it might happen to correspond to ADDR or EFF: *address someone*.PAT, *dig a hole*.PAT. The same approach is applied to verbs with three and four participants. Panevová [117] says: " If, in the cognitive stratum, an action has not among its elements an item that could be the base of Ag [later called ACT, S.C.] and/or PAT (for the tectogrammatical level), then the "free position" is filled, in its frame, according to the arrows from Fig. 6.3[3]. In case of a possible choice, the position of PAT is occupied by what otherwise (with a verb having a larger number of participants) would function as EFF, while ADDR and ORIG remain unshifted" ([117], part I, p. 29).

[3]The figure has been copied from [117].

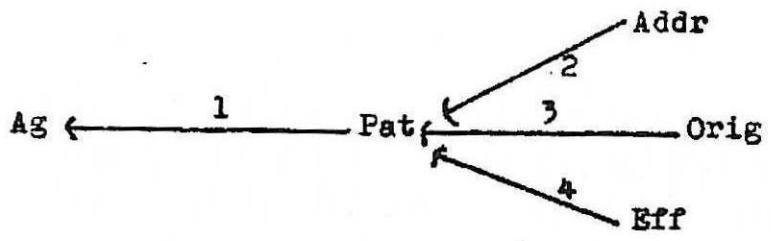


Figure 6.3: Shifting

6.3.5 Quasi-Valency Complements

The recent development of the valency theory of FGD [89] introduced a category of a few quasi-valency complements, which were earlier regarded as free modifications:

- **Obstacle** (OBST)

  (74) *Sleeping Beauty pricked herself on a thorn.*

- **Mediator** (MED)

  (75) *John brought the dog by its collar.*

- **Difference** (DIFF)

  (76) *Move two steps higher.*

  (77) *Our team won by two goals.*

- **Intent** (INTT)

  (78) *John went swimming.*

  (79) *John stayed for lunch.*

OBST and MED were distinguished among the MEANS adverbials as two separate groups during the annotation of the Prague Dependency Treebank [54]. In the Czech data they are clearly identified by the preposition by which they are governed. The labels DIFF and INTT were part of the label set from the initial stages of the PEDT 2.0 annotation, but large-scale annotation casts a different light onto their syntactic behaviour. The features that these free complements share with the inner participants are:

1. they are governed (their morphemic shape is determined) by their verbal heads

2. they occur with a limited class of verbs

3. they cannot repeat.

However, they also have the typical features of free modifications:

1. they are semantically homogeneous

2. they do not underlie the 'shifting'

3. they are mostly optional.

These features can also be observed by ADDR and ORIG. The authors suggest that ADDR and ORIG could also be moved to the new group of quasi-valency complements. The final decision whether ADDR and ORIG also belong to quasi-valency complements was yet not made.

6.3.6 Discussion

In a valency lexicon, each verb is divided into lexical units (LU's) (originally called frames). Originally, it was the syntactic pattern rather than a possible difference in meaning (conditioned e.g. by different collocates) that defined an LU. Panevová herself says: "More than one meaning is distinguished only if this distinction is made necessary by a difference in verbal frames; from the lexicographical point of view, this distinction is not made in a systematic way" ([117], part II, p. 17). This means that two or more different readings of a verb often remained merged, as far as their valency frame was identical. In the current routine lexicographic work on the Czech Vallex lexicon, figurative meanings are mostly separated and provided with different frames [106]. For instance, when one of the arguments of the more concrete reading is an obligatory free modification (e.g. *depart from a place*.DIR1), the figurative meaning is mostly rendered by a separate frame with a different functor (*depart from a fact*.PAT).

The original FGD valency theory is primarily based on syntactic criteria. In case of conflict, additional semantic criteria are employed, such as animacy of the second and third participant. However, the predefined semantic criteria do not quite hold when confronted with a large amount of data. Animacy would be, for instance, the main criterion in assigning the functors ACT ADDR PAT (shifting) versus ACT PAT EFF (the more semantic consideration) in a verb that requires two objects, out of which one is not a typical 'dative' object, but at the same time it is not really a result of anything – simply another object without any semantic description. Then, according to the original theory, the frame ACT ADDR PAT would be preferred with the second object being primarily animate, while the frame ACT PAT EFF would be preferred with both objects inanimate. This would result in intuitively inappropriate splittings in verbs

whose readings are not affected by the animacy of their collocates. This principle was difficult to follow in practice.

This conflict can be illustrated by a comparison of two versions of the frame entry for *spojit (connect/ link/ associate/ unite / unify)* in the Czech valency lexicon Vallex (Version 1.0 and Version 2.5). The old version (Fig. 6.4) follows the original theory, the new one (Fig. 6.5) does not. Since the second version (Vallex 2.0), the animacy criterion has not been followed, and the resulting entry has become both simpler and lexicographically more adequate.

spojit **pf.**

1 spojit$_1$ ≈ **zkombinovat; sloučit**
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$ $\mathbf{EFF}_{s+7}^{obl}$ $\mathrm{MANN}_{pod+4,pod+7}^{typ}$
-example: spojit procházku s nákupem; projekt spojil dohromady muzeum s divadlem
-asp.counterpart: spojovat$_1$ impf.
-class: combining

2 spojit$_2$ ≈ **sjednotit**
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_4^{obl}$ $\mathbf{EFF}_{do+2,v+4}^{obl}$ $\mathrm{BEN}_{proti+3}^{typ}$
-example: spojil obyvatele do sdružení
-asp.counterpart: spojovat$_2$ impf.
-class: change

3 spojit$_3$ ≈ **ztotožnit**
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{PAT}_{4,že}^{obl}$ $\mathbf{EFF}_{s+7}^{obl}$
-example: spojili tuto ideu s vrcholem filozofie; spojil (si) Petra s průšvihy
-asp.counterpart: spojovat$_3$ impf.
-class: combining

4 spojit$_4$ ≈ **přepojit telefonicky**
-frame: $\mathbf{ACT}_1^{obl}$ $\mathbf{ADDR}_4^{obl}$ $\mathbf{PAT}_{s+7}^{obl}$ MANN^{typ}
-example: spojit telefonicky Petra s Pavlem
-asp.counterpart: spojovat$_4$ impf.
-class: social interaction

Figure 6.4: The original version of the first 4 readings of the entry *spojit* (Vallex 1.0)

Fig. 6.4 shows the original entry of the Czech verb *spojit (connect/ link/ associate/ unite / unify)* in Vallex 1.0 ([91] and [167]). Valency frame 4 *(establish a telephonic connection)* is the only one of the sample that is primarily designed as a frame for two animate objects. The second example contains two grammatically inanimate objects, which are, nevertheless, cognitively animate (two offices were connected = two people from the respective offices got the chance to have a conversation on the phone)). In all

the preceding readings, a mixture of animate and inanimate objects or two inanimate objects are assumed, although the verb *spojit* in all the three readings can be easily used with two animate objects. Readings 1 and 2 even appear to merge when used with animate objects:

(80) *Svatba Josefa a Karolíny spojila Blažíčkovy a Vomáčkovy.*
The marriage of Josef and Karolína unified the families Blažíček and Vomáčka.

(81) *Pravicové i levicové politiky spojuje touha po moci.*
What the right-oriented politicians have in common with the left wing is their desire for power.

The most recent release of Vallex (Vallex 2.5, [94]) unifies the functor assignment into ACT ADDR PAT for the first three readings from Vallex 1.0 (Fig. 6.5).

spojovat^{impf}, spojit^{pf}

1 ≈ **impf: kombinovat; slučovat; sjednocovat** **pf: zkombinovat; sloučit; sjednotit**
-frame: **ACT**$_1^{obl}$ **ADDR**$_{s+7}^{obl}$ **PAT**$_4^{obl}$ **EFF**$_{do+2,v+4}^{opt}$
-example: impf: spojovat obyvatele do sdružení; spojovali se proti rodičům; spojovat procházky s nákupy; spojovat dvě lana uzlem pf: spojil obyvatele do sdružení; spojit procházku s nákupem; projekt spojil dohromady muzeum s divadlem
-rfl: pass: impf: lana se spojovala uzlem pf: lana se spojila uzlem
-rcp: ADDR-PAT: impf: spojovat spolu provaz a lano pf: spojit spolu obchody a galerii v jeden celek
-class: combining

2 ≈ **impf: ztotožňovat; být společným rysem; mít společný rys** **pf: ztotožnit; být společným rysem; mít společný rys**
-frame: **ACT**$_1^{obl}$ **ADDR**$_{s+7}^{obl}$ **PAT**$_4^{obl}$
-example: impf: byli zvyklí spojovat tuto ideu s vrcholem filozofie, spojoval (si) Petra s průšvihy; tyto dva romány spojuje láska autorů k dávné vlasti pf: spojili tuto ideu s vrcholem filozofie, spojil (si) Petra s průšvihy; tyto dva romány spojila láska autorů k dávné vlasti
-rfl: pass: impf: tato myšlenka se spojuje s jeho jménem pf: tato myšlenka se spojila s jeho jménem
-class: combining

3 ≈ **impf: přepojovat telefonicky** **pf: přepojit telefonicky** (idiom)
-frame: **ACT**$_1^{obl}$ **ADDR**$_{s+7}^{obl}$ **PAT**$_4^{obl}$ MANN^{typ}
-example: impf: spojovat telefonicky Petra s Pavlem pf: spojit telefonicky Petra s Pavlem
-rfl: pass: impf: právě se telefonicky spojují jejich kanceláře pf: kanceláře se telefonicky spojí
-rcp: ADDR-PAT:
-class: social interaction

Figure 6.5: The new version of the entry *spojit*- the first 3 readings in Vallex 2.5

The animacy criterion appears to work well for distinguishing between ADDR and DIR3 or ORIG and DIR1, as well as with a few verbs that denote exclusively human transactions, e.g. *lend* and *buy*, in which any grammatically inanimate objects that can be explained with metonymy (e.g. people-organizations). However, in many verbs the

borderline between readings does not go in parallel with the animacy/inanimacy of the inner participants.

6.3.7 Noun Valency in FGD

When characterizing the valency behaviour of nouns, Panevová [119] starts from the fact that nouns never require their inner participants to be expressed in the surface shape of the sentence. The introspective dialog test is the dominant decision basis for defining the valency frame (see Section 6.3.3). The nouns are divided into 4 basic groups:

1. syntactic derivates
2. lexical derivates
3. event nouns
4. 'nouny' nouns

The syntactic derivates of verbs (nominalizations with the suffixes *-ní* and *-tí* in Czech, present participles in e.g. English and Swedish) often inherit the frames of the corresponding verbs. A large part of Panevová's paper [119] deals with the possible morphological realizations of nominalization participants in Czech. These Czech-specific details are perhaps irrelevant for languages with very reduced inflection such as Swedish and English, but the general legacy of this study is to be kept in mind: "as the nominalization, in contrast with the finite verb, always results in reduction (undoubtedly a reduction of morphological categories), also a reduction of argument slots can be expected."[4] And, as Panevová already noted earlier ([116], p. 17), about repeating free modifications, the semantically-based valency pattern provides more slots than can be realized on the surface. The valency frame lists possible slots in a linear way, although it is unlikely that all of them or any combination of them can occur on the surface simultaneously. The ability of a given noun to take the given types of complements "is connected with the recursive properties of the language as a whole and with its potential infiniteness, which contrasts with the restricted and finite character of performance."

The complicated system of case transformations in Czech nominalizations described in [119] suggests that any lexicographical processing of nominalizations must pay attention to the commonly accepted surface representations, as to which participants may appear together on the surface. In other words: the semantic reasoning is

[4]"Vzhledem k tomu, že nominalizace oproti konstrukci s verbem finitem vždy znamená redukci (nesporná je redukce morfologických kategorií), lze předpokládat, že půjde i o redukci počtu argumentů (valenčních míst)." Translation S.C.

inevitably based on introspection, but especially with regard to noun processing, the observations of the actual surface 'filling' of the respective inner participants, as well as their combinations, must be corpus-based, and alternatives should be provided with relative frequency counts (Cf. Chapter 7.).

The lexical derivates are mainly derivates from verbs that are built by means of a productive suffix, (e.g *-er, -ence/-ance, -ee,* etc. in English). They do not necessarily denote events, and some types have one built-in participant (e.g. ACT in *teacher*). Many lexical derivates are names of artifacts, which are actually regarded as having a built-in Effect. This built-in participant is not reflected in the current FGD-based annotation. Interestingly, however, the recent release of NomBank [104] even includes it into the entry with a special markup. The event nouns, which are often a transition between the syntactic and the semantic derivation, partly inherit the frames of their corresponding verbs, but deviations are possible.

The current FGD basically recognizes the conclusions of Piťha [129]. There are a few noun-specific complements:

- **MAT (Material)**
  This is an inner participant that is obligatory with nouns that explicitly denote measures or parts of a whole, such as *part, end, half.* It is an optional participant in nouns that often denote measures/parts of a whole *(cup, basket)*.

- **APP (Appurtenance)**
  This is an obligatory free modification in nouns that denote relations; e.g. *brother, feature*.

- **AUTH (Author)**
  This is a free modification of artifacts; e.g. *Zipf's law, Tolstoi's novels*

- **ID (Identity)**
  This is a free modification of artifacts (names of artifacts, trade marks, series numbers); ; e.g. *Kent cigarettes, the comedy 'Guess Who's Coming To Dinner?', Opel 307*

The noun participants are very often realized by anaphorical pronouns or indirectly in the broader context. In addition to the noun-specific complements, nouns can also take the same complements as verbs.

In the current data annotation, nouns have their frames in the lexicon, but the missing participants are not inserted into the data according to the lexicon.

6.3.8 FGD-valency for Learners of Basic Verbs

Considering the FGD valency theory in connection with basic verbs, it appears that the main benefit of FGD is the distinction between semantic and syntactic obliga-

toriness. This feature is important both for human users and computers in NLP applications, since it specifies which type of information is to be inferred or looked for in the context.

On the other hand, the current FGD-based lexicons give very sparse information on which complements of the respective verbs tend to be explicitly realized on the surface and which tend to be elided or implicitly present in the context or regarded as common knowledge. The Czech valency lexicons list possible morphosyntactic representations of each valency complement, but they do not provide explicit information on their frequency. In noun entries the lexicons do not specify which slots and which surface forms can and which can not coexist simultaneously.

The semantic labeling of the inner participants is rather coarse and, in addition, the cognitive roles the participants might have are obscured by the shifting, and thus not actually helpful when the learner learns by associating morphosyntactic forms to cognitive roles.

Finally, the FGD-based lexicon gives hardly any hints regarding collocations, apart from very lexicalized idioms. At least basic verbs, whose meaning in context so often depends on the collocates, deserve more explicit information about the meaning shifts of identical syntactic realizations in different contexts (Cf. Chapter 7).

Chapter 7

Corpus Pattern Analysis

"Lexicons of the future will be application-driven, and will pay much more detailed attention to the connection between meaning and use. To do this, they will focus on determining the probabilities, and associating them with prototypical contexts, rather than seeking to cover all possible meanings and all possible uses."

Patrick Hanks

7.1 Theory of Norms and Exploitations (TNE)

Hanks [57] proposes Corpus Pattern Analysis (CPA) as a standard for building verb entries. As this standard has been adopted in Swe-VALLEX, a short summary of CPA will be provided.

CPA is the lexicographical implementation of the **Theory of Norms and Exploitations (TNE)**. It focuses on relating syntactic patterns to selectional preferences[1]. In case of verbs, it seeks to list all usual syntactic complementation types of a given verb (obtained by querying a large corpus) and to group them according to selectional preferences. Each group of collocates referring to the same selectional preference forms a *lexical set*. Combinations of lexical sets repeatedly co-occurring in the same syntactic patterns correspond to what is commonly called *lexical units* or *readings*, *meanings* or *senses*. TNE calls them *meaning potentials*. The core idea of TNE is that "a word in isolation, strictly speaking, does not have a meaning; it only has a meaning potential, or rather a cluster of meaning potentials, any one or more of which, may be activated by the context when a word is used in context" [57].

[1]Emphasis has been laid on the difference between *preferences* and *restrictions*: "A restriction prevents or forbids you from doing something, whereas it is often the case that locutions excluded by a selectional preference are nevertheless perfectly grammatical, psychologically acceptable, and communicatively adequate" [57].

7.2 Defining Lexical Sets

A lexical set is a set of lexical items (typically nouns) that occur in the same valency slot or argument position in relation to a given collocate. A lexical set, therefore, comprises both norm-compliant uses and exploitations of the norm. The 'established norm' cannot be rigorously defined. It is nothing but an empirical value. As Hanks puts it: "…because of the flexible, variable nature of the lexicon, even attempting a full and accurate description of the norm for any given usage may be impossible, for principled reasons: how can one define a phenomenon whose boundaries are shifting and variable?" Non-conventional verb uses are supposed to be metaphorical exploitations of the norms. Determining a lexical set is a performance of linguistic intuition, backed by plentiful lexical evidence.

The selectional preferences that define a lexical set are expressed in the form of non-formalized semantic labels (*mnemonics*). The mnemonics must apply the appropriate degree of generalization to be able to distinguish the given meaning potential from others. Hanks discusses the example of the label *Human*, which is notoriously known from numerous attempts on semantic feature analysis. The label *Human* includes some features humans share with animals (sleeping, eating, fearing, etc.) and others shared by institutions, nations, even computers (cognitive activities). The verb in question may activate only one of them. In such case the label *Human* would be under-specifying. The opposite extreme would be labeling the lexical sets in a way specific to the verb in question. Lexical sets do not have the same function as frame elements of FrameNet (see Section 9.6). Hanks's solution of the semantic labeling in his own proposed 'pattern dictionary' is discussed below.

Lexical sets, too, should only reflect norms, not exploitations. To use one of Hanks's examples again: one meaning potential of the English verb *to urge* is mainly associated with horses: *to urge a horse up the path* etc. Finding evidence of e.g. camels being urged somewhere can result in labeling the lexical set with *Steed* to emphasize the importance of being ridden to the exact information on animal species. However, a sentence in which the driver urges his car somewhere, will be regarded as exploitation of the horse-ride shaped sentences. The car has adopted the steed-feature through a metaphorical transfer. The sentence acquires then the additional meaning of the driver imposing his will on his car, like a rider imposes his will on the horse, which apparently was the communicational aim of selecting the verb *to urge* in connection with a car. On the other hand, an exploitation can be allowed to develop into a new norm, as Hanks's 3rd meaning potential of *to urge* illustrates: *urging practitioners towards greater involvement* has the same syntactic pattern as *to urge a horse somewhere.* It is now quite normal to urge people in a particular direction; much of the metaphor has already got lost by the conventionalization. Nevertheless, both complementations are regularly populated with different lexical sets. Whereas the first pattern included steeds as direct objects and adverbials of (spatial) direction, this pattern is typically

populated with persons and institutions as direct objects and with adverbials of (intentional) direction.

In a pattern dictionary proposed by Hanks and Pustejovsky [59], two different types of labels can be used to define a lexical set. The first label, called **semantic type**, belongs to an – almost – closed set of the Brandeis Semantic Ontology [136]. It renders an intrinsic attribute of the given noun at a rather general level, such as: [[Human]], [[Institution]], [[Location]], [[Body Part]], [[Vehicle]]. The semantic types can alternate. Alternations are regular choices of types within an overall pattern; e.g.: *[[Human|Institution]] negotiate…* [58]. The current version of the Pattern Dictionary of English verbs, however, uses its own ontology.

The second type of label is **semantic role**. The semantic role renders the attribute assigned to the noun by the selection of a given verb in that particular context; i.e. 'what implicatures does the noun obtain by being combined just with that particular verb'. For instance, the slots opened by the verb *sentence* are both intrinsically [[Human]], but with this particular verb, the subject [[Human]] is a *Judge*, whereas the object [[Human]] is a *Convict*. *Judge* and *Convict* are semantic roles. The set of semantic roles is not closed.

More verbs can happen to have identically labelled lexical sets. However, this does not necessarily imply that the respective lexical sets are populated by the same nouns. Hanks and Jezek [58] say that lexical sets are 'shimmering'. This is to say that "the membership of the lexical set changes from verb to verb: some words drop out while other come in, just as predicated by Wittgenstein (family resemblances). Different verbs select different prototypical members of a semantic type even if the rest of the set remains the same."

Hanks and Jezek illustrate this phenomenon on *wash* and *amputate*. Both typically select [[Body Part]] as their direct object. One can wash any body part, but the typical collocates of *wash* are *face|hands|hair*, which one does not have amputated (at least face and hair impossibly). In such cases, not even semantic roles can help to make the lexical set more specific.

This gives an implication for building ontologies: as the lexical sets "shimmer according to what we predicate of them", and therefore "a node in the ontology (i.e. a semantic type) is not to be thought of as an address for 'all and only' the lexical items that belong to that node. Rather, it is an address for lexical items that typically belong to that node. The ontology is thus best conceived, not as a rigid yes/no structure, but as a statistically based structure of shimmering lexical sets" [58].

On the contrast of *to urge* to its near synonym *to incite* Hanks shows the necessity to consider the lexical sets in relation to the Good-Bad axis. *Incite*, unlike *urge*, typically occurs with bad things as direct objects. When occurring with a neutral expression like *John incited Barry to speak to Astrid* (Hanks's example, too), the default interpretation must be that Barry's speaking to Astrid (or speaking to Astrid in general) was something bad. This, as well as other implications and presuppositions is

part of the information which a lexicon is supposed to mediate to the user, as far as this information is possible to retrieve without the risk of over-interpretation, against which Hanks himself [57] warns: "It may therefore be preferable to approach teasing out such implications as a matter of identifying mutual beliefs by the traditional techniques of introspection and comparison of intuition, rather than through computational analysis of texts".

7.3 Applying CPA

The objectives of CPA, in the first instance, is to compile a pattern dictionary [59] of the content words of a language – or at any rate, the verbs. A pattern dictionary differs from a pattern grammar (e.g. [68]) in its level of generalization. A pattern dictionary associates one or more patterns with each word: a pattern grammar seeks a more abstract level of generalization.

Hanks suggests the following steps in determining the patterns of a verb [57]:

- Use some statistical procedure [...] to identify statistically significant collocates of the target word, sorted as far as possible by clause roles.
- Sort this first list of collocates into relevant sets for purposes of meaning discrimination; devise approximate intensional criteria for set membership.
- Give each set a name (coined ad-hoc as a mnemonic, and bringing with it no theoretical baggage).
- Sort more concordance lines into groups, according to the intensional criteria just mentioned; extend the sets; refine the intensional criteria; refine the lexical-set mnemonics. Repeat indefinitely as new data becomes available.
- Note correlations among different sets in particular clause roles, with a view to specifying the meaning potentials of the target word.
- (Optionally) add a numeric value such as the number of occurrences or (better) the number of different texts in which each set of pair or group of sets is found (Cf. e.g. [50]).
- Explain the relation of any ad-hoc set members to bona-fide set members by appealing to criteria of ellipsis, stylistics, rhetoric, metaphor, etc.

This set of instructions has been applied in Swe-VALLEX. The application of CPA on the Swedish data is discussed in more detail in Section 16.3.

Chapter 8

Lexical Functions

8.1 Brief Overview

The lexicon proposed by this study was significantly inspired by already existing collocational dictionaries that have paid systematic attention to LVCs, namely *The BBI Combinatory Dictionary of English Word Combinations*. [108], as well as *Tolkovo-kombinatornyj slovar sovremennovo russkovo jazyka* [69] and *Dictionnaire explicatif et combinatoire du français contemporain* [100], out of which the latter two are modeling 'institutionalized' lexical relations by the so-called Lexical Functions. A brief overview of these dictionaries has been given in Section 9.

Lexical Functions are part of the Meaning-Text-Theory developed by Igor Mel'čuk and his collaborators [99], [77]. There are two elementary types of LFs – paradigmatic and syntagmatic – and this study concerns only the latter. In terms of collocations, when two lexical units are collocates, one is usually the base that "selects" the other lexical unit to render a certain meaning together. MTT aims to capture it by the mathematical functional notation: LFi(X) = Y, where X is the keyword (the collocational base) and Y the value of the LFi (the collocate). LFs can assign one value or a set of values to a given keyword. The values stand in the same lexical relation to the keyword but they are not necessarily synonymous. The LFs describe the semantic relation between the keyword and the values. For further reference see [163].

As the data of PNL presents numerous instances of LF-combinations, most illustrative examples in this chapter will be English and directly taken from [163].

8.2 Lexical Functions and FGD

Lexical Functions are closely associated with valency. For light verbs and predicate nouns it is crucial to determine the participants in the light-verb frame as well as in the event that is denoted by the predicate noun. Often, but not necessarily, there is

a corresponding lexical verb from which the noun had been derived (or vice versa), e.g. *ledning*: *leda*, *fråga*: *fråga*, etc., and it just inherits the frame of the verb. Othervise, the nominal event would have to be paraphrased with a verb that has no word-formation relation to the given noun. Mel'čuk calls the nominal event in a LVC "the underlying expression," which specifies the corresponding situation 'L': a full verbal form meaning 'L' ([101], p. 61).

LFs Oper and Labor use numerical indexes to indicate which of the participants of the nominal event ("the underlying expression") is identical with the subject of the light verb. Meaning – Text Theory refers to participants as DSyntAs and labels them with numbers. As we are working within the FGD framework, we will have to relate what LFs use from MTT's valency theory to the valency theory of FGD.

In terms of FGD, the entire LVC would be drawn as a light-verb deep valency frame, in which the predicate noun is the deep frame slot with the deep frame slot filler CPHR (or sometimes DPHR). The predicate noun itself is a deep frame evoker. It has again its own deep frame slots whose fillers denote the cognitive roles of the participants. Now, one of the deep frame slots belonging to the predicate noun is in grammatical coreference with the Actor of the light verb. Its surface frame slot filler will get the substitutional t-lemma QCor (quasi-coreference).

The numerical index of the LF refers to the number of the particular deep frame slot of the predicate noun whose surface frame slot filler has got the substitutional t-lemma QCor. Deep frame slot fillers, unlike DSyntAs in MTT, are not labeled by numbers but they are assigned functors. Presumably, only functors of inner participants and no functors of free modifications will participate in describing LVCs. To be fully just to FGD, the LFs would actually have to carry functor names instead of numerical indexes when used within the FGD framework. However, the functors of inner participants can be easily displayed as numbers: ACT corresponds to 1 and PAT corresponds to 2. ADDR, ORIG and EFF correspond to 3, 4 and 5 respectively. As for the principle of shifting, any ADDR, ORIG or EFF 'shifted' to PAT is regarded as PAT and is to be assigned the label 2.

8.3 Basic Lexical Functions in LVC Description

Lexical Functions are, according to Mel'čuk ([101], p. 60), characterized by the following relation between the verb and the noun: "The support verbs serve to link, on the DSynt-level, (the name of) a DSyntA of L to L itself; they thus play an important semantico-syntactic role and can be loosely called 'semi-auxiliaries' ". The verbs "play an important communicative role: they are used to express the communicative organization, or perspective, of the sentence. [...] Therefore, although they are semantically empty [...], they are by no means asemantic. In addition, they carry all grammatical verbal caegories which must be expressed in a sentence (mood, tense,

person and number, etc.)."

The following LFs are specific to LVCs; their keywords are the predicate nouns and their values are by definition verbs: Oper_i, $\text{Labor}_{i,j}$ and Func_i. For the purpose of PNL, Copul was added to be observed as well, though it does not belong to the LVC-describing Lexical Functions.

8.3.1 Oper_i

The Actor (and subject) of the light verb stands in grammatical coreference with the *i*-th inner participant of L, i.e. of the predicate noun in question. The predicate noun is direct object or indirect object (if the verb cannot have a direct object) or the strongest prepositional object of the light verb (if the verb cannot take non-prepositional objects). In FGD, it gets the functor CPHR, which in this particular constellation substitutes for the Patient, which is nr. 2 in the hiearchy of FGD-inner participants. This corresponds to the original definition of MTT, saying that "The DSyntA I of this verb (and its SSynt-Subject) is the phrase that is described in the GP [*government pattern,* S.C.] of L as the *i*-th DSyntA of L, and Oper_i's DSyntA II (= its main SSynt-Object) is L itself."

Typically, Oper_i comes realized as Oper_1 or Oper_2. Constructions like *to give an order*, *to lend support* make good Oper_1 examples. Having a closer look at *to give an order* from the perspective of FGD, the construcion has the following features:

- The construction can be paraphrased as *X gives Y X's order to do z*. The Actor (at the same time subject on a-level) of the verb corefers with the Actor (at the same time subject on the a-level) of the event described by the noun *order*. (Let us assume the following deep valency frame in *order*: X's.ACT *order* to/for Y.ADDR to do z.PAT.

- The noun *order* is direct object of the verb *to give*. It has the functor CPHR – which has substituted PAT, the functor the noun would have got if it hadn't been a predicate noun in an LVC.

Oper_2 differs from Oper_1 in that the Actor (and subject on the a-level) of the verb corefers with the Patient (and object on the a-level) of the event denoted by the noun, e.g. in *to come under X's control*. Let us assume the following deep valency frame in the noun *control*: X's.ACT *control* over Y.PAT. The Actor (and subject on the a-level) of the verb *to come* corefers with the Patient (and object on the a-level) of the event denoted by the noun *control*. The construction can be paraphrased as *Y comes under X's control over Y*.

Though Mel'čuk [101] only mentions indexes 1 and 2 in connection with the Oper-LF, PNL introduces also Oper_3. Oper_3 is used for LVCs whose predicate noun's deep valency frame consists of ACT, ADDR and PAT. This would be e.g. the case of

to receive an order. *Order* will have the following deep valency frame: X's.ACT order to Y.ADDR to do z.PAT. The entire LVC *to receive an order* could be paraphrased in this way: Y receives X's order to Y to do Z. The Actor (and subject on the a-level) of the verb *to receive* corefers with the Addressee (and object on the a-level) of the event denoted by the noun *order*.

So far, neither Oper_4 nor Oper_5 have occurred in PNL, but they are formally possible.

8.3.2 $\text{Labor}_{i,j}$

In Labor, the predicate noun is the prepositional object of a transitive light verb. The Actor (and subject on the a-level) of the light verb stands in grammatical coreference with the *i*-th participant of the event L denoted by the predicate noun. The direct object of the light verb stands in grammatical coreference with the *j*-th participant of the event L denoted by the predicate noun. This FGD-adopted definition originates in [101] saying that with $\text{Labor}_{i,j}$ "The DSyntA I f this verb (and its SSynt-Subject) is the *i*-th DSyntA of L, its DSyntA II (= its main SSynt-Object) is the *j*-th DSyntA of L, and its further DSyntA (= its second or third SSynt-Object) is L itself."

If *interrogation* combines the verb *to subject* as predicate noun, they belong to the $\text{Labor}_{1,2}$ LF as keyword and value. *Interrogation* is the prepositional object of *to subject*. In FGD, the light verb frame of the verb to subject would consist of ACT, PAT and CPHR. The event L denoted by the noun *interrogation* would have the following participants: X's.ACT interrogation of Y.PAT. The LF $\text{Labor}_{1,2}$ can be paraphrased as *X subjected Y to X's interrogation of Y*. The Actor (and subject on the a-level) of the light verb *to subject* stands in grammatical coreference with the Actor (and subject on the a-level) of the event L denoted by the predicate noun *interrogation* and, at the same time, the Patient (and direct object on the a-level) of the light verb *to subject* stands in grammatical coreference with the Patient of the event L denoted by the predicate noun *interrogation*.

8.3.3 Func_i

With the LF Func_i the predicate noun is the Actor (and subject on the a-level) of the light verb it joins in a LVC. The index refers to the number of the participant in the deep valency frame of the event L that is denoted by the predicate noun. in MTT's terms, "The DSyntA I of this verb (and its SSynt-Subject) is L itself, and its DSyntA II (= its main SSynt-Object) is the *i*-th DSyntA of L" ([101], p. 61), e.g. the event L denoted by the noun *blow* would have two participants, possibly an Actor (X's blow, blow from X) and a Patient (blow to Y). e.g. the sentence *The blow comes from X.* could be paraphrased as X's.ACT blow to Y.PAT comes from X.DIR1. It is the Actor of the event L denoted by the predicate noun *blow* that corefers with a member of the

light verb's frame. Therefore the relation will get the index 1. In *The blow falls upon Y* it is the Patient of the event L denoted by the predicate noun *blow* that corefers with a member of the light verb's frame, and therefore this relation will be classified as Func_2.

The MTT's convention, which was fully adopted by PNL, says that the Func-index is 0 when the predicate noun joins an intransitive verb, as in Func_0(*the war*) =*is on*, or an intransitive LU of a verb, when a verb has intransitive as well as transitive LUs.

Unlike MTT, FGD does not regard the relation between a verb and a noun collocate as a LVC and the keyword of the Lexical Function Func keeps the functor ACT, not replacing it by CPHR.

8.3.4 Copul

The LF Copul captures verbs that acquire the meaning of a copula. In the Swedish data it is typically constructions as *ligga sjuk, sitta modell, gå vakt* etc. Though MTT does not count such constructions as LVCs, and neither does FGD mark the nouns with the functor CPHR, PNL observes them together with LVCs, loosely associating them to LVCs.

8.3.5 Cross-linguistic Comparison of the Predicate Noun *disposal* in Swedish vs. in Czech

1. *ställa något till förfogande = dát něco k dispozici*
   give something at someone's disposal

2. *få/ha något till förfogande = dostat-mít něco k dispozici*
   get/have something at one's disposal

3. *något står till förfogande = něco je k dispozici*
   something is at someone's disposal

In both languages the deep valency frame of the noun is identical:

- *Y's.ACT förfogande över z.PAT = Y-ova.ACT dispozice se zPAT*

The first sentence is to be (language-specifically) paraphrased as follows:

- *X ställer z till Ys förfogande över z för Y.*[1]

[1] *för Y* is part of the verbal deep valency frame and is to be assigned the functor ADDR. It corresponds to Czech dative object.

- *X dává Y-ovi z k Y-ově dispozici se z.*

In both languages this LVC would be classified as CausLabor$_{2,3}$ (more about Caus see below, Section 8.5). The sentences can be paraphrased as *X causes that Y has z at Y's disposal*.

The second sentence would be classified as Labor$_{1,2}$. The sentences can be paraphrased as

- *Y.ACT får/har z.PAT till Y's.ACT förfogande över z.PAT*
- Y.*ACT dostane/má z.PAT k Y-ově.ACT dispozici se z.PAT.*

The respective verbal Actors stand in grammatical coreference with the respective nominal Actors and the respective verbal Patients stand in grammatical coreference with the respective nominal Patients.[2]

The third sentence is to be classified as Oper$_2$ LF. The sentences can be paraphrased as

- *z.ACT står till Ys.ACT förfogande över z.PAT*
- *z.ACT je Y-ovi.ADDR k Y-ově dispozici se z.PAT.*

The Actors of the respective light verbs stand in grammatical coreference with the Patients of the respective predicate nouns.

8.4 Phasal Lexical Functions

There are three supplementary Lexical Functions that can specify the basic Lexical Functions Oper, Labor, Func and Copul:

[2]Note that Swedish does not allow for ADDR as direct object (or at least strongly prefers other forms of ADDR to direct object) in this LVC (no occurrences in PAROLE), which makes the coreferential relations quite difficult to resolve. The entire sentence gets the regular tectogrammatical representation if the Addressee is realized on the surface of the verbal frame, e.g. in the sentence *I en farmartjänst ställer de sina maskiner och kunskap til förfogande för industrier och kommuner för exempelvis grönyteskötsel*. Then the phrase *för industrier och kommuner* actually stands in grammatical coreference with the Actor of the predicate noun *förfogande*, i.e. the Actor of the predicate noun *förfogande* will never be realized on the surface together with the verbal ADDR, and it will obtain the substitutional tectogrammatical lemma QCor. A non-standard situation occurs when the nominal ACT is realized on the surface, e.g. in the sentence *I stället ställer Nato staber och militära resurser till Västeuropeiska unionen VEU:s förfogande vid kriser – operationer…* Then the verb is not allowed to realize the ADDR-participant on the surface, and the verbal ADDR gets the substitutional t-lemma QCor. The coreferential arrow pointing up in the tectogrammatical tree structure is hypothetically possible but has not occurred during the annotation of PDT, which implies that the issue has not been sufficiently explored yet.

- Incep (for inchoative/inceptive events)
- Cont (for continuing events)
- Fin (for finishing/ceasing events).

Examples:

1. *to open fire on X* = IncepOper_1(*fire*)
2. *to fall under the power of X* = IncepOper_2(*power*)
3. *to lose one's power over X* = FinOper_1(*power*)
4. *to get out of X's control* = FinOper_2(*control*)
5. *to retain one's power over X* = ContOper_1(*power*)

8.5 Causative Lexical Functions

Like phasal Lexical Functions, causative Lexical Functions specify basic Lexical Functions by observing additional semantic features that have to do with causativity. The causative Lexical Functions are:

1. Caus (for causative verbs)
2. Liqu (for verbs with the meaning of stopping an event)
3. Perm (for verbs with the meaning of permitting an event)

Examples:

1. *to lead X to an/the opinion* = CausOper_1*(opinion)*
2. *to stop aggression* = LiquFunc_0(*aggression*)
3. *to condone aggression* = PermFunc_0(*agression*)

They can even be used as basic LFs with verbs as keywords, even in Swedish: Caus(*falla*) = *fälla*, Caus(*sitta*)= *sätta*, etc.

Chapter 9

Examples of Valency Lexicons and Collocational Lexicons

The complex structures of Swe-VALLEX/PNL draw on a huge amount of lexicographical research as well as lexicographical work performed by various lexicographers in the past and in the present. This Chapter mentions (with gratitude) the most essential sources of inspiration for Swe-VALLEX/PNL. Lexicons primarily designed for human use as well as lexical sources for NLP applications (and combinations of both) are represented here, yet a thorough analysis and mutual comparison of them would go beyond the scope of this study. Therefore, only a very simplified overview of their most distinctive features regarding valency description and collocation sorting is given.

9.1 BBI

The BBI Dictionary of English Word Combinations[108] is explicitly dedicated to learners of English as a foreign language. It distinguishes two types of collocations, which it seeks to capture:

- grammatical collocations
- lexical collocations.

By the expression *grammatical collocations*, BBI understands "a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as infinitive or clause", intuitively regarded by native speakers, as opposed to 'free combinations', which "consist of elements that are joined in accordance with the general rules of English syntax and freely allow substitution". Lexical collocations "normally do not contain prepositions, infinitives, or clauses. Typical lexical collocations consist of nouns, adjectives, verbs, and adverbs" and are contrasted with free

lexical combinations, "in which the two elements do not repeatedly co-occur; the elements are not bound specifically to each other; they occur with other lexical items freely".

Each grammatical as well as lexical collocation is characterized by a combination of letter G or L and a number. BBI observes 8 types of grammatical collocations:

1. G1: noun + preposition combinations, e.g. *blockade against*
2. G2: noun + *to* + infinitive, e.g. *attempt to do it*[1]
3. G3: noun + *that* clause, e.g. *agreement that*
4. G4: preposition + noun combinations, e.g. *in advance, to somebody's advantage*
5. G5: adjective + preposition combinations that occur in the predicate of as set-off attributives (verbless clauses), e.g. *angry at*
6. G6: predicate adjectives + *to* + infinitive, e.g. *ready to go*.[2]
7. G7: adjective + *that* clause, e.g. *She was afraid that she would fail the examination.*
8. G8: The G8 group consists of 19 verb patterns, distinguished by capital letters A-S. The observed patterns are the following:
   (a) A: dative movement transformation, e.g. *He gave his brother a book – He gave a book to his brother*.
   (b) B: transitive verbs with indirect object that do not allow the indirect object in dative, e.g. *He mentioned the book to her – *He mentioned her the book.*
   (c) C: transitive verbs allowing a preposition-less indirect object in alternation with prepositional object with *for*, e.g. *She bought her husband a shirt – She bought a shirt for her husband.*
   (d) D: verb + preposition, e.g. *to work as, to adhere to*
   (e) E: verb + *to* + infinitive, e.g. *He decided to come.*
   (f) F: verb + infinitive without *to*, e.g. *help (He helped her climb the stairs.)*
   (g) G: verb + *-ing*, e.g. *enjoy watching TV*
   (h) H: verb + direct object + *to* + infinitive, e.g. *She asked me to come*.

[1] Surprisingly, BBI observes this pattern even in contexts that appear to suggest that the infinitive is not syntactically dependent on the noun, e.g. in sentences with expletive *it*:*It was a pleasure to do it = To do it was a pleasure.*, which does not seem quite appropriate.

[2] G6 raises the same question with predicate adjectives as G2 does with nouns, and so does G7.

(i) I: verb + direct object + infinitive without *to*, e.g. *She heard them leave.*

(j) J: passive use of the verbs of the group I, e.g. *We were made to get up.*

(k) K: verbs + possessive pronoun or noun in genitive + gerund, e.g. *Please excuse my waking you so early.*

(l) L: verb + *that* noun clause, e.g. *He informed his students that the examination had been canceled.*

(m) M: verb + direct object + *to be* + adjective or past participle or noun/pronoun, e.g. *We consider her to be very capable.*

(n) N: verb + direct object + adjective or past participle or noun/pronoun, e.g. *She dyed her hair red.*

(o) O: verb + two preposition-less objects, e.g. *The police fined him fifty dollars.*

(p) P: verb + obligatory adverbial , e.g. *He carried himself with dignity. – *He carried himself.*

(q) Q: verb + wh-word: *He asked how to do it.*

(r) R: transitive verb often expressing emotions preceded by the dummy *it* + *to* + infinitive or *that*-clause, e.g. *It surprised me to learn of her decision, It surprised me that our offer was rejected.*

(s) S: intransitive verb + predicate noun or predicate adjective, e.g.: *She became an engineer.*

In terms of FGD, BBI's 'grammatical collocations' partly correspond to definitions of deep valency frames (notes on obligatory syntactic complementations, e.g. G8-P), and partly to constraints on surface frame slot fillers (e.g. object complement can only be realized with *to*-infinitive, G8-M; required prepositions in G1 etc.) Some groups (e.g. G8-A, B, C) observe syntactic alternations of English verbs, as proposed and sorted by Levin [85]. This feature is to be included in the next version of the Czech Vallex.

Besides the 8 'grammatical collocation' types, BBI introduces 7 'lexical collocation' types. This list, together with a short explanation and examples, are given below.

1. L1: transitive verb + noun/pronoun or prepositional phrase. The verb denotes *creation* or *activation* of the nominal part, e.g. *to come to an agreement, to set a record*, but also *to compose music*.

2. L2: transitive verb + noun/pronoun or prepositional phrase. The verb denotes *eradication* or *nullification* of the nominal part, e.g. *reject an appeal, lift a blockade, break a code*.

3. L3: adjective + noun, e.g. *strong tea, reckless abandon*. Even nouns in adjectival positions have been entered: *house arrest, jet engine*.

4. L4: noun + verb. The noun is the subject of the verb: *blood circulates, blizzards rage*.

5. L5: noun + *of* + noun. The first noun denotes a unit: *a swarm of bees, a pack of dogs*.

6. L6: adverb + adjective, e.g. *strictly accurate, sound asleep*.

7. L7: verb + adverb, e.g. *to affect deeply, to anchor firmly*.

BBI's collocation types resemble Lexical Functions but they are much more general. e.g. L1 and L2 do not only capture LVCs but all unpredictable combinations of verbs with effected objects[3].

9.2 Combinatorial Explanatory Dictionaries

The attributes 'combinatorial' and 'explanatory' identify a dictionary as one elaborated within the Meaning-Text Theory, a formal theory of natural language (for further reference see e.g. [77]), which emphasizes the development of highly structured lexica. It regards the lexicon as the pivot element in formal description of natural language. Therefore, MTT-based dictionaries contain vast amount of lexical information, useful for the human user as well as various computational applications.

With the help of [102], [77], and [131] a short description of the main features of combinatorial explanatory dictionaries (CEDs) will be given here. CEDs are production-oriented dictionaries, i.e. they seek to provide the user (be it a human or a non-human one) with all lexical information needed to correctly use the given lexical unit in context. 'Lexical unit' is one reading or one sense of the given lexeme. The question how senses are distinguished in CEDs is left aside here.

Each entry comprises ten types of information:

1. spelling, spelling variants

2. pronunciation

3. information on part of speech, reference to a declension or conjugation pattern and an explicit list of its irregular forms

4. stylistics

[3]On the difference between *effected* and *affected object* see [6] and [137].

5. definition and connotations

6. the word's government pattern GP (corresponds to FGD's deep valency frames and surface valency frames)

7. information on combinatorial potential of the lexical unit, described by means of Lexical Functions

8. examples

9. phraseology

10. informal comments on culture-specific issues affecting the use of the given lexical unit, additional explanations of the government pattern etc.

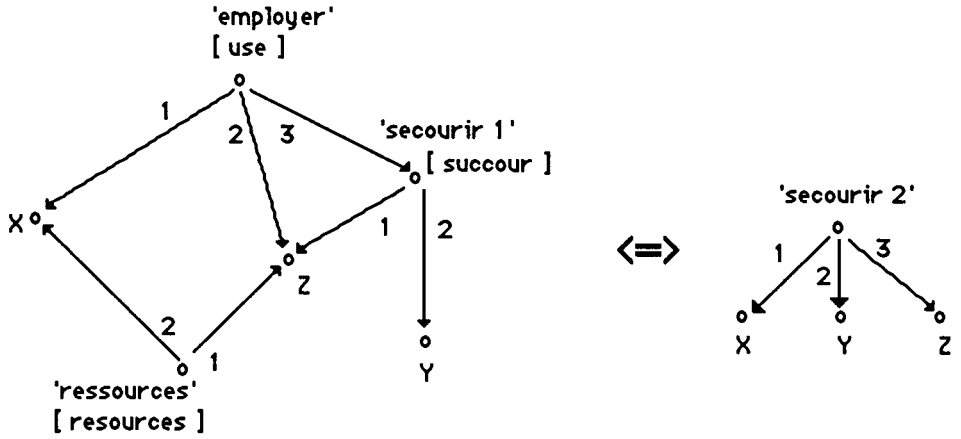
Definitions, GP and combinatorial potential will be described in more detail.

9.2.1 Definitions in CEDs

Definitions in CEDs are shaped as pairs of dependency graphs. The second graph (being placed to the right) depicts the lexical unit and its semantic relations towards its actants, i.e its complementations required by the GP (to be described below). The complementations are marked with capital letters. The first graph depicts the semantic decomposition of the relations between the lexical unit and its actants, as well as relations between the respective actants. It is in fact a graph of a sentence, in which semantically simpler (i.e. more general) words are used to express the same utterance as a sentence with the lexical unit in question. The left-hand definitions do not use any rigidly formalized interlingua, yet it is assumed that a closed list of recurring decomposition patterns is going to be the result of processing a larger data set. The words repeatedly used in the semantic decomposition are then expected to form a list of semantic primitives as a by-product (Fig. 9.1). The figure is taken from [102].

9.2.2 Government Pattern

Each GP appears in the form of a table. Its first row bears the names of semantic actants of the given lexical unit. The second row contains numbers of their corresponding surface syntax matches. The lowest row assigns each actant its relevant syntactic and morphological constraints (lists possible prepositions and parts of speech that can follow them etc.) For illustration see Fig. 9.2, copied from [102], p. 270.



Definition of the French verb SECOURIR 2

What we mean by a (lexicographic) definition of a lexical meaning is an equivalence between this lexical meaning itself taken together with its Sem-actants (the righthand part of the figure) and its semantic decomposition observing the Maximal Block Principle (the lefthand part).

In English, the network in the lefthand part can be read as: 'X uses Z which is X's resources in order to SECOURIR 1 Y'

Figure 9.1: A definition in the explanatory combinatorial dictionary

9.2.3 Lexical Combinatorics

CEDs keep strictly apart linguistic restrictions of lexical cooccurrence from those resulting from the common knowledge. As already Mel'čuk ([99], p. 47 glosses), "the analysis of meaning itself goes beyond the scope of MTT: it does not distinguish 'normal' meanings from absurdities, contradictions or trivialities. Discovering that something is stupid or absurd or detecting contradictions is by no means a linguistic task". The truly linguistic lexical constraints are captured by means of Lexical Functions. Lexical Functions have been described in detail in Section 8, therefore no more details will be given here.

9.2.4 Monolingual and Multilingual CEDs

So far there have been two major combinatorial explanatory dictionaries, Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka (An Explanatory Combinato-

X	Y	Z	W
I	II	III	IV
1. N	1. N	1. à N 2. à V 3. dans N 4. pour N 5. pour V	1. de N 2. avec N 3. par N 4. Adv

Government pattern of the French verb AIDER 2

X, Y, Z and W are the semantic actants of AIDER 2: X is the person who helps, Y the person who receives help, Z the activity of Y in which he needs help, and W the resources by which X helps Y.
I, II, III and IV are the Deep-Syntactic actants of AIDER 2: I refers to the noun phrase that expresses X etc.

Figure 9.2: A government pattern in the explanatory combinatorial dictionary

rial Dictionary of the Contemporary Russian Language) [69]) and Dictionnaire explicatif et combinatoire du français contemporain [100]. Attempts on linking monolingual MTT-based lexical databases into one multililingual open-source combinatorial explanatory dictionary have been reported [145]. More information can be obtained at http://www.papillon-dictionary.org.

9.3 Dictionary of Czech Phraseology and Idiomatics

The third volume – *Verbal Expressions* – of the *Dictionary of Czech Phraseology and Idiomatics* (*Slovník české frazeologie a idiomatiky. Výrazy slovesné*, [20]) is a comprehensive work that lists, analyses and explains about 20,000 verbal phrasemes and idioms. Under verbal phrasemes and idioms the authors understand expressions that contain a verb and act as verbal phrases, most typically as complex predicates. The verbal expressions are grouped according to their syntactic form. The following groups have been observed:

1. verb + noun
2. verb + abstract noun

3. verb + adjective

4. verb + adverb

5. verb + subordinate clause

6. verb + synsemantic words, e.g. prepositions, pronouns, numerals, interjections or particles

The group most relevant for this study is the one defined by the combination of a verb and an abstract noun. Its members make up about one half of the dictionary (i.e. approx. 10,000 entries). They are regarded as 'quasiphrasemes' [20], p. 10 and a special entry type was introduced for their description. They are usually semantically transparent. Thus they do not need extensive explanation but they represent quite rigid collocation patterns (paradigms). Therefore, emphasis is laid on listing the possible verbal collocates of the respective nouns in question. These combinations are lemmatized under noun lemmas. Each noun lemma contains information on its valency. It may contain a semantic gloss and/or stylistic remarks.

The following features are distinguished with the verb + abstract noun combinations:

1. the syntactic position that the noun takes (Obj-Subj)

2. their phase meaning (inchoative-IN, durative-DUR, terminative-TERM) for the noun as subject and the noun as object respectively

3. causativity (**K** in front of the phrasal marker = causative)

4. synonymous one-verb expressions

5. obligatory or very commonly inserted adjectival modifiers of the noun.

The group verb + abstract noun is considered to belong on the periphery of phraseology ([20], p. 11), since its members represent a continuum between recurrent combinations and actual phrasemes and idioms. By applying the same sort of entry to constructions as *dostat zájem* on the one hand and *nevědět si rady* on the other hand, the authors avoid drawing an 'artificial'[4] line between phrasemes and that collocation type, for which this study generally uses the term 'light verb constructions'.

[4]"Zastoupení obojího typu hesel se substantivem abstraktním tak zamezuje umělému dělení této oblasti a umožňuje i komplexní studium jejich úlohy ve frazeologii a jazyce vůbec." [20], p. 11.

9.4 PropBank, NomBank

PropBank [114], [115] is the common label for a deep-syntactic corpus annotation above the Penn Treebank corpus, Wall Street Journal Section [107], as well as for a lexicon of deep valency frames of verbs occurring in PTB – WSJ. In 2005, the first version of NomBank[103], [1], the nominal counterpart of PropBank was released, followed by the June-2006 release. NomBank is the extension and continuation of NOMLEX, the LF-inspired, 1000-entry lexicon of nominalizations [96]. This Section describes only PropBank in detail.

The verb entries of the PropBank Lexicon are divided into rolesets. Rolesets roughly correspond to senses. They rely on syntactic rather than semantic criteria, which are considered to be 'subjective and potentially unlimited' ([133] p. 3), while syntactic distinctions be 'rigorous and objective' Therefore rolesets are much more coarse-grained than e.g. WordNet [41] senses. Each roleset includes a set of labeled arguments ('roles') and one or more example sentences, in which combinations of the roles are rendered by the surface syntax. Rolesets are numbered within the respective entries ('roleset-ID's"). In addition, each roleset has a definition-like description attached ("roleset names"). e.g. the verb *to yell* has only one roleset, *yell.01*, which is labeled as *"to cry out loudly"*. The verb *to abandon*, on the other hand, has three rolesets (*abandon.01.-03*).They are labeled as *"leave behind", "exchange", "surrender, give$_o$ver"*, respectively. The PropBank-Lexicon comprises about 2000 roleset names, in which about 4600 rolesets are grouped. (There are 3323 verb entries in the PropBank Lexicon. Some of them also include phrasal verbs. Phrasal verbs do not have entries of their own but are displayed as rolesets). PropBank's conception of valency derives from Levin's assumption [85] that the syntactic alternations that verbs participate in are not arbitrary but reflect underlying semantic components of the events denoted by each given verb. Semantically related verbs can be grouped into classes according to which alternations they take part in. The roleset names group semantically and syntactically related verb senses into classes like the Levin classes. The PropBank classes cut somewhat across the Levin verb classes in accordance with the valency behavior of the data in PTB-WSJ.

Each roleset introduces an enumeration of arguments (roles). The arguments are divided into 'numbered arguments' and 'adjuncts' [133]. The numbered arguments are arguments that take part in the syntactic alternations analyzed by Levin [85], and can become syntactic subjects. The adjuncts are optional, often rendered by prepositional groups and adverbs. Each argument has two parts: the argument number and a semantic descriptor specific to the given roleset. e.g. *to yell* would acquire the following arguments:

Arg0:*Yeller*
Arg1:*Utterance*
Arg2:*Hearer*

The first roleset of *to abandon (abandon.01 "leave behind")* will have the following arguments:

Arg0:*abandoner*

Arg1:*thing abandoned, left behind*

Arg2:*attribute of arg1*

The arguments do not have to be all present on the surface of a sentence at the same time. Thus the first example sentence *And they believe the Big Board.**Arg0**, under Mr. Phelan, has abandoned their interest.**Arg1*** contains only Arg0 and Arg1, while the second example sentence contains all three: *John.**Arg0** abandoned his pursuit.**Arg1** of an Olympic gold medal as a waste.**Arg2** of time.* Considering the syntactic alternations as pairs of alternation realizations, one can often identify alternation realizations in the rolesets. Each example sentence is provided with a supplementary comment, which even sometimes suggests which alternation realization the given sentence represents. Yet these comments are not formalized, nor is it explicitly stated by which alternations the respective rolesets are defined. About one half of PropBank entries are mapped onto VerbNet [79], in which relevant alternations are listed for each verb entry. However, the linking between the PropBank-Lexicon and VerbNet does not extend to the respective sentences. Besides, the PropBank verb classes (i.e. the roleset names) do not correspond to the VerbNetverb classes (i.e. the original Levin classes), and thus the example sentences in the PropBank-Lexicon do not necessarily show the same alternation patterns as the corresponding entry in VerbNet.

9.5 VALLEX, PDT-VALLEX

This Section is dedicated to the VALLEX-twins, VALLEX 1.0 and PDT-VALLEX. Both lexicons have been elaborated within the FGD framework and their common features deeply affected the data structure of Swe-VALLEX, being its main source of inspiration. Though they have a lot in common, there are some differences between them, which do not allow for common description. A comparison of the two lexicons will be given right at the beginning of this section:

9.5.1 Differences between VALLEX and PDT-VALLEX

VALLEX has been built as a machine-readable lexicon for public use. PDT-VALLEX has been developed during the annotation of PDT (for details see [53]) as a supporting annotation tool. VALLEX includes only verbs, while PDT-VALLEX includes also nouns and adjectives.

Being independent of the corpus annotation work, VALLEX has been built proceeding from lemma to lemma. A verb's uses were investigated in text corpora and compared to information in existing lexicons, as a result of which deep valency frames

of the given verb were defined. The deep valency frame slot fillers were then completed with surface frame slot filler constraints, and the entire frame was enriched with examples.

Unlike Vallex, PDT-VALLEX-frames were created in parallel with the data annotation. When the annotated corpus did not contain a given reading of a verb, it simply remained missing in the lexicon. As more and more data had been processed, the PDT-VALLEX-frames were altered, in order to match the data in the best possible way and to secure the highest possible consistency of annotation. As new valency frames of certain types of nouns and adjectives were arising at the later stages of PDT-annotation, they were entered into the lexicon.

9.5.2 Structure of VALLEX

The Czech valency lexicon describes the valency behaviour of each lexeme (verb, noun, adjective or adverb) in form of valency frames, which roughly correspond to senses. Like rolesets in the PropBank-Lexicon, the valency frames primarily rely on syntactic criteria though the syntactic criteria are sometimes modified to take account of the semantics of the given verb (see below).

A valency frame in the strict sense consists of inner participants and obligatory free modifications [120]. Free modifications are prototypically optional and do not belong to the valency frame in the strict sense, though some frames require a free modification as an obligatory slot in the frame (e.g. direction in verbs of movement, see Section 6.3.2). Both the obligatory and the optional inner participants belong to the valency frame in the strict sense. Like the free modifications, the inner participants have semantic labels according to the cognitive roles they typically perform: ACT (Actor), PAT (Patient), ADDR(Addressee), ORIG (Origin), and EFF (Effect). However, if a verb only has one inner participant, it is automatically labeled with ACT. A two-participant verb always has an ACT and a PAT (see Section 6.3.4).

PDT-VALLEX and VALLEX are very similar in structure: each lexeme corresponds to one entry. The entry is divided into valency frames. A valency frame is modeled as a sequence of frame slots. Each frame slot corresponds to one complementation of the verb in question. Each slot is assigned a functor according to its semantic relation to the governing verb. Each slot includes an enumeration of its surface forms. Each frame is supported by at least one example sentence.

PDT-VALLEX notes only the valency frames in the strict sense (i.e. obligatory or optional inner participants and obligatory free modifications), while VALLEX also lists optional free modifications typical of the given frame. When delimiting the respective valency frames, syntactic as well as semantic criteria are adopted. Therefore a verb can have two valency frames with identical distribution of functors. Lopatková [88] notes that "the change in morphemic realization signalizes the possibility of different meanings; on the other hand, particular complementation in a valency frame

can have morphemic variants (if the meaning is 'sufficiently close')".

Additional syntactic information is attached to each frame. It concerns mainly reflexivity and reciprocity conditions, verb control, aspectual characteristics of the given frame and its aspectual counterparts. Besides that, information on possible diatheses is added. Each frame is also classified on the scale of idiomaticity (values 'primary usage' – 'secondary usage' – 'idiomatic usage'). Frames of 'frozen collocations and idioms" [88] differ from ordinary frames in that they contain one slot with the functor CPHR or DPHR. DPHR is reserved "for the dependent parts of collocations with which the complementation is lexically limited to a single word (or to a restricted set of words) and the collocation cannot be syntactically analyzed. The CPHR functor is mainly used for marking predicate nouns in LVCs.

Some frames had been tentatively sorted into semantic-syntactic classes, out of which already 15 classes arose to be of help during consistency checking in VALLEX 1.0. The idea of Levin-like sorting of the verbs into classes based on the verbs' syntactic and semantic behavior significantly affected the recent activities around VALLEX. Žabokrtský [166] suggested a new structure of VALLEX, which would capture regular syntactic alternations. His point of departure was the fact that even very subtle changes of meaning, triggered by syntactic shifts, require their own frames. Consequently, the lexicon grows bigger than could be intuitively expected. e.g. each memeber of the pair of Levin's example sentences would have different functors assigned, and thus it would require its own frame:

1. *The sun.ACT radiates heat.PAT*
2. *Heat.ACT radiates from the sun.DIR1.*

The drawbacks are obvious:

- Though the shift 'ACT – DIR1 and PAT – ACT' possibly applies to many verbs, it will not be explicitely indicated. There will be no evidence that this shift is a regularly occurring phenomenon.
- Two frames instead of one common increases the space requirements of the lexicon.
- Every manually created frame can contain inconsistencies and/or errors.

Therefore, Žabokrtský suggests putting up a list of alternations to observe in clusters of lexical units. (In case of VALLEX 1.0 it means in clusters of the current frames). Whenever a pair of alternation realizations was found, only one member of the pair per alternation would be selected (in a more or less arbitrary way) to be explicitly stated in the lexicon as *basic lexical unit (BLU).* The other pair member would become its *derived lexical unit (DLU).* DLU would be generated 'on demand' from BLU by means

of a transformation rule, which would be applied to DLU. DLU would of course have to contain a list of such alternation rules. In order not to cause difficulties for human users, examples of DLUs would be listed under the relevant BLU together with the name of the alternation [92]. Mutual relations among lexical units within a cluster of lexical units (CLU) can be much more complicated than described here. A given LU can be a BLU of one or more alternations and in addition it can be a DLU of yet another alternation at the same time.

VALLEX distinguishes two alternation types:

- 'syntactically-based' alternations
- 'semanticall-based' alternations.

'Syntactically-based' alternations comprise constructions listed by Czech grammars (e.g. the two types of passivization, reciprocity, the resultative construction *mít něco uděláno, dostat něco uděláno*, the causative *dát/nechat si něco udělat* and some constructions "regular enough to be covered by general rules" like dispositional modality (e.g. *Matematika se mi učí dobře*) and impersonal constructions (*To se ti to mluví*).

'Semantically-based' alternations 'Semantically based' alternations are exemplified in [92]. It is the case of *vyjít kopec.PAT – vyjít na kopec.DIR3*, *poslat dopis mamince.ADDR – poslat peníze do Indie.DIR3*, etc.

9.5.3 Other Language Versions

An English version of VALLEX has been built during the annotation of the Prague English Dependency Treebank (PCEDT), the English counterpart of the Prague Czech-English Dependency Treebank [32]. EngValLex is based on the annotation scheme of VALLEX 1.0 and was obtained by a semi-automatical conversion of the PropBank – Lexicon, and subsequently manually adjusted. For more details on EngValLex see [25] and [144].

Swe-VALLEX is also mainly based on the annotation scheme of VALLEX 1.0. Its structure has been analyzed in more detail in Section 16.3.

9.6 FrameNet

The FrameNet project is perhaps the best-known project in the field of valency description. It is based on the theory of *frame semantics* [43], [76],[44]. Frame semantics investigates valency (or 'predicate argument structure') in verbs, nouns and adjectives, by grouping predicate arguments into *frames* and providing them with semantic labels. The frames are schematic representations of events denoted by the predicating words. The arguments have semantic roles specific to the given word or a set of

words; the words associated to one common frame share the same arguments. e.g. the commercial-transaction frame relates *buy* to *sell*, *spend*, *pay* and *cost* by assigning identical semantic labels to their arguments:

- *Buyer bought Goods from Seller for Money*
- *Buyer paid Seller Money for Goods*
- *Buyer sold Goods to Buyer for Money*
- *Seller sold Buyer Goods for Money*
- *Buyer spent Money on Goods (Seller not expressed)*
- *Gods cost Buyer Money (Seller not expressed)*

The arguments assigned the semantic roles are called *frame elements*. Frame elements are rendered by surface syntax structures.

The purpose of FrameNet corpus annotation is to process an amount of predicating words, i.e. assigning them to frames and encoding information on surface syntax realizations of frame elements in various frames, in order to capture and sort human knowledge about which semantic roles are typically represented, how they are realized in surface syntax and how they are related in various word classes across domains.

9.7 Svenskt Språkbruk

Svenskt Språkbruk – ordbok över konstruktioner och fraser [31] is a large production-oriented monolingual dictionary for human users. Its central goal is to treat words in combinations. The following types of combinations are distinguished:

- *constructions*, i.e. valency patterns
- *phrases*, i.e. collocations
- *idioms*, i.e. semantically non-compositional units
- *pragmatical phrases*, i.e. cliché-like phrases and discourse markers.

Under constructions the authors understand grammatical constructions with virtually unlimited collocation potential. For instance, the construction *kasta ngt* indicates that *kasta* is a transitive verb and requires an object. Phrases are defined as stable word combinations where the lemmatized word requires a certain collocate or a restricted list of collocates. A phrase does not need to be an idiom. Idioms follow the usual characteristics of idioms and pragmatical phrases are evaluative expressions

used in certain situations, e.g. *det är inte mitt jobb att…* to indicate that one is not going to get involved in something.

The following issues are regarded as vital for correct language production:

- use of prepositions
- use of correct collocates in a collocation, e.g. correct adjectival attributes with nouns (***djup*** *snö*), correct verbs governing an object (***sätta upp*** *en affisch*), etc.
- correct use of metaphorical uses
- correct use of evaluative phrases.

This dictionary is organized in a very systematic way and examples had been taken from a corpus. It has been used as a measure during the experimental work on Swe-VALLEX/PNL entries (see Section 16).

9.8 PAROLE-SIMPLE

Though finally not directly projected into Swe-VALLEX/PNL, the list of lexicographical inspirations of this work would be incomplete without mentioning the Swedish and the Danish branches of the PAROLE-SIMPLE project [153], [123].

The EU-project SIMPLE (*Semantic Information for Multifunctional Plurilingual Lexica*) had the goal to provide harmonised semantic lexicons for NLP-tasks for 12 EU-languages, among others Danish and Swedish. The SIMPLE lexicons provided linguistic description of each lexical entry on the morphological, the syntactic and the semantic level. The morphological and the syntactic layer were language-specific, unlike the semantic level, which had been gradually unified into an ontology with extended qualia structure (for further reference on the ontology see [135]) to ensure interchangeability of the data. The application of Pustejovsky's generative-lexicon framework to multilingual data gave rise to a number of interesting papers concerning e.g. collocation- and valency description [13], [14], metaphorical transfer and event structure [124], [110].

9.9 VerbaLex

The lexicon of Czech verb valencies VerbaLex [64] merges three independent valency frame lexicons into the so-called *complex valency frames*. It is a combination of the Czech WordNet [162] valency frames dictionary, VALLEX 1.0 [90], and the syntactic lexicon BRIEF [113]. Entries in VerbaLex enrich the lemmas with synonymic relations and with common valency frame. Apart from the VALLEX functors, it uses a second set of more fine-grained semantic labels.

Part III

Grammar and Corpus – the Case of Swedish

Chapter 10

The Role of Verbs in the Vocabulary

10.1 Verb Statistics

Together with nouns, adjectives and adverbs, verbs belong to the so-called open-class parts of speech (unlike e.g. prepositions). Open-class parts of speech are those which obtain new members as the language develops. Environmental and cultural changes, new experiences, discoveries and challenges give rise to new nouns, verbs, adjectives and adverbs, but hardly ever to new (primary) prepositions, determiners, modals or numerals.[1] While nouns typically denote entities, verbs typically report on relations between them. Verbs thus represent the syntactic core of the sentence. Nouns and verbs are believed to be the only universal parts of speech in human languages; i.e. all known human languages have verbs and nouns, though their syntactic profile in general might be very different from that of Indoeuropean languages.

Viberg (e.g. [156], [161], [160], and many other titles) has performed a typological research on Swedish lexicology, proceeding from a general overview of Swedish "lexical profile" [156] to the functions that various basic lexemes have in Swedish compared to their equivalents in other European languages. He studied in detail concrete verbs like *dra (to pull)*, *ge (to give)*, *göra (to do, to make), gå (to go),* as well as entire semantic fields of verbs like verbs of physical contact, mental verbs etc.

Viberg's analysis of the most frequent lexemes in Swedish [156] shows an interesting fact: there are many more nouns in the language than there are verbs. The Swedish frequency dictionary [2] contains 39 486 nouns but about 8,5 times fewer verbs (4 649). This implies that verbs must have the ability to fit into many more different contexts than nouns. In accordance with Zipf's law is Viberg's observation that almost one half (45,5%) of verb occurrences is represented by the 20 most frequent

[1]Though Hanks [55] observes that, although verbs are traditionally classified as 'open class', new ones are actually rather rare compared with new nouns, names, and MWEs.

verbs. Almost every second verb used in the language is then one of the top-twenty.[2] There is an evident preference for frequent verbs, which is in accordance with Sinclair's claims [149] mentioned in Section 1 and with the observations made by Hanks (Section 3.1).

Viberg divides the verbs, nouns, adjectives and adverbs in the respective top-twenty lists into semantic fields. Let us have a look at the verb list:

10.2 The 20 Most Frequent Verbs in Swedish

1. *är/vara (to be)*
2. *ha (to have)*
3. *kunna (can)*
4. *ska (shall)*
5. *få (to get)*
6. *bli (to become)*
7. *komma (to come)*
8. *göra (to do, to make)*
9. *finnas* (existential *to be*, lit. *to be found*. Similar to the German *es gibt*.)
10. *ta (to take)*
11. *säga (to say)*
12. *gå (to go)*
13. *ge (to give)*
14. *se (to see)*
15. *måste (must)*
16. *vilja (to want)*
17. *stå (to stand)*

[2]The 20 most frequent nouns cover only 8,1% of noun occurrences, the 20 most frequent adjectives cover 24,2% of adjective occurrences and the 20 most frequent adverbs have similar rate as verbs – 42,1% of adverbial occurrences.

18. *visa (to show)*
19. *böra (ought)*
20. *gälla (to apply, to be valid)*

If we ignore the copulas *vara* and *bli* and the modal verbs *kunna, ska, måste* and *böra*, we get a list of lexical verbs, of which *ha* is also used to make the perfect tense (in the same way as English and German). The verb *finnas* is used in existential phrases, and thus it is strictly speaking not a lexical verb any more, but rather the regularly derived deponential form of *finna (to find)*, which is a lexical verb. The semantic motivation of *finnas* as existential verb is obvious, and thus *finnas* will be observed in relation to its active lexical counterpart in Swe-VALLEX/PNL.

The lexical verbs can then be divided up as follows:

1. SPATIAL VERBS:
   (a) BODY POSITION: *stå*
   (b) MOVEMENT: *komma, gå*
2. OWNERSHIP: *ha, få, ta, ge*
3. PRODUCTION: *göra*
4. VERBAL COMMUNICATION: *säga*
5. METALINGUISTIC EXPRESSIONS: *gälla*
6. PERCEPTION: *se, visa*
7. WISH: *vilja*
8. EXISTENCE: *finnas*

The list of the 20 most frequent verbs in Swedish does not substantially differ from analogical lists that Viberg set up for 10 other European languages (English, German, French, Spanish, Italian, Romanian, Polish, Russian, Finnish and Hungarian). The expanded 50-top-frequent lists confirm the dominant occurrence of spatial verbs (the Swedish list of spatial verbs adds *följa (to follow), sätta (to place in a sitting position), ställa (to put vertically), lägga (to put horizontally), dra (to pull) and lämna (to leave)*). Most of the 30 additional Swedish verbs belonged to the fields PERCEPTION and COGNITION. Three new semantic fields have been introduced: MANIPULATION (*hålla – to hold*), QUANTITY (*öka – to increase*) and ORGANIC LIFE (*leva – to live*). While the fields QUANTITY and ORGANIC LIFE contain

words that ended up at the very bottom of the list, the field MANIPULATION contains a verb that ranked 22nd, and is thus evidently relevant. Besides, the MANIPULATION field correponds to verbs of physical control mentioned by Bybee, Perkins and Pagliuca [75]. The PERCEPTION, COGNITION and VERBAL COMMUNICATION fields appear to be less relevant for Swe-VALLEX/PNL as they hardly ever enter Light Verb Constructions (see Chapter 4), which was set as one major criterion for making a verb a lexical entry in a lexicon. Instead of proceeding from verb to verb strictly according to frequency, Swe-VALLEX/PNL rather looks 'deeper' into the fields SPATIAL VERBS, MANIPULATION and PRODUCTION. For more detail on selecting the entry candidates see Section 15.5.

Viberg calls the cross-linguistically predominant frequent verbs *unmarked lexical elements* (*omarkerade lexikala element*, as opposed to language-specific lexical elements). He characterizes them with the following features:

1. They are simple stems rather than derivations or compound words.
2. They have a phonologically simple form.
3. Their conjugated forms are often irregular.[3]
4. They occur in the respective languages with high frequency.
5. Typologically, they have a broad distribution (they exist as equivalents in many languages).
6. They have many "secondary" meanings[4].
7. They have a significant potential to become grammatical markers.
8. They act as syntactic prototypes (i.e. they allow for many valency patterns and occur in more compound words and derivations).
9. They are preferred at the early stages of first as well as of second language acquisition.

Viberg makes one more good point in his cross-linguistic comparison of "typologically unmarked lexical elements": he compares the universality of the respective verbs in other languages to Swedish. Thus he is able to draw conclusions as to whether

[3]This can indicate that the respective forms are acquired by rote learning and remain further unanalyzed by speakers (cf. [17]).

[4]Many studies on this have been published in the Scandinavian area also by other authors. Among others [37], [42], [61], [71], [73], [97], [128], [138].

Verb	English Equivalent	Phrase	Literal English Equivalent	Grammatical Meaning
komma	to come	X kommer att verb-a, X kommer verb-a	X comes to verb, X comes verb	future marker, opposite to ska (shall)
hålla	to hold	X håller på (med) att verb-a, X håller på och verb-ar	X holds on (with) to verb, X holds on and verb-s	progressive, discourse backgrounding, sometimes 'something nearly happened'
få	to get	X får verb-a, X får Y att verb-a, X får Y verb-at, X fick verb-ad Y	X gets verb, X gets Y to verb, X gets Y verb-ed, X got verb-ed Y	modal: may, modal: must, causativity, successfully finished action of X

Verb	English Equivalent	Phrase	English Equivalent	Grammatical Meaning
ligga	to lie	X ligger och verb-ar	X lies and verb-s	atelic event
sitta	to sit	X sitter och verb-ar	X sits and verb-s	atelic event
stå	to stand	X står och verb-ar	X stands and verb-s	atelic event
gå	to go	X gick och verb-ade	X went and verb-ed	ingressive telic event
ta	to take	X tog och verb-ade	X took and verb-ed	ingressive telic event

a verb from a given language matches a particular verb in Swedish or whether a selection must be made from a cluster of context-dependent partial equivalents; e.g. *to put* vs. *ställa – sätta – lägga (to put vertically – to put so that it 'sits' – to put horizontally)*.

The next sections draw on cross-linguistic studies of Swedish basic verbs by Viberg. Selected cases will be presented to illustrate the different degrees of grammaticalization in basic verbs. The spectrum ranges from well-established grammatical constructions ('well-established' means mentioned by SAG [152]) to "recurring strategies for building discourses" [66]. Subsequently some selected basic verbs will be examined as a whole. The basic information including the example sentences stems from SAG (Vol. 4 – Satser och meningar). Additional remarks have been made on the basis of original corpus research.

Leaving aside the verb *ha (to have)* as a perfective auxiliary verb and phasal verbs like *fortsätta (to go on), börja (to start)* and *sluta (to finish)*, Swedish employs the following lexical verbs to render grammatical meanings:

In addition to *hålla på* SAG names a few other verbs that specify or emphasize telicity/atelicity in events:

SAG makes a distinction between *inchoative* and *ingressive*. *Ingressive* means just "starting" while *inchoative* implies a change of state. This difference has not been considered in other places in this thesis and *inchoative* has been used not only for verb phrases denoting transitions but certainly also for verb phrases denoting processes.

In addition to the verbs *ligga, sitta, stå, gå* and *ta*, also the verb *vara (to be)* can be used in pseudocoordination with lexical verbs. It is primarily used to make telic events atelic. Example 82 shows how the primarily resultative construction *få något verb-at (get something done by somebody else)* is 'imperfectivized' by the use of the pseudocoordination with *vara*:

(82) *Han är och får sina tänder undersökta.*
lit. *He is and gets his teeth examined.* (He is having his teeth examined.)

Chapter 11

Komma att

The verb *komma* in combination with the infinitive of a verb is primarily perceived as a grammatical marker of the future tense, along with the auxiliary *skola*, though their use is to a large extent mutually exclusive. Apart from expressing a certain kind of future, the structure *komma+infinitive* has acquired a few more conventionalized uses, which might have developed from the future meaning. This Section discusses the use of the construction *komma+infinitive*.

11.1 Future

There are three ways[1] to express future in Swedish, including *komma+infinitive*:

1. **Using the present tense, possibly with a suitable temporal adverbial:**

   (83) *Vi hittar den, det vet jag.* (SAG)[2]
   almost lit. *We find it, I know that.*

   (84) *Snart sitter du också bakom ett sådant här skrivbord.* (SAG)
   lit. *Soon you also sit behind such a desk.*

2. **Using the future marker *skola (skall, ska)*:** *Skola* with future meaning implies that the event is going to happen and mostly that it is under the subject's control or that it has been decided by somebody else and the subject is going to follow the other person's will:

   (85) *I kväll ska vi gå på bio.* (SAG)
   lit. *We shall go to the movies tonight.*

[1]not including *tänka*, which expresses exclusively agentive human intentions like e.g. the English verb *intend*

[2]Examples marked with 'SAG' come from [152].

(86) *Böckerna ska stå i vardagsrummet.* (SAG)
lit. *The books will stand in the living room.*

Occasionally it is used interchangeably with *komma att*, especially when the speaker draws on an observation of regularities or some revelation:

(87) *Att döma av naturens tecken skall vintern bli kall.* (SAG)
lit. *To judge from the natural signs, the winter will be cold.*

3. **Using the future marker *komma att*:** When used as future marker, *komma att* is in the present tense. Then it means that the mentioned event is going to take place in the future. At the same time, it indicates the speaker's certainty on what he is asserting:

(88) *Flygplanet kommer att landa kl 3.* (SAG)
The plane is expected to land at 3 o'clock.

(89) *Träden kommer att vara så här höga om fem år.* (SAG)
The trees are going to be this tall in 5 years.

(90) *Det kommer (kanske) att sitta en del studenter i salongen.* (SAG)
There will (perhaps) be some students sitting in the lounge.

Unlike the 'planning' verb *skola*, *komma att* says nothing about the speaker's will or preference, even when it is used in the first person, referring back to the speaker:

(91) *Jag kommer att sitta bland åhörarna när du talar.* (SAG)

(92) *I am going to be sitting in the audience while you are talking.*

Here it rather indicates the speaker's certainty, which follows from the speaker's knowledge of the situation or from indices the speaker has perceived and evaluated.

The recent years brought a variant of *komma att* without the infinitive particle, which is gradually spreading from spoken Swedish into the written form. PAROLE found 178 hits, of which all 5 preterital hits and two present tense hits were irrelevant. Two relevant infinitive and no supine[3] hits occurred. A random search among extended concordances showed that *komma+infinitive* without *att* was used in direct speech (e.g. 'spoken' language) as well as in clearly written-style contexts.

[3]Swedish uses a non-inflected form of the past participle to build the perfect past tense, which is called 'supinum' by Swedish grammars.

Fig. 11.1 shows the list of basic verb forms (i.e. infinitives without the infinitive marker *att*) in descending frequency order. Their distribution does not show substantial differences from the collocate list that can be obtained for *komma* with the infinitive marker *att*.

lemma	Freq
p/n komma bliva	20
p/n komma vara	9
p/n komma få	9
p/n komma öka	7
p/n komma ta	7
p/n komma göra	7
p/n komma fortsätta	5
p/n komma se	4
p/n komma kunna	4
p/n komma klara	4
p/n komma ge	4
p/n komma finna	4
p/n komma uppgå	3
p/n komma sälja	3
p/n komma släppa	3
p/n komma hända	3
p/n komma ha	3

Showing page 1
Next >

Figure 11.1: The most frequent verbs governed by *komma* without the infinitive marker *att* in PAROLE

Below are a few examples of *komma* without *att*:

(93) *Makt är något som alltid* ***kommer finnas****, och naturligtvis behövs den.*
Power is something that is always going to exist, and of course it is necessary.

(94) *Vare sig om det är Shockwave, Java eller vad som än* ***kommer vinna****, så* ***kommer*** *vi* ***se*** *mycket mer av interaktivitet på nätet.*
Be it Shockwave, Java or whatever is going to win, we are going to face much more interactivity in the net.

Komma att in the future meaning is not used in the past tense (when referring back before the moment of the utterance, only *skola* is eligible) except when the speaker wants to emphasize that the event was not planned (or the event is by

its nature impossible to be planned, cf. Section 11.2). In that case it is always governed by *skola (skulle)*:

(95) *Han trodde att det skulle komma att vara flera invandrare närvarande.* (SAG)
lit. *He had believed that it should come to be more immigrants present.*
He had believed that there would be more immigrants present.

(96) *N.N. skulle komma att bli världens främste tenor.* (SAG)
lit. *N.N. should come to become the world's leading tenor.*
(= It so happened that N.N. has become world's leading tenor.)

A combination with the *skola*-future tense is marginally possible (only two instances were found in PAROLE, and one of them was also used in SAG):

(97) *Man förväntar sig nu att den nya politiken* ***ska komma att inledas*** *med frigivningen av Mbeki.* (SAG, PAROLE)
lit. *One expects now that the new the policy will come to be started with release of Mbeki.*
The new policy is expected to start with the release of Mbeki.

(98) *Demonstrationen, som arrangerades av fackföreningsrörelsen, var en massiv protest mot den borgerliga regeringens sparplaner som framför allt, säger man, ska komma att drabba den sociala välfärden.*
lit. *The demonstration, which was organized by the Trade Union, was a massive protest against the right-wing government's saving plans which above all, it is being said, will come to affect the social welfare.*
The demonstration, which was organized by the Trade Union, was a massive protest against the right-wing government's saving plans which above all, it is being said, are going to affect the social welfare.

(99) *Om han ska komma att skjuta sig? var det då någon som sade.* (Intercorp)
lit. *If he will come to shoot himself? was there anyway someone who said.*
What if he happens to shoot himself dead? said someone finally.

What the three concordances have in common is reported speech introduced by a verb of saying or a mental process, which SAG [152] does not comment on. The working hypothesis was that *ska* is added to mark future since *kommer att* is used as a non-volitionality marker and has lost its temporal meaning (cf. Section 11.2, below), and hence to play down the asserting value of the proposition by explicitly saying that the outcome is not known yet. Since PAROLE and Intercorp delivered too sparse evidence, an additional search in the

respective corpora within Konkordanser [151] was performed. Konkordanser gave some 30 concordances. Given that Konkordanser currently comprises approximately 150 million tokens, this construction has proven fairly marginal. It occurred mainly in the more recent press corpora, and it combined with the following introducers:

- *veta (know)*
- *bero på (depend)*
- *tro (believe)*
- *rädslan för/oro för (fear/anxiety that)*
- *hoppas (hope)*
- *undra (wonder)*

Several more examples from Konkordanser:

(100) *Vad säger du? Likpredikan? –Ja, just så, Charlotte. Har aldrig Charlotte undrat över vad prästen ska komma att säga på graven, då Charlotte är död?*
lit. *What are you saying? Funeral oration? – Yes, exactly, Charlotte. Has Charlotte never wondered about what the priest will come to say over the grave when Charlotte is dead?*
What are you saying? Funeral oration? – Yes, exactly, Charlotte. Has Charlotte never wondered about what the priest is going to say over the grave when Charlotte is dead?

A number of other concordances corresponded to Examples 95 and 96 (above). All of them were clearly in historical present.

(101) *I denna första del skildras hur fem män möts och slår sig ner i det som senare ska komma att bli Stockholm.*
The first volume depicts how five men meet and settle in a place that will later come to be called Stockholm.

11.2 Accident, Coincidence

Komma att, when used in the past imperfect tense (*preteritum*) without *skola*, acquires a non-temporal meaning indicating that an event happened unintended or by coincidence[4] .

[4]"komma att " ([152], Volume 4, p. 246)

(102) *Jag* ***kom att lämna*** *boken på bordet över natten.* (SAG)
lit. *I came to leave the book on the table over the night.*
I happened to leave the book on the table over night.

The corpus examples[5], however, suggest that *kom att* is also used to emphasize a given event's deviance from an expectation, which either might be the speaker's personal expectation, or a generally accepted norm. E.g. Example 103 clearly indicates that for the speaker (or generally) it is not common to take part in a wedding breakfast where the female participants have nothing on but underwear – which was exactly what the ladies in question did since they felt concern for their dresses as the breakfast was taking place outdoors on muddy ground. The context (a cheerful gathering within the family and best friends) makes the moment understandable, but it remains a startling picture anyway – which was probably the narrator's intention. The same applies to Example 104. Governments do not usually have official dinners in someone's kitchen – not even in Iceland, where both the events take place.

(103) *…måltiden kom att* ***avnjutas*** *i nätta små dessouer…*
lit. …the meal came to be enjoyed in cute tiny underwear…
…the meal turned out to be enjoyed in cute tiny underwear….

(104) *Regeringsmiddagen, som kom att* ***inträffa*** *i vårt kök…*
the government dinner that happened to take place in our kitchen

In combination with atelic predicates the construction *kom att* acquires a more general inchoative meaning, which is motivated by the conception that the event was not at all bound to take place.[6]

(105) *Så* ***kom*** *han* ***att heta*** *Habermeier.* (SAG)
So he came to be called Habermeier.

(106) *De dog […] eftersom de kom att* ***beundra*** *sig själva för mycket i kampen.*
lit. *They died […] because they came to admire themselves too much in the battle.*
They died […] because they could not resist preening during the battle.

Occasionally the coincidence/accident or non-volitionality can be projected onto the infinitive or onto wh-clauses.

(107) ***Att*** *in en pressad situation* ***komma att glömma*** *ett namn kan hända vem som helst.* (SAG)
lit. ***To*** *in a pressing situation* ***come to forget*** *a name can happen to anyone.*
Anyone can happen to forget a name when under pressure.

[5]taken from Intercorp, cf. Chapter 15

[6]"Eftersom *komma* innebär att handlingen också kunde ha uteblivit, får *komma* 'råka' ofta ingressiv betydelse. Detta är speciellt tydligt, när den underordnade verbfrasen anger en oavgränsad aktion." ([152], Volume 4, p. 246)

(108) *…hur den andra delen av ätten kom att* ***ställa sig****…*
lit. *…how the other part of the family came to react…*
…how the other part of the family would react

11.3 A Contrastive Corpus Analysis of the Non-Future Uses

11.3.1 Significant Collocates

Two corpora were searched for the simple past form of *komma* immediately followed by *att*: PAROLE and a parallel corpus of Swedish and Czech (cf. Section 15.2). Thanks to the morphosyntactic tagging, the PAROLE search could be narrowed down to verb infinitives immediately following *att*. PAROLE yielded 555 hits. The parallel corpus (henceforth Intercorp) yielded 128 hits.[7] The top-ten in the distribution of verb collocates immediately following the infinitive marker *att* in PAROLE had a significant overlap with the top-ten in Intercorp (see Fig. 11.2 and Fig. 11.3).

The top-ranking verbs in both corpora were *tänka (to think)* and *bli* (*to become*, also regularly used as passive auxiliary). In neither corpus was *bli* normally used as passive auxiliary (only two cases in Intercorp). Almost all uses of *bli* were copula predicates. The number of occurrences decreased approximately by a double from the third position, and then the decrease was continuous in both corpora.

The corpora had three other verbs in common within the top-ten: *handla, få* and *ligga*. No significant similarities occurred below the twelfth position (*stå* in both corpora). In the variation of the query with the perfect tense (*kommit*, the supine form of *komma*), the top three were *bli, kallas (to be called)* and *tänka* in PAROLE, which yielded 266 hits. The verb *tänka* was the only significant collocate of *kommit att* in Intercorp (11 occurrences from 36 relevant hits). Other verbs had only 1–2 occurrences each.

The search results suggest a rather generalized distribution of the construction *komma att* in the past form, where *tänka* seems to stand out as by far the most significant lexical-verb collocate. This observation will be discussed later in more detail.

11.3.2 Coincidence in Czech

A closer look at the parallel corpus reveals that the assumed expression of coincidence/accident by *komma att* with other verbs than *tänka* has no universal Czech

[7] A beta-version of the software Paraconc [5] was used for the corpus searches. Paraconc uses its own query syntax, which won't be presented here. Instead, ParaConc queries, if made explicit, will be presented in the Corpus Query Language (CQL) format used in PAROLE as well as in the Czech SYN-corpora.

equivalent. Unfortunately, the corpus did not contain enough Czech-to-Swedish translations to give an idea of which Czech context features can make a Swedish translator decide to use *komma att*.

It is very common that neither a syntactic or morphological, nor a lexical translation equivalent of *komma att* can be traced in the Czech translations.

(109) *Vi kom att bli vänner för livet.*
lit. *We came to become friends for lifetime.*
Stali se z nás přátelé na život a na smrt.
lit. *Became of us friends on life and death.*

(110) *…hur Isak kom att finna Karin…*
lit. *…how Isak came to find Karin…*
Jak Izák našel Karin…
lit. *How Izák found Karin…*

(111) *…hans båda norska samtida kom att behärska den svenska scenen…*
lit. *…his both Norwegian contemporaries came to dominate the Swedish scene…*
……švédskou scénu ovládli jeho dva norští současníci…
lit. *…the Swedish scene* (accusative) *dominated his two Norwegian contemporaries.*

(112) *Det var den kvällen Ruben kom att berätta om vad han […] visste om judarnas öden…*
lit. *It was that evening Ruben came to narrate about what he knew about the Jews' fates…*
Toho večera Ruben vyprávěl, co […] věděl o osudu Židů…
lit. *That evening Ruben narrated what he knew about the fate of the Jews…*

(113) *Hon kom för tidigt, bara någon månad efter vigseln, och skammen kom att häfta vid henne genom uppväxtåren.*
lit. *She arrived too early, just some month after the wedding, and the shame came to stick to her through her young years.*
Narodila se příliš brzy, jen nějaký měsíc po svatbě, a hanba ji pronásledovala během celého dětství a dospívání.
lit. *She was born too early, just some month after the wedding, and the shame was haunting her through the entire childhood and youth.*

(114) *Och jag kom att leva i en tid som inte erkände sorgen…*
lit. *And I came to live in a time that did not recognize sorrow…*
A já jsem žila v době, která smutek neuznávala…

lit. *And I lived in a time that sorrow* (accusative) *did not recognize…*

(115) *Hon stirrade på Martin Beck med runda blå ögon, som kom att se onaturligt stora ut genom de tjocka glasen.*
lit. *She stared at Martin Beck with round blue eyes, which came to look unnaturally large through the thick glasses.*
Zírala na Martina Becka kulatýma modrýma očima, které za tlustými skly brýlí vypadaly nepřirozeně veliké.
lit. *She stared at Martin Beck with round blue eyes, which through the glasses of thick spectacles looked unnaturally large.*

(116) *Det som Eskil fick se när de gick runt gjorde dock att han kom att ändra sin uppfattning i grunden.*
lit. *What Eskil got to see when they went around did nevertheless that he came to change his mind completely.*
To, co Eskil viděl, když vše spolu obcházeli, způsobilo, že změnil od základu své mínění.
What Eskil saw when (they) were going round together, caused that (he) completely changed his mind.

As the examples show, the 'no-Czech equivalent' variant is not associated with a particular event structure of the given Swedish predicate. The examples show Swedish predicates denoting transitions (*bli vänner*), telic processes (*finna, behärska, erkänna, ändra sin uppfattning* (in the sense of *approve, accept*)), as well as states (*häfta, se ut*).

When the nominal part of a copula predicate with *bli* is a noun, Czech has two options: the less agentive construction *Y becomes of X* and the regular *X becomes Y*, which occurred more frequently in the corpus. No preference for the less agentive construction could be observed, but the data was too sparse to make a qualified judgement on this issue.

However, there are also cases of lexical expression of the coincidence/accident feature represented in the corpus:

(117) *…för att han kom att höra ett telefonsamtal.*
lit. *.. because he came to hear a phonecall.*
…proto, že ***náhodou zaslechl*** *telefonní hovor.*
lit. *…because he* ***accidentally caught*** *a phonecall.*

(118) *Fredrika Bremer […] kom att skriva sina första anspråkslösa noveller, nämligen för att få pengar…*
lit. *Fredrika Bremer […] came to write her first easy-to-read tales, namely in order to get money…*
Fredrika Bremer […] začala psát své prvé nenáročné povídky ***čistě náhodou*** *– aby si opatřila peníze...*

lit. *Fredrika Bremer [...] started to write her first easy-to-read tales purely by concidence – in order to get money...*

In 117 there are two lexical indicators of coincidence: both the adverb *accidentally* and the verb *zaslechnout (to catch)*, which is a derivation of *slyšet (to hear)*, meaning *to catch accidentally by ear, perhaps just a part, not quite clearly, not from trustworthy sources*, etc.

Another option is for the Czech translation to use a different verb, which has perfective aspect, denotes a transition and focuses its accomplishment, while the original Swedish verb denotes a process, to which *komma att* contributes the inchoativity feature.

Example:

(119) *...också det namnet på planen som – mer eller mindre ironiskt – kom att* ***användas*** *i Israel...*
lit. *...also the name of the plan, which – more or less ironically – came to* ***be used*** *in Israel...*
Jméno plánu ***se*** *také více či méně ironicky –* ***ujalo*** *v Izraeli...*
lit. *...name of the plane* ***took hold*** *– more or less ironically – in Israel...*

Sometimes *komma att* used in the past tense but related to the future is translated by the Czech modal *mít*. This use of *mít* approximately corresponds to *should/ought to*, but is perhaps closest to the German *sollen*, with which it shares the ambiguity of moral obligation versus predestination or logical outcome in modality shifts. This is astonishing, given that *komma att* in the past tense is expected to express coincidence/accident:

(120) *Levertin var den första av judisk härstamning som kom att inta en viktig plats i svensk litteratur.*
lit. *Levertin was the first of Jewish origin who came to take an important position in Swedish literature.*
Levertin byl prvním spisovatelem židovského původu, který měl ve švédské literatuře zaujmout významné místo.
lit. *Levertin was the first writer of Jewish origin, who should in the Swedish literature occupy an important position.*

Such concordances are numerous in Intercorp. They remain interesting despite the fact that they all originate from the same Swedish original (a textbook of Swedish literary history) and they all were translated by the same Czech translator, who might simply be wrong. Is there any sensible motivation for using *měl* (3rd person singular, past, masculine) as a syntactic equivalent of *kom att*? Can *kom att*, primarily a coincidence emphasizer, ever acquire the meaning of logical outcome/predestination? If the answer should be 'yes', one hint can be hidden in the **future** use of *komma att*.

When used in the present tense (cf. Example 91), *komma att* expresses the speaker's certainty about the future, which follows from the speaker's knowledge of the situation or from the indices that the speaker has perceived and evaluated, and it says nothing about the speaker's plans or preferences for the future.[8]. The expression of future by *kommer att* has thus a flavour of objectivity. The construction *kom att* (past tense) refers to a future outcome, which is, nevertheless, past for both the speaker and the listener, and therefore it is already known and in a way objectively existent. Hence, the general implicature of *kom att* governing an event is that the event (or state) has happened (or begun), which makes the entire statement true (cf. Example 101 and also 95, 96). As a result, *kom att* helps even present a clearly evaluative statement (e.g. Example 120) as an objective truth. That *komma att* can be used as a rhetoric trick to increase the credibility of a statement appears far more likely than that a literature textbook would say about one of the national classics that he 'accidentally took an important position in the Swedish literature'. The Czech formulation with the modal *měl* itself is perfectly idiomatic and very common in this genre and context.

There is an interesting difference in the distribution of *komma att* in the perfect tense (perfektum) and in the simple past tense (preteritum). The perfect tense variant is quite rare compared to the simple past tense variant in both PAROLE and Intercorp. The perfect tense variant *kommit att* is apparently restricted to (or at least preferred with):

- states and transitions

(121) *Efter Daytonavtalet hösten 1995 har dock frontrapporterna kommit att ändras.*
lit. *After the Dayton Treaty in autumn 1995 have yet the front reports come to change.*
(The reports are still coming and they change every time.)[9]

(122) *Stockholms Saxofonkvartett har alltmer kommit att framstå som en av de ledande ensemblerna för nutida musik.*
lit. *Stockholms Saxofonkvartett has more and more come to appear as one of the leading ensembles for contemporary music.*
(It is becoming more and more clear that Stockholms Saxofonkvartett is one of the leading ensembles specializing in contemporary music.)

[8]"*Komma* anger inte bara att den underordnade aktionen är framtid utan att talaren så gott detta låter sig göra garanterar aktionens fakticitet. […] *Komma* utsäger […] ingenting om de deltagande personernas beslut, överenskommelse eller planering utan anger bara en mer eller mindre säker prognos eller kalkyl på grundval av vad den som gör bedömningen har för kunskaper om de faktorer som påverkar aktionen (inklusive om egna och andras planeringar och åtaganden) eller om upplysingar som tyder på att aktionen kan förutsägas. […]. Inte sällan anger *komma* att talaren anser sig ha goda skäl att anta att aktionen kommer att äga rum." ([152], Volume 4, p. 244–245)

[9]The non-italics lines are not translations but only glosses or paraphrases.

- passive verb forms with unexpressed agents

  (123) *Denna insikt har senare kommit att kallas Lucas-kritiken.*
  lit. *This insight has later come to be called Lucas-criticism.*
  (This insight was later called Lucas-criticism, but the term spread gradually and it is still in use.)

  (124) *På senare år har en diskussion kommit att föras om möjliga intressekonflikter....*
  lit. *In the recent years has a discussion come to be led about the possible conflicts of interests...*
  (There were gradually more and more discussions and people became more and more interested, and the discussion is still going on.)

- adverbials denoting time spans[10]

  (125) *Under flera decennier har sjukdomsbegreppet kommit att dominera som förklaringsmodell.*
  lit. *For several decades has the notion of illness come to dominate as an explanation model.*
  (For several decades has the notion of illness dominated as an explanation model.)

  (126) *Så småningom har glasnost också kommit att innefatta det levande förflutna och en kritisk inställning till den vedertagna Brezjnevska [...] synen på historien.*
  lit. *Gradually has glasnost also come to include the living past and a critical attitude towards the abandoned Brezhnjev-like conception of history.*
  (Glasnost has been increasingly including the living past etc., which has been a continuous process.)

  (127) *Med åren har hon kommit att uppskatta Sverige.*
  lit. *With the years has she come to appreciate Sweden.*
  (As the years go by, she appreciates Sweden more and more.)

- other modifiers expressing graduality (e.g. in extent)

  (128) *Sedan dess har religionerna kommit att spela en allt mer politisk roll i världen.*
  lit. *Since then religions have come to play a more and more political role*

[10] *idag (today)* was used only in the sense of *nowadays*

in the world.
(Nowadays the religions play a political role in the world, which was a process that started long ago.)

(129) *I dagens högteknologiska samhälle har saker kommit att betyda allt mer.*
lit. *In today's high-tech society have things come to mean more and more.*
(In today's high-tech society things are becoming more and more important.)

After finding a few occurrences of *kommit att* in Intercorp, PAROLE was queried for the sequences

```
[lemma="ha"] [word="kommit"] [word="att"]
[lemma="ha"] [] {1,3} [word="kommit"] [word="att"] ,
```

which together yielded approximately 200 hits.

As the examples above show, verbs denoting processes were clearly found less often than verbs denoting states and transitions, and in most of these occurrences the agents were inherently non-volitional (inanimate or at least non-human, mostly abstract entities), or the verbs themselves were inherently non-agentive, as e.g. *get the insight*.

(130) *Men när de [gudstjänster och dogmer] blir självändamål, som människor skall underordna sig, då har de kommit att skymma Gud.*
lit. *But when the worships and dogmas become an objective in their own right, they start to put God in the shade.*
(When the worships and dogmas become a self-purpose, they will start to shade God.)

(131) *Och i frågor som rörde handel, tullar och skatt hade Kelderek snart kommit att märka att hans egen instinkt […] var bra mycket säkrare än baronernas.*
lit. *And in the issues that related to trade, tolls and taxes had Kelderek soon come to notice that his own instinct was far safer than the barons'.*
(It was becoming all clearer to Kelderek that he had a better instinct for trading, tolls and taxes than the barons.)

(132) *Sedan den stunden när jag vände mig om och såg Tuginda stå bredvid baronens grav har jag kommit att inse en mängd saker…*
lit. *Since that moment when I turned back and saw Tuginda stand beside the baron's grave have I come to understand a lot of things.*
(Since that moment I started to understand a lot of things better and better.)

The sparse parallel corpus evidence suggests that it is the adverbials that affect the Czech verb form rather than *kommit at*. The absence of any extra inchoativity marker in the Czech translation is logical since the inchoativity feature is implicitly present in

the imperfective present form in verbs that denote transitions or telic processes (like Example 133), and it is irrelevant in atelic processes (134). The PAROLE examples above (Examples 121–132) could mostly be translated to Czech using imperfective verb forms and present tense, except the past perfect forms *(**hade** kommit att)*.

(133) *Polisväsendet har visserligen under senare år i allt högre grad kommit att politiseras, samtidigt som kåren som sådan allt oftare utnyttjats i politiska syften.*
lit. *The police has certainly during the recent years to a growing extent come to be politicized, at the same time that the corps are increasingly often abused for political purposes.*
…se policie čím dál víc politizuje a zároveň se sbor jako takový využívá stále víc pro politické cíle…
lit. *The police is becoming more and more political and the corps as such are increasingly more often abused for political purposes.*

(134) *…den antihumanitära cynism som **allt mer** kommit att kännetecknа vad som kallades Det Mänskligare Samhället…*
lit. *…the antihumanitarian cynism which [has] increasingly come to characterize what was called The More Human Society…*
*…nehumánní cynismus, kterým se **stále víc** vyznačuje naše takzvaná zlidštěná společnost…*
lit. *…the antihumanitarian cynism which increasingly characterizes our so-called Humanized Society…*

11.3.3 *Kom att* in Result Clauses

Komma att also sometimes governs the predicates of result clauses, though marginally. The comparison of the queries

```
[lemma!="bliva"& lemma!="vara" & word!="det" & word!=","]

 [word="så"] [word="att"] [] [tag="V@II.S" & lemma="komma"]
 [word="att"]
```

versus

```
[word!="Inte"][lemma!="bliva"& lemma!="vara" & word!="det"

& word!=","] [word="så"] [word="att"] [] [tag="V@II.S"
& lemma!="komma"]
```

showed that the result clause with *komm att* governing the predicate only occurred 9 times in PAROLE, which is approx. 8,5 times less frequently than the result clause

without *kom att* (769 hits in PAROLE). To keep the number of concordances reasonably low for the non-*kom-att* query, the search was narrowed down to exactly one token between *att* and the finite verb. With the span expanded to 4 tokens, only 2 additional concordances with *kom att* were gained. Was the insertion of *kom att* in result clauses just incidental, or could it have a reason? To get an idea, the PAROLE query was complemented with an Intercorp query in order to retrieve Czech parallel sentences.

The Intercorp query about *så att* yielded 5 relevant hits of result clauses in which predicates were governed by *kom att*, which was just a tiny fraction of all sentences with *så att*. This finding corresponded to what PAROLE had revealed. The investigation of the Czech translations (Swedish was the target language in all pairs) brought an astonishing observation: two concordances with *kom att* were translated into Czech as purpose clauses ***(tak, aby...),***. That was very fascinating, considering that *kom att* adds the semantic feature accident/coincidence. One would not expect a sentence with this construction to be interpreted as a purpose clause, since a result that was achieved on purpose is hardly coincidental! Unsuprisingly, the syntactic expression of purpose by the subordinator *för att* is incompatible with the coincidence marker *komma att* in the subordinate clause, as has shown the PAROLE query

```
[word="för"] [word="att"] []{0,5} [lemma="komma"] [word="att"],
```

which gave no single relevant hit, and neither did a similar query in Intercorp.

The first possible explanation is that the Czech translator has simply overinterpreted the original. However, the Czech translations sound idiomatic and make sense. Given we know the outcome of a targeted action and given we positively know that the outcome was desired, is the result automatically a purpose? Then the Swedish result clause subordinator would be polysemous. Or is there a semantic distinction between the same outcome expressed as a result, no matter whether wanted or not, and as a purpose?

The hypothetical distinction between a good result and an achieved purpose is that a purposeful activity is explicitly targeted towards the particular goal expressed by the purpose clause, while a result might be recognized as good even though the preceding activity leading to it must not have been consciously targeted at this particular goal. On the other hand, the purpose clause does not explicitly state whether the goal in question was achieved or not, while the result clause does. In practice, the speaker who reports on a third person's action, rarely knows whether or not the given volitional action of the third person was consciously targeted at the given good outcome (which is existent). This is exactly where a coincidence marker would fit!

As shown above, Czech does not have a syntactic means to express coincidence, and therefore the Swedish coincidence marker gets often lost in translation. The Czech speaker has two options for this case: either a result clause (facticity explicit, evaluation supressed), or a purpose clause (evaluation explicit, factitity supressed). Hence,

for a Czech speaker, the expression of a good outcome of a volitional activity that was possibly but not evidently targeted specifically to this outcome, is a matter of focus. The Swedish speaker, on the other hand, is not forced to make focus decisions at all, but he has the additional option of playing down the implicit purposivity of ambiguous result clauses by avoiding to make a statement about whether or not the good outcome was consciously pursued by the action denoted by the main clause. Nevertheless, the use of *kom att* is – as is even evident from the statistics – by no means obligatory.

The Intercorp concordances with English translations illustrate how differently the Czech translators treated the ambiguous result clauses. The prototypical Czech purposive construction is *(tak)…, aby (so…as to)*, and the result construction is *(tak…), že/až (so…) that/as much…. as to* or *takže (so that/such that)*.

(135) *Han flyttade stegen så att den kom att* ***hänga*** *tätt bredvid fönstret*
lit. *He moved the ladder so that it came to hang tight beside the window.*
He moved the ladder so that it was hanging immediately beside the window.
Přemístil žebřík ***tak, aby*** *byl co nejblíže okna.*
He moved the ladder ***so as [it] be*** *as close to the window as possible.*

(136) *Martin Beck vred sig i stolen så att han kom att sitta mitt emot henne.*
Martin Beck turned in his chair so att he came to sit opposite her.
Martin Beck turned in his chair so that he was sitting opposite her.
Martin Beck se otočil v křesle ***tak, aby*** *seděl přesně proti ní*
Martin Beck turned in his chair ***so as to*** *sit opposite to her*

(137) *Nu var det så att bröllopet försiggick i ett tält, och under regnandet hade marken därinne förvandlats till rena gyttjan, vilket fick bruden och några av de mer frigjorda väninnorna att ta av skor och klänningar, så att måltiden kom att* ***avnjutas*** *i nätta små dessouer…*
lit. *Now it was so that the wedding took place in a tent, and during the rain the soil inside had turned into pure mud, which got the bride and some of the more liberal girlfriends to take off the shoes and dresses, so that the meal came to be enjoyed in cute tiny underwear.*
To přimělo nevěstu a několik jejích svobodomyslnějších kamarádek, aby si sundaly střevíce a róby, takže potom spokojeně hodovaly v titěrném prádélku.
That got the bride and some of her more liberal girlfriends to take off the shoes and dresses, so that they enjoyed the meal in cute tiny underwear.

(138) *Hon drog upp benen så att fötterna i de stora grå raggsockorna kom att vila mot kanten av fåtöljsitsen.*
lit. *She pulled up her legs so that the feet in the big grey knitted socks came to rest*

at the edge of the armchair seat.
Skrčila nohy, až se chodidly opírala o okraj křesla…
She pulled the feet as much that she was leaning with her feet against the edge of the armchair seat.

(139) *Hon flyttade sig ett halvt steg så att hon kom att stå vänd snett emot honom med högra benet.*
lit. *She moved a half step so that she came to stand diagonally opposite him with the right leg.*
Udělala půlkrok kupředu, zastavila se šikmo před ním s pravou nohou mezi jeho nohama.
She made a half step forward, stopped diagonally in front of him with her right foot between his feet.

The only feature the nine concordances evidently had in common was that the respective contexts suggested that the result was possibly conceived as a satisfactory state by the agent. The first thing to prove was whether any argument of the predicate in the result clause was coreferential with an argument of the main clause predicate. This has not proved necessary. Another thing to prove was the agentivity/volitionality of the agent in the result clause. This was not necessary either. On the other hand, all agents of the main clause predicates were human, and therefore agentive/volitional. Unexpressed agents (see Example 143) are agentive/volitional by default. The PAROLE examples below illustrate the possible combinations discussed, disregarding the presence of *kom att* in the subordinate clause.

(140) *Ada Mason […] tog hastigt på sig mrs Ketterings kläder, fäste några falska rödbruna lockar vid öronen och sminkade sig så att hon kom att likna henne så mycket som möjligt.*
Main clause: Agentive/volitional agent, controlled action.
Result clause: Agent coreferential with the main-clause agent, desired result.
Ada Mason hastily put on Mrs. Kettering's clothes, attached some fake red curls around her ears and made up her face so as to be as like her as possible.

(141) *Inifrån skruvade han upp en trälucka och fällde ut den så att den kom att bilda ett litet bord eller en disk.*
Main clause: Agentive/volitional agent, controlled action.
Result clause: Agent non-volitional, coreferential with the main-clause patient, desired result.
He screwed off a wooden desk from the port, and he placed it so that it resembled a little table or a plate.

(142) *Syster Cilla […] arrangerade rocken och filtar kring kroppen så att ingen kom att undra över dess nakenhet.*
Main clause: Agentive/volitional agent, controlled action.

Result clause: Agentive/volitional agent, not coreferential with any argument of the main-clause predicate, no other argument coreferential with any other main-clause predicate argument, desired result.
Sister Cilla arranged the coat and the blankets around the body [corpse] so that nobody wondered about its nakedness.

(143) *Eftersom detta var oacceptabelt för testkonstruktörerna justerades testfrågorna så att de kom att ge samma medelvärde för båda könen.*
Main clause: Unexpressed agentive/volitional agent, controlled action.
Result clause: Non-agentive/non-volitional agent, coreferential with the main-clause predicate patient, desired result.

The evident preference for desirable events in the *så att – kom(mer) att* constructions might also have syntactic roots. In Swedish, the prototypical purpose clause is introduced by the subordinator *för att*, followed by either a finite clause or an infinitive clause. Yet SAG names one case which might be the clue: *för att* may be replaced by *så att* and the auxiliary *skola* when expressing a desire. SAG's examples make it more intelligible ([152], Volume 4, p. 126, footnote 2):

(144) *Han ställde sig i dörren så att hon skulle se honom.*
He placed himself into the door so that she could see him.

PAROLE has yielded 258 matches for the query

```
[word="så"] [word="att"] []{1,5} [word="skulle"]
```

and they all appear to render purpose, in accordance with the footnote in SAG. The structure

```
[word="så"] [word="att"] []{1,5} [word="kom"] [word="att"]
```

thus can also be an interference from the *så att – skola* structure.

11.3.4 Czech Equivalents of *komma att tänka*

So far, only the widely generalized uses of *komma att* have been discussed. However, the prominent position of the verb *tänka* in the collocate list suggests that something specific to this particular word may be going on here: the corpus evidence shows that it is systematically translated as *napadnout (something occurs to someone)*, *uvědomit si (to realize),* and, most often, as *vzpomenout si (recall, remember)*, although the verb *tänka* would be normally translated into Czech as *myslet*. The fact that, when in the construction with *komma att* it is regularly being translated by verbs completely different from *myslet*, the first-choice equivalent of *tänka*, along with the significant

difference in cooccurrence frequency between *tänka* and other collocates of *komma att*, suggests that the employment of *kom att* is the predominant means of modifying event structure in the particular case of *tänka*. This has lexicographical consequences. It definitely ought to be noted under the lemmas *vzpomenout si*, the impersonal *napadnout* and under the lemma *uvědomit si* in a Czech-Swedish lexicon, and it should not be missing under the lemma *tänka* in a Swedish-Czech lexicon.

Example:

(145) *…att jag fortfarande långt efteråt började stamma och gjorde stavfel när jag* ***kom att tänka*** *på det.*
lit. *… that I even much later started to stutter and made spelling mistakes when(ever) I came to think of it.*
…že dlouho ještě když jsem si na to ***vzpomenul****, tak jsem z toho koktal a dělal pravopisný chyby*
that I even long afterwards started to stammer and make spelling errors whenever I recalled it.

The combination of *tänka* and *aldrig (never)* triggers the perfect tense of *komma att*. *Jag har aldrig kommit att tänka på det* (*I have never thought of it* or sooner *It never occurred to me*) has probably become a lexicalized phrase.

word	Freq
kom att bli	71
kom att tänka	54
kom att kallas	20
kom att handla	17
kom att spela	13
kom att få	10
kom att stanna	9
kom att ligga	7
kom att ägna	6
kom att tillhöra	6
kom att betyda	6
kom att stå	5
kom att omfatta	5
kom att känna	5
kom att hamna	5
kom att gå	5
kom att bilda	5
kom att vara	4
kom att utvecklas	4
kom att tala	4
kom att påverka	4
kom att präglas	4
kom att prägla	4
kom att personifiera	4
kom att markera	4
kom att leda	4
kom att dominera	4
kom att vila	3
kom att verka	3
kom att ta	3
kom att möta	3
kom att likna	3
kom att leva	3
kom att intressera	3
kom att innehålla	3
kom att ingå	3
kom att göra	3
kom att ge	3
kom att följa	3
kom att framstå	3

Figure 11.2: The most frequent verbs governed by the simple past tense of *komma* with the infinitive marker *att* in PAROLE

26	20.3125%	tänka
12	9.3750%	bli
5	3.9063%	få
3	2.3438%	inta
3	2.3438%	utöva
2	1.5625%	fungera
2	1.5625%	fästa
2	1.5625%	handla
2	1.5625%	ligga
2	1.5625%	mer
2	1.5625%	minnas
2	1.5625%	stå
1	0.7813%	användas
1	0.7813%	avlägsna
1	0.7813%	avlösa
1	0.7813%	avnjutas
1	0.7813%	begripa
1	0.7813%	behärska
1	0.7813%	benämna
1	0.7813%	berätta
1	0.7813%	bestämmas
1	0.7813%	bestå
1	0.7813%	beundra
1	0.7813%	bilda
1	0.7813%	den
1	0.7813%	deras
1	0.7813%	det
1	0.7813%	dominera
1	0.7813%	dra
1	0.7813%	ersätta
1	0.7813%	finna
1	0.7813%	föra
1	0.7813%	förstå
1	0.7813%	förälska
1	0.7813%	han
1	0.7813%	hedlund
1	0.7813%	hjälpa
1	0.7813%	häfta
1	0.7813%	hänga
1	0.7813%	höra

Figure 11.3: The most frequent verbs governed by the simple past tense of *komma* with the infinitive marker *att* in the parallel corpus

Chapter 12

Hålla på …

12.1 Valency Patterns

The verb *hålla* enriched with the particle *på* is known to have grammaticalized uses. It has three valency patterns, in which the lexical verb is represented by the hypothetical verb *verba (to verb)*:

1. *X håller på med Y (Y = noun)* (lit. *X holds on with Y*)
2. *X håller på (med) att verb-a* (lit. *X holds on (with) to verb*)
3. *X håller på och verb-ar* (lit. *X holds on and verb-s*)

12.2 *X håller på med Y*

The first use is a lexical one, with a very broad sense of *to be busy with, to be engaged in* etc. It often replaces a more specific verb. A few examples from PAROLE give an impression of how unspecific the verb *hålla på med Y* can be:

(146) *Han* ***höll på med*** *idrott.*
He went in for sports.

(147) *Han* ***höll på med*** *sin nya visa.*
He was composing his new song.

(148) *Och det är ingen som* ***håller på med*** *narkotika längre heller, va?*
And there's nobody who takes drugs any more, is there?

(149) *Mamma* ***håller på med*** *lunchen.*
Mum is busy cooking lunch.

This use is most likely the cognitive source for the grammaticalized use of *hålla på* in combination with a verbal clause. The patterns *X håller på (med) att verb-a* and *X håller på och verb-ar* are sometimes interchangeable. The latter construction is more recent (yet dating back to the 18th century, as noted by Pihlström [128]) and it was regarded as a less formal stylistic variant of the former until the late sixties. However, Pihlström pointed out cases that show a clear semantic distinction. Eleven years later, SAG [152] associated each form with a different grammatical meaning, though noting cases in which the two constructions are still partly interchangeable. Here the two constructions will be also treated separately.

12.3 *X håller på (med) att verb-a*

The primary grammatical meaning of *X håller på (med) att verb-a* is progressive. Its use as a progressive is restricted to verbs denoting processes (i.e. neither transitions nor states):

(150) *Jag **håller på att** lära dom gilla min musik.*[1]
I am teaching them to like my music.

(151) *En ny rutin **håller på att** etableras.*
A new routine is being established.

(152) **Boken **håller på att** ligga på bordet.*
The book is lying on the table.

(153) **Chefen **håller på att** avgå.*
The boss is resigning.

Concordances from PAROLE show that the preposition *med (with)* is on the verge of disappearing from this pattern. The 9 (!) concordances indicate restriction to human agentive subjects (also observed by SAG) and verbs denoting processes in unbounded events:

(154) *nån full jävel **höll på med att** skjuta tändstickor mot högtalarna i taket*
some drunken git was shooting matches at the loudspeakers in the ceiling

(155) *Rolf Nygren har **hållit på med att** sälja EMI:s musik i 21 år.*
Rolf Nygren has been selling EMI's music for 21 years.

(156) *När alla satt sig ner och som bäst **höll på med att** äta och dricka föll det en sten på granntomten.*
When everyone had sat down and they were in the middle of eating and drinking, a stone fell onto the neighbor's lot.

[1]These examples have been taken from SAG.

This restrictive tendency moves its semantics back, closer to the apparent source *X håller på med Y (Y = noun)*[2] as described above. However, it is still able to carry out the functions of a grammatical construction like discourse backgrounding (which proves the last example sentence).

The variant *X håller på att verb-a* is the predominant one. It has two grammatical meanings:

1. **progressivity**

2. **tendentiality**.

12.4 Progressivity

The progressive use approximately corresponds to the English gerund *to be verb-ing*. It is used for backgrounding events in the discourse and to indicate ongoing processes. Unlike the English gerund it is unacceptable with verbs denoting states and with verbs denoting transitions (see above). The progressive meaning is only activated in combination with atelic verbs. The combination with telic verbs yields the tendential meaning (see below). It can be used together with verbs in passive.

The progressive meaning can be rendered by *X håller på att verb-a* as well as by the coordinated construction *X håller på och verb-ar*. Pihlström observed speakers' preference for the coordinated construction, even though it had not yet been accepted as standard in the 80's. SAG does not comment on the respective variants' stylistic values but adds the same observation. According to SAG, some speakers even make a sharp semantic distinction between the two variants in that they exclusively associate *X håller på att verb-a* with tendentiality and *X håller på och verb-ar* with progressivity. However, SAG mentions another tendency that goes against this semantic distinction: the coordinated construction is strongly preferred with animate agentive subjects although it is still considered odd with inanimate non-agentive subjects:

(157) *Klimatet* ***håller på att*** *bli varmare.*
?Klimatet ***håller på och*** *blir varmare.*
The climate is becoming warmer.

PAROLE contains only 118 instances of **X håller på och verb-ar**, out of which indeed only one's subject is inanimate (a computer) but it is agentive:

(158) *och att en dator nu* ***höll på och*** *smälte svaren.*
and that then a computer was digesting (i.e. processing) the answers.

[2]Pihlström [128] infers this path of development from the entry in SAOB [132].

Progressivity marking is typical in telic verbal clauses in the past tense when the context indicates that the described event was prevented from reaching the expected terminal point ([152], p. 340):

(159) *Karin* ***höll på att tvätta håret*** *men blev avbruten.*
?Karin ***tvättade håret*** *men blev avbruten.*
Karin was washing her hair but was interrupted.

12.5 Progressivity Hides Constancy

Interestingly, the construction *hålla på och* as well as *hålla på att* (though less frequently) appears to acquire the meaning *constantly* (which is a sort of contrary of both progressivity and tendentiality).

The parallel corpus has yielded one spectacular example. It is the Swedish translation of a text originally written by B. Hrabal in very colloquial Czech:

(160) *proto se taky náš farář musel* ***v jednom tahu*** *modlit, aby nebyl tak zlej...*
därför måste också vår präst ***hålla på och*** *be* ***stup i ett****, så att han inte skulle vara så elak ...*
and that's why our priest had to be praying ***all the time*** *in order not to be so evil...*

The Swedish idiom *stup i ett* is perfectly equivalent to the Czech *v jednom tahu*. However, the translator added the *hålla på* construction partly to emphasize that the priest had been praying constantly or very often, but also as an indication of colloquial register[3].

More sentences containing at combination of *hålla på* and atelic verbs were sought in PAROLE, which might be bearing the semantic component of constancy. No unambiguous declarative sentence in the past tense has been found that would be acceptable without a disambiguating adverbial. Most hits (approx. 30) were propositions with low factitity[4], i.e. negative sentences, sometimes with the imperative *ska (should, ought to)*, questions, and infinitives.

All the instances from PAROLE seem to be quotations of direct speech or free indirect speech[5], which suggests that this use of *hålla på* is still confined to spoken language. Exceptions will be discussed below.

[3] This assumption was confirmed by the translator in personal communication [84].

[4] (See 5.3, Fig. 5.1).

[5] Wikipedia (http://en.wikipedia.org): **"Free indirect speech** (or **free indirect discourse** or **free indirect style**) is a style of third person narration which has some of the characteristics of direct speech. Passages written using free indirect speech are often ambiguous as to whether they convey the views of the narrator or of the character the narrator is describing. Free indirect speech is contrasted with direct speech and indirect speech."

Here are a few sentences from PAROLE in which *hålla på* could be substituted with *hela tiden (all the time):*

(161) *I princip tyckte hon det verkade botten att* ***hålla på och*** *knega mellan nio och fem .*
Basically she meant that it appeared miserable to keep working from nine to five.

(162) *Ni ska inte* ***hålla på och*** *larva er sådär, för jag har ingenting att skämmas för.*
You are not supposed to keep acting like this because I have nothing to be ashamed of.

(163) *Men i längden så kan vi ju inte* ***hålla på att*** *bara försvara oss.*
But for a longer time we can't just keep defending ourselves.

(164) *Är det slut? — Det vet jag inte heller. Varför ska du* ***hålla på och*** *fråga så där?*
Is that the end? – I don't know, either. Why do you keep interrogating me like this?

(165) *Men jag tyckte det var lika bra att vara kvar och inte* ***hålla på att*** *bråka.*
But I meant the best thing to do was to stay there and not to keep fighting.

PAROLE yields just one instance of a positive declarative sentence in the present tense, and, in this particular case, *hålla på* was disambiguated by temporal adverbials in the close context (cf. Example 160):

(166) *"Det var* ***alltid*** *bara som du inbillade dig." "Du förnekar det* ***fortfarande****. Det är otroligt." "Det är otroligt att du* ***fortfarande håller på och*** *ältar det. Jag gillade henne aldrig."*
"You had just been fancying it". "You ***still keep denying*** *it. It's incredible". "It's incredible that you* ***still keep agonizing*** *over that. I* ***never*** *liked her."*

How is it that a progressive construction has acquired just the opposite meaning? The progressive *hålla på* is the default interpretation of *hålla på* with atelic verbs. It appears in positive as well as in negative declarative clauses, questions etc., in all tenses. On the other hand, the 'constancy' *hålla på* seems to almost exclusively appear in negations, questions and infinitives (this is at least what the corpus evidence says). It is negation that gives a clue for the semantic change. In a negated progressive sentence, it is not just a single moment of the given event that is negated, but it is the entire event. For instance, the sentence *De håller på att bråka (They are fighting)* focuses just one moment in the ongoing action. The same goes for the progressive aspect as a discourse backgrounder: *De höll på att bråka när jag kom (They were just fighting when I arrived)*. However, the negation of the sentence predicate says that **the entire event** does not take place (at the moment of reference), not that **a single moment of the event** does not take place at the moment of reference. This is best perceived in

the imperative; for instance, by saying: *Don't be doing*, the speaker necessarily means: "**Stop** doing that you have been doing just long enough to annoy me". Implicitly, the event really must have been taking place.

Nevertheless, the relation between progressivity and facticity also works the other way round: when the 'constancy' *hålla på* is employed in a negative imperative with an event, it suggests that the given event is actually taking place and should be stopped [6]. In written language, the reader naturally has no way to decide whether the given event is just taking place or not. By employing the 'constancy' *hålla på* the speaker adds some kind of asserting modality.

A search in Google returned a few additional interesting examples.

(167) *Vi **ska inte hålla på** och keynesianskt försöka mota konjunkturer. Vi ska bygga en robust arbetsmarknad*
och en stabil privat konsumtion som bägge ska klara att anpassa sig till chocker. [7]
We are not supposed to be reaching for conjunctures in a Keynesian fashion. We are supposed to build a robust labour market....

(168) *De latinamerikanska, asiatiska och afrikanska staterna **ska inte hålla på och**blanda sig i USA´s och Europas affärer hela tiden!*[8] *The Southamerican, Asian and African states should not permanently get involved in the USA's and Europe's affairs!*

The speakers in these examples virtually underspecify the actual event(s). What they do instead is label them with expressions that are evaluative, with clear (here negative) connotations: *reach for conjuncturalisms instead of building a solid labour market, get involved in someone else's affairs without being invited.*

---

[6]According to [84], the authentic sentence

(1) *Jag har sett som min uppgift att övertyga mitt eget folk om att vi inte kan hålla på och förtrycka ett annat folk.*
I have considered it to be my task to convince my own nation that we cannot keep suppressing another nation.

really assumes that the suppressing is taking place. It would be unacceptable to say

(2) **...att vi inte kan hålla på och förtrycka ett annat folk genom att börja bygga järnvägar på deras mark.*
**...that we cannot keep suppressing another nation by starting to build railways in their territory.*

[7]Quoted from Google, 2006-09-19, URL<http://forum.svt.se/jive/svt/report.jspa?messageID=84154>.

[8]Quoted from Google, 2006-09-19, URL<debatt.passagen.se/show.fcgi?category=3500000000000014&conference...>.

Another notable thing with these examples is that the 'constancy' *hålla på* can be used with general statements (indicated by *man (one)* and *en…som (a…which)*[9]

(169) *En sida som riktar sig till barn och ungdomar* ***ska inte hålla på och*** *göra reklam för pornografi, säger Mats Albinsson på Rädda Barnens hotline.*[10]
A homepage targeted at kids is not supposed to be advertizing pornography, says Mats Albinsson at the Rescue Kids' hotline.

(170) *Man* ***ska inte hålla på och*** *hetsa mot folkgrupper.*[11]
One should not incite hatred for particular groups of people.

The transition of the progressive *hålla på* into a 'constant' *hålla på* is a good example of a not yet completed context-induced reinterpretation (see Section 3.5.2, p. 34). The focal sense A is clearly progressivity. The conversational implicature associated with A is 'X is happening at the time of reference'. Constancy is the non-focal sense B. The conversational implicature associated with B is 'X has been happening to the time of reference'. The corpus evidence suggests that the sense B is still bound to contexts where ambiguity is not likely to arise (negative statements, infinitives, questions and declarative clauses with disambiguating adverbials).

12.6 Tendentiality

When used with telic verbal clauses to mark tendentiality, only the variant *hålla på att* (never including *med*) is used. Tendentiality in the present tense means that a telic event is taking place and is nearing completion [152], p. 335:

(171) *Allt slammet* ***håller på att*** *rinna ut i bäcken. Vi måste göra något*!
All the mud is about to flow out into the pond. We have to do something!

Used in the past tense, the construction can be ambiguous when no contextual clues are given. Beside the meaning 'taking place and nearing completion' it is often used to indicate that an event almost but not completely reached its terminal point (typically due to an interruption) ([152], p. 335):

[9]NB: No evidence of this generalized use in *spoken* language could be found as there is no publicly available corpus of spoken Swedish. Similarly, in written English *don't go doing something* can be used regardless of whether the mentioned event is taking place at the given moment or not, at least with general statements. A random look at Google returned a homepage advising children how to prevent insect bites and stings: *Don't go annoying ants by stomping on their nests*. (Quoted 2006-09-17 from URL<http://www.cyh.com/HealthTopics/HealthTopicDetailsKids.aspx?p=335&np=285&id=1707>.) This contrasts with *Don't go on doing something*, which necessarily implies an ongoing process.

[10]Quoted from Google, 2006-09-19, URL<www.tiger.se/blog/archives/2004/03/23/index.html >.

[11]Quoted from Google, 2006-09-19, URL<www.fritext.se/hets.html – 5k>.

(172) *Bertil* ***höll på att*** *vattna ihjäl mina blommor.*
lit. *Bertil had almost watered my flowers to death.*

Sometimes the underspecified context only focuses on the final stage of the event which precedes its completion (the example comes from SAG):

(173) *Han* ***höll på att*** *tömma brunnen (när de kom).* (SAG)
He had almost emptied the well when they arrived.

(174) *Fan, det var tur …, sa Olle . — Vadå? — Såg du inte? Den höll på att sitta.*
Damn it, what a luck…, Olle said. – What? – Didn't you see? It was already getting there. (the hockey puck into the goal)

The ambiguity is especially typical of events that can be conceptualized both as a telic process and as a transition:

(175) *Patienten höll på att dö.* (SAG)
The patient was dying.
The patient almost died.

Some typical examples of the sense 'something nearly happened' from PAROLE follow:

(176) *Jag* ***höll på att*** *ta bussen häromdan men bestämde mig för att gå istället.*
I nearly took the bus the other day but then I decided to walk instead.

(177) *Hon* ***höll på att*** *ropa : det finns ett badrum till, men svalde orden av någon sorts skamkänsla.*
She almost exclaimed: There's another bathroom here, but a sort of shyness made her remain silent.

(178) *Ett riskfyllt steg som* ***höll på att*** *leda till inbördeskrig 1990, när emiratet Sharja ville legalisera alkohol.*
A risky step, which nearly resulted in civil war in 1990 when the emirate Sharja wanted to legalize alcohol.

(179) *Det finns inte en enda skylt som talar om att många barn rör sig här. Min dotter* ***höll på att*** *bli påkörd häromdagen.*
There is not a single sign saying that there are many kids around here. My daughter was almost hit by a car the other day.

Chapter 13

Pseudo-coordinations with *ligga, sitta, stå*

SAG names the lexical verbs *sitta (to sit), ligga (to lie), stå (to stand), komma (to come), gå (to go)* and *ta (to take)* as the constituting members of 'pseudo-coordinations' (*pseudosamordning*, [152], pp. 335f.) with other lexical verbs, conjuncted by *och (and)*:

(180) *Han* ***sitter och*** *läser.*
lit. *He sits and reads.*
He is reading.

These verbs are named together with the copula verb *vara (to be)* :

(181) *Han* ***är och*** *får sina tänder undersökta.*(SAG)
lit. *He is and has his teeth examined.*
He is having his teeth examined.

Viberg [161] uses the term *periphrastic progressive* for the coordinated construction with *gå*, but this seems somewhat underspecified. Therefore either the literal translation of *pseudosamordning – pseudo-coordination*, or *coordinated construction* will be used, in accordance with SAG. The construction *hålla på och verb-a* is also a pseudo-coordination (examples see Chapter 12).

The three spatial verbs (*sitta, ligga* and *stå*) have also been named by Pihlström [128] as progressivity markers along with *hålla på*. PAROLE as well as the parallel corpus was searched to investigate the degree of their grammaticalization and the semantic components activated in the respective concordances.

When determining the degree of grammaticalization, the following assumptions were made:

1. Progressivity is generally associated with location (cf. [75]). Thus the semantic component primarily activated in the grammaticalized uses of the three spatial verbs *ligga, sitta* and *stå* will be the one 'to be located somewhere'. This use

(unlike the verticality/horizontality opposition) normally requires a locative adverbial:

(182) **En katt satt. *There was a cat sitting.*

(183) *En katt satt på mattan.*
There was a cat sitting on the mat.

If a grammatically well-formed sentence with a coordination contains a locative adverbial, it is interesting to see whether the locative adverbial is governed by the first verb or by the entire pseudo-coordination. When the locative adverbial immediately follows the first verb (the spatial verb), followed by *och* and the second verb, it is apparently governed only by the first verb[1]:

(184) *Hon stod på balkongen och rökte.*
She was standing on the balcony, smoking.

There is no reason to call this type of combination a pseudo-coordination, as the surface syntax suggests that it is a regular coordination of two predicates. The first predicate, being used in the 'to be located somewhere' sense, has a complete valency frame and does not seem to be undergoing any semantic change. When, on the other hand, the locative adverbial is apparently governed by the entire coordination, being placed after the second verb, its use can be more grammaticalized than if the locative adverbial is governed only by the first verb, which is the reason why this sort of coordinations is called a *pseudo-coordination*.

Moreover, in the 'to be located somewhere' component, progressivity can also be anchored in the 'movement block' component of *stå, ligga, sitta*.[2] A sentence like 185

(185) *Eller fattade de inte, att han stod och gjorde sig rolig?*
lit. *Or did they not figure that he stood and kidded?.*
Or did they not figure out that he was kidding?

apparently does not require any location adverbial. Therefore sentences lacking location adverbials found in PAROLE comprise potentially many grammaticalized uses and they will be paid special attention here.

[1]The adverbs *här (here), där (there), nu (now)* and *då (then)*, when used as deictic markers, make an exception. Cf. Section 13.4

[2]Semantic components of *ligga, sitta* and *stå* have been thoroughly analysed by Jakobsson [71].

2. To show that the meaning of the first verb has become generalized, spatial verbs should be able to occur in pseudo-coordination with verbs denoting events for which spatial orientation is irrelevant, or even with verbs denoting events that themselves imply a different spatial orientation. However, in the latter case it must be clear from the context that the spatial verb is really not being used to indicate an unusual spatial orientation of the subject (as would be the case with e.g. *he was standing and sleeping*).

3. Another piece of evidence for generalized use is a subject that is not stereotypically associated with spatial orientation, typically an abstract entity. The opposition animate/inanimate does not really seem to show the generalization since all three spatial verbs are normally used with inanimate concrete subjects (though *sitta* only to a limited extent).[3]

4. Grammaticalized pseudo-coordinations are expected to be found in sentences with low facticity (cf. Section 12.5).

Corpus queries (below) have shown that pseudo-coordinated constructions with *sitta, ligga* and *stå* do not have as generalized distribution as *hålla på*. On the other hand, they can be used as progressivity markers with verbs denoting states and transitions, which *hålla på* cannot.

13.1 The Profile of *X står och verb-ar* in PAROLE

PAROLE returned 776 instances of *stå* with a morphologically congruent lexical verb separated by *och*. (This query consisted only of the uninterrupted sequence *stå och verb-a*.)

The Word Sketch for verb coordinations (see Section 15.3) displays the following groups of verbs as the most typical collocates of *stå och*:

- verbs of seeing
- verbs of speaking
- verbs of waiting

Seeing as well as speaking are human activitites typically associated with standing or sitting[4].

[3] *Sitta* with inanimate subjects requires a location adverbial as it is only used with inanimate subjects in the sense 'to be fastened somewhere'.

[4] The Word Sketch for *sitta* returns a very similar result: the first 20 most significant collocates of *sitta* are typically verbs of perception (seeing and listening), verbs of speaking and verbs of cognition (*to think, to meditate*), as well as verbs denoting typical sitting activities like sewing, eating and drinking, reading. Like with *stå*, waiting lies second with *sitta*.

The Word Sketch for *stå*in PAROLE contains a few apparent idiomatic expressions or frozen collocations, which are counterexamples of generalized use:

- *stå och stampa*:

  (186) *…i en tid när den japanska ekonomin **står och stampar.***
  *…in a time when the Japanese economy **is stuck.***

- *stå och falla:*

  (187) *…tidningen **står och faller** med läsarnas intresse.*
  *…the newspaper **depends on** the readers' interest.*

- *stå och väga:*

  (188) *Det är klart att Masters är roligare om det är två spelare som **det står och väger mellan**.*
  *Obviously The Masters is more fun when there are two players **to be decided between.***

  (189) ***Det står och väger** just nu, säger Patrick Englund . Pengarna i Schweiz eller den utbildning jag påbörjat här hemma.*
  *Right now **it depends**, says Patrick Englund. The money in Switzerland or the education I have started here at home.*

  (190) *Det **stod och vägde** en bit in på uppploppet men klassige Sleepwalker hade de bästa krafterna till slut…*
  *The final meters **were going to decide** for a while, but the excellent Sleepwalker [horse] had finally more power…*

- *stå och dra/småputtra/svälla:*

  (191) *Låt det hela **stå och dra** i kylskåpet minst 1 timme.*
  *Let it [the meal] **stand** in the refrigerator at least for 1 hour.*

  (192) *Låt smeten **stå och svälla** 15 minuter.*
  *Let the dough **expand** for 15 minutes.*

  (193) *…blir soppan bara godare ju längre den får **stå och småputtra***
  *…the soup becomes all the better the longer it can **simmer***

The inspection of all 776 concordances returned three more collocations of this type: a door or a window can 'stand and flap' in the wind, old useless stuff can 'stand/lie and be trash' and people who are not so sick that they are only able to lie down can 'stand and go'

(194) *…tak sjunker, rutor är trasiga , en dörr* ***står och slår****.*
…the roofs are sinking, the window panes are broken, a door is slamming.

(195) *Har du en gammal radio- eller TV-apparat som* ***står och skräpar*** *, så tag kontakt med Radiomuseet.*
If you have an old radio or a TV that ***stands in your way****, you should contact the Broadcasting Museum.*

(196) *På var sida om honom stod fem eller sex slavpojkar – alla […] som hade krafter att stå och gå.*
On each side of him stood five or six young male slaves – all of whom could stand on their feet.

The pseudo-coordination *stå och verb-a* was found a few times with telic verbal clauses, resulting in an iterative meaning, e.g.:

(197) *Alla som stått och kastat på en vakande fisk vet att den inte hugger på vad som helst.*
Everybody who has ever been casting for a fish that is awake knows that it would not bite into just anything.

It can even be used with verbs denoting transitions, to emphasize the process and to neutralize the inherent terminal point of the event in question:

(198) *den nya mjölken utanför dörren hade stått och surnat.*
the fresh milk in front of the door had been going sour.

Unlike *hålla på att*, neither *stå och* nor *ligga och* in the sense 'it nearly happened' was found in the concordances.

Out of a random sample of 100 instances (a subset of the 776), 75 lacked locational adverbials. A closer analysis revealed that this *stå* was rather the opposite of *to go* than to *to sit* or *to lie*, which suggests that the movement-blocking anchoring is a very commonly activated semantic component.

Typical instances of activated movement-blocking follow:

(199) *Med spelad nonchalans sparkar hon med fötterna mot trottoarkanten. — Här kan vi inte stå och ruttna bort.*
Withfeigned nonchalance she kicks the edge of the kerb: We can't be rotting here.

(200) *Det kostar pengar att låta taxin stå och vänta.*
It costs money to have the cab wait.

(201) *den kloke gubbe, som stod och sa att min mors plötsliga död var förutbestämd.*
the clever man who was asserting that my Mother's sudden death was not an accident.

The movement-blocking *stå* combines with four sorts of verbal collocates:

1. verbs denoting activities typically associated with standing, e.g.:

   (202) *Många kvinnor satt vid symaskinen eller stod och strök.*
   Many women were sitting at sewing machines or were ironing.

2. verbs typically associated with standing or sitting. The speaker either knows the relevant position or he infers the most stereotypical one for the given context:

   (203) *Nu vill hon inte ens stå och vänta på hissen tillsammans med honom.*
   Now she would even hate to be waiting for the lift with him beside her.

   (204) *Det är en skön avkoppling att sitta och vänta på att fisken ska nappa.*
   It is a pleasant relaxation to be (sitting and) waiting for the fish to be caught.

   Even if we do not know anything about what the building looks like in the first sentence, we would guess that there are no chairs in front of the lift, since the stereotype is that one is standing when waiting for a lift. On the other hand, fishermen are typically sitting when waiting for fish to get caught.

3. verbs denoting events for which spatial information is irrelevant (even though it is known or it could be inferred from the context):

   (205) *Kyrkans roll har i detta läge blivit att stå och bjuda ut sin religion*
   In this situation the church's role had become to keep promoting their religion

   (206) *Eller fattade de inte, att han stod och gjorde sig rolig?*
   lit. *Or did they not figure that he stood and kidded?.*
   Or did they not figure that he kept on kidding?

(207) *Skäms inte, Fräulein! Varför står ni och skäms?*
Don't be shy, miss! Why are you shy?

(208) *Vad står du och tänker på?*
What are you thinking about?

4. verbs denoting events that themselves imply a certain spatial orientation which is apparently incompatible with the one provided by the spatial verb:

(209) *Och kommer inte att tänka på det förrän alla utestängda rejvare står och dansar precis där jag vill dansa.*
And it won't occur to [me] until all the ravers outside are dancing exactly on the spot where I want to dance.

(210) *När doktorn kom stod jag och hoppade i sängen för att få in luft i lungorna*
When the doctor arrived I was jumping up and down on my bed to get some air into my lungs

Sometimes a preference for *stå* or *sitta* or *ligga* results from activating a more subtle component of the given verb's stereotypical image, which has to do with the 'movement-block' sense. e.g. *to sit* and *to stand* with human subjects is often associated with laziness, resignation, lack of enterpreneurial spirit or a temporary inability to react:

(211) *Jag är glad att du inte bara sitter och väntar på att hon ska ringa*
I am pleased that you don't just sit and wait for her to give you a call

(212) *När någon går mig på nerverna , bara står och glor så där eller stirrar snett kan han få en örfil eller något .*
lit. *When someone irritates me, just stands and stares so or looks down at me, he may get a blow or something.*

Standing as well as *lying* is possibly associated with things being untouched by human hand (i.e. unused)) for a long time:

(213) *Det visar sig att kranen i badrummet ovanpå har stått och droppat och att vattnet runnit ner genom den trasiga badrumsmattan.*
It becomes obvious that the tap in the bathroom upstairs has been leaking and water has been running down through the worn-out bathroom mat.

(214) *Äpplena ska inte behöva ligga och ruttna på fälten längre.*
The apples should not be rotting on the fields any more.

Constancy in *stå och verb-a*

Unfortunately PAROLE returns too little evidence of negative imperative clauses and questions with the pseudo-coordination with *stå* to make it possible to decide whether this construction has the same 'constantivizing' power as *hålla på*, though some few instances suggest that it possibly could:

(215) *Eller fattade de inte, att han stod och gjorde sig rolig?*
lit. *Or did they not figure that he stood and kidded?.*
Or did they not figure that he kept on kidding?

(216) *Med spelad nonchalans sparkar hon med fötterna mot trottoarkanten. — Här kan vi inte stå och ruttna bort.*
With feigned nonchalance she kicks the edge of the kerb: We can't be rotting here.

(217) *Men hon kunde ju inte stå och kamma sej hela natten . – Hej , sa hon igen och klev in i vardagsrummet.*
But she couldn't anyway keep brushing her hair all night. – Hi, she said again and entered the living room.

Also, in *stå och verb-a* when used as illustrated in 215, 216, and 217, the semantic component activated is most likely the 'movement block'. The predominance of the movement-block component resembles in a remarkable way the progressivity-constancy shift described with *hålla på* – cf. Section 12.5.

13.2 The Profile of *X sitter och verb-ar* in PAROLE

There are 922 instances of pseudo-coordination in PAROLE (uninterrupted sentence). Predictably, all of them have human (incl. institutions) or at least animate subjects (a butterfly and birds, in total just two concordances). They form neither idiomatic expressions nor frozen collocations as *stå* does, except perhaps *sitta och hänga*[5]. The list of the most typical collocates given by the Word Sketch shows a balanced distribution of *sitta* among verbs denoting events stereotypically associated with sitting like handicrafts, eating, drinking, listening to music, talking, performance watching etc.

[5] *sitta och hänga* or even *stå och hänga* is an emphatic variant of *hänga*, in the sense of spending time in one place doing nothing. Typically, a youth is hanging around in a pub or at a railway station, etc.

Unlike *stå*, PAROLE yielded no verbs that would denote events in which incompatible spatial orientation would have to be integrated. A tentative search in Google gave about 6 000 occurrences of *satt och åkte*. The human agents were typically sitting in cars, driving. Example 218 even indicates that the emphasized semantic component of the event of driving was constancy (i.e. the ride took a long time) rather than progressivity, Cf. Section 12.5):

(218) *Efter racet slängde jag mig i bilen och satt och åkte hela vägen ner till Schweiz.*[6]
lit. *After the race I hopped into the car and sat and drove all way down to Switzerland.*
After the race I hopped into the car and kept driving all way south to Switzerland.

Another verb apparently semantically incompatible with *sitta* tested in Google was *hoppa*. Google returned about 2000 instances:

(219) *Jag satt och hoppade på stolen i ren nervositet!*[7]
lit. *I sat and hopped in my chair of genuine nervosity.*
I was hopping in my chair of genuine nervosity.

A few PAROLE concordances comprise verbs denoting events for which spatial orientation is irrelevant:

(220) *Du behöver inte sitta och tro att jag inte gör det.*
lit. *You don't have to sit and believe that I don't.*

(221) *Du sitter och vet allting …*
lit. *You sit and know everything…*

13.3 The Profile of *X ligger och verb-ar* in PAROLE

The uninterrupted sequence *ligga och verb-a* was found in PAROLE 258 times. The most typical verbal collocates of *ligga* are *sova (to sleep*, 37 occurrences), *flyta (to flow*, 12 occurrences) and *lyssna (to listen*, 9 occurrences).

Out of the three spatial verbs *ligga* is most commonly used with non-human subjects. Some verbs that ranked high in the Word Sketch for *ligga* are typically used with non-human subjects: *flyta (to flow)* – 2nd, *skräpa (to be somewhere as trash)* – 3rd, *gro (to sprout)* – 4th, *ruttna (to rot)* – 6th.

Ligga is most typically used with atelic verbal phrases to mark progressivity:

[6]quoted from Google, 2006-09-23, URL<www.powerbar-europe.com – 25k>
[7]quoted from Google, 2006-09-23, URL <typtrettio.blogs.se/2006/02/ – 61k>

(222) *De flesta av de omkomna var män som låg och sov vid kollisionen.*
Most victims were men who had been sleeping at the moment of the collision.

(223) *alla väskor och byxor som låg och skräpade på bänkarna*
all the bags and pants that had been lying as trash on the banks

Some contexts indicate constancy (see Section 12.5):

(224) *Nu har vi kört en tuff serie, har bra tempo och Dackarna har ju legat och fått stryk hela året, säger Tommy Rander till GP .*
Now we [a sports team] have completed a tough series, we have a good rhytm while Dackarna [another sports team] have been getting licked for a year, says Tommy Rander for GP.

(225) *Idén till den showen har legat och grott i två år och nu funderar Nicklas redan på nästa drömprojekt.*
The idea of the show has been sprouting for two years and now Nicklas is already thinking of the next dream project.

(226) *Äpplena ska inte behöva ligga och ruttna på fälten längre.*
The apples should not be rotting on the fields any more.

Used with telic verb phrases, the pseudo-coordination seems to be neutralizing the inherent terminal point of the event:

(227) *Medan han låg och återhämtade sig, vilade han ögonen på Lauritz gåva.*
While he was recovering, he kept looking at the present from Lauritz.

(228) *Man kände det som om man låg och tog upp plats för en massa andra som ville leva men blev livsfarligt sjuka ändå.*
One felt oneself to be taking the place of many others who wanted to live but became seriously ill anyway.

The verb *ligga*, when used with human subjects, is used in contexts typically associated with horizontal position: sleeping, listening to music/rain drops in bed, sighing and weeping alone at night etc. No instances were found in PAROLE where the second verb would denote an event that itself is associated with a different spatial orientation. Example 224 only shows a somewhat metaphorical expression of a team's long-lasting bad results by the stereotypical picture of a kid lying on the father's knees being whipped.

13.4 Deictic Markers in Pseudo-coordinations

Pseudo-coordinations, mainly in direct speech, are optionally emphasized by the deictic adverbs *nu (now), här (here)*, and *där (there), då (then)*:

(229) *Men här står vi och pratar och jag ber dej inte ens stiga fram och sitta – du är väl hungrig kan jag tro?*
lit. *But here we stand and chat and I don't even ask you to come in and sit down – you must be hungry, aren't you?*

(230) *Nu står du och drömmer igen, brukade flickorna reta mej.*
lit. *Now you stand and dream again, the girls used to tease me.*

(231) *Hon är elak ! Och där står hon och vill att jag ska förstå henne.*
lit. *She is evil! And there she stands and wants that I shall understand her.*

(232) *Nej här sitter jag och pratar en massa om mig själv…*
Oh no, I keep talking [too much] about myself…

The usage of deictic markers in Swedish spatial pseudo-coordinations deserves one additional remark related to Czech: spoken Czech can make use of pseudo-coordinations with *ligga, sitta* and *stå*, too. The insertion of a deictic locative adverb (*tady* and *tam* – *here* and *there*) is also extremely frequent. However, in Czech the deictic markers are even frequently used with simplex verbs, which is not acceptable in Swedish:

(233) *Ty tady vyspáváš, zatímco já musím pracovat!*
lit. *You here sleep while I have to work!*

(234) *Du ligger och sover medan jag får arbeta!*
Här/Nu ligger du och sover medan jag får arbeta!
***Du sover här medan jag får arbeta!**

The locative anchor in progressive meanings is evident in Swedish as well as in Czech, yet the deictic markers in Swedish seem to be only a supplement of progressive constructions. Czech speakers have to learn this to avoid errors in Swedish language production.

The following example comes from the Swedish translation of a text by Bohumil Hrabal. The Swedish translator preserved the deictic marker *here* but had to place the verb in a pseudo-coordination with *sitta*.[8]

[8]The lexical verbs *skotačit* and *prata strunt* are not equivalent. The translator either used a different text version, or he was simply wrong. The Czech verb *skotačit* means *to frisk*, while the Swedish verb

(235) *a já* ***tady*** *v sedumdesáti letech s vámi skotačím, jak císař se Šratovou…*
och ***här sitter*** *jag sjutti år gammal* ***och pratar strunt*** *med er, precis som kejsarn med Schrattskan…*
and here I go, 70 years old, dancing/chatting with you, like the Emperor with the Schrattwoman…

In the examples of Czech translations from Swedish originals the spatial verbs are missing, since they are not obligatorily used with deictic markers in Czech.

(236) *A ty mi* ***tady*** *budeš vykládat něco o domově důchodců?*
Och du ***kommer hit och talar*** *om ålderdomshem.*
And you are talking to me about an old people's home?

(237) *Poslyš, co mě* ***tady zpovídáš*** *o věcech, které snad všichni víme.*
Hör nu, varför ***står du här och frågar*** *om saker som vi allesammans vet?*
Listen, why are you questioning me about things we all are familiar with?

This is another hint for translations from Swedish to Czech: spatial pseudo-coordinations are probably used less frequently in Czech than they are in Swedish. A longer translation from Swedish into Czech, in which all spatial pseudo-coordinations were literally translated, would possibly raise the native speaker's attention, even if there had occurred no instances of (Swedish) second verbs denoting events incompatible with the spatial orientation expressed by the first verb or spatially irrelevant events. Even Swedish combinations of a spatial verb and a second verb for which the spatial orientation expressed by the first verb is typical can have various equivalents in Czech. Pseudo-coordinations are often replaced by deictic markers and manifold marked imperfectivizing verb derivations (e.g. *spát (to sleep) – vyspávat (approx. to be outsleeping)*), or even just by using an unmarked imperfective verb form.

A search in the Czech National Corpus (SYN2000) revealed that pseudo-coordinations are frequent mainly in colloquial Czech, but no cases of two semantically incompatible verbs in pseudo-coordination have been observed. However, a more detailed study of this issue would require creating a subcorpus by eliminating translations from SYN2000 and comparing matches of the following queries:

1. deictic markers – lexical verb – *a (and)* – lexical verb
2. no deictic marker – lexical verb – *a (and)* – lexical verb
3. deictic markers – lexical verb – no *a (and)* – no verb

prata strunt means *to talk bullshit*. Speaking as a social act, like in a conversation, is intuitively associated with sitting, and that is why sitting has formed the pseudo-coordination. Nevertheless, the lacking correspondence of these two lexical verbs in these two parallel texts does not invalidate the legacy of this general observation!

The first verbs in pseudo-coordinations would not be confined to just *sedět (to sit), stát (to stand)* and *ležet (to lie)* since they can be replaced with colloquial variants (e.g. *dřepět*), which are likely to occur. As the queries are very general, the processing of the concordances would be very time-consuming. Relating deictic markers to pseudo-coordinations in Czech and a quantitative comparison of their frequencies in Czech and Swedish goes beyond the scope of this thesis.

Chapter 14

Pseudo-coordinations with *ta*

The verb *ta (to take)* is normally (i.e. in its most cognitively salient use) a transitive verb. Changes in the prototypical valency frame may suggest possible semantic changes indicating the process of grammaticalization in the verb *ta*. The assumption is that *ta* keeps its original transitive valency frame in regular predicate coordinations, but that it possibly can become intransitive when used to modify the event structure of another verb, with which it creates a pseudocoordination.

1. The verb *ta* has an object and the second verb has another object.

   (238) *grupper som tar gisslan och utövar utpressning*
   groups that take the hostage and exploit him

2. The verb *ta* has an object which is obviously not shared by the second verb.

   (239) *Hon tog brickan och gick uppför trappan.*
   She took the tray and went upstairs.

3. The verb *ta* has an object and the second verb formally has a direct object but the object of the second verb actually refers to the direct object of the first verb.

   (240) *Jag tog badrumsmattan och la den i badkaret.*
   I took the bathroom mat and laid it into the bath tub.

4. The verb *ta* as well as the second verb have a generalized direct subject.[1]

   (241) *en förmåga att ta och ge.*
   an ability to take and to give.

[1]This is actually a subset of 3.

A second verb and the verb *ta* in coordination form a grammaticalized pseudo-coordination when:

1. Neither the verb *ta* nor the second verb governs any direct object.

   (242) *Nej, nu måste jag nog ta och gå, sa så Ulla och såg bekymrat på sin klocka.*
   lit. *No, now I must well take and go, …*
   Oh no, I certainly have to be leaving now, Ulla said and took a worried look at her watch.

2. The second verb governs a direct object which is not shared by the verb *ta*.

   (243) *Jag får ta och ringa dom.*
   lit. *I have to take and phone them.*
   I have to give them a call.

   (244) *Ska vi inte ta och lägga oss?*
   lit. *Should we not take and lay us?*
   Shouldn't we go to bed now?

   (245) *Det ska jag faktiskt ta och fråga honom om.*
   lit. *That will I actually take and ask him about.*
   This is actually what I will ask him about.

Some contexts can be ambiguous as to whether the direct object is governed by the second verb or by the entire pseudo-coordination[2]:

(246) *Jag ska ta och linda foten.*
lit. *I will take and wrap the foot.*

(247) *Här, ta och stoppa den [lax] i kylen!*
lit. *Here, take and put it [salmon] in the fridge!*

14.1 The Profile of *X tar och verb-ar* in PAROLE

There are 96 occurrences of *ta and verb-a* in PAROLE, of which approx. 12 were ruled out as irrelevant (mostly the *ta och ge* examples and *ta* as a light verb in coordination with another predicate). Most of the relevant hits were direct speech or free indirect speech, which suggests that this construction is still rather colloquial.

It typically occurs with telic events[3].

[2]This occured approx. 5 times in PAROLE, i.e. in 20% of cases

[3]processes as well as transitions

(248) *Skulle du inte ta och sätta dig?*
Won't you sit down?

(249) *Man börjar fundera på om man inte skulle ta och hyra Den lilla sjöjungfrun på video.*
One starts to think about whether one shouldn't get The Little Mermaid hired on video.

The pseudo-coordination *ta och verb-a* also bears a distinct flavour of sudden or spontaneous intentionality.

(250) *…men jag som är så dålig i gympa får väl ta och bli något annat.*
…but as I am so bad at sports I will have to become something else [i.e., not a sports teacher].

Sometimes the lexical verb denotes a prototypically non-intentional event (e.g., a transition as *bli*, see 250), or the agent can be non-volitional. PAROLE even yielded one concordance with a non-human agent (251):

(251) *Och handlarn bara borrade in sina gräsögon i hans, han kände hur det tog och vred om, tills allting började snurra. Han hade glömt bort vad han skulle köpa!*
lit. *…. he felt how it took and twisted until everything started to buzz….*
And the storekeeper just drilled his grass-coloured eyes into his; he felt how the world came to twist until everything was buzzing. He had forgotten what he was going to buy!

Ta och vrida in combination with a non-volitional agent in Example 251 emphasizes the sudden beginning of the event.

Verbs that denote atelic events are also acceptable in the pseudocoordination with *ta*, but they are far less frequent (3 in PAROLE). They express the beginning of an atelic event.

(252) *Om man skulle ta och följa efter någon på kul.*
If one should start to follow a person just for fun.

(253) *Om man skulle ta och städa…*
If one should start to tidy up…

(254) *Nej , ska vi inte ta och prata om nåt annat.*
No, shouldn't we start to talk about something else…

14.2 Czech Equivalents

Most typically, the pseudo-coordination has no explicit Czech counterpart with telic events. Similarly to the pseudo-coordinations with *ligga, sitta* and *stå*, *ta och verb-a* occurs in the Swedish translation of a colloquial text.

(255) *Vy máte čas, tak pojedete s mojí paničkou na nádraží*
hej! har ni tid så ta och åk med min fru till stationen
You have time, so you will take my wife to the station

(256) *a podepište revers*
ta och skriv under en revers
and sign the declaration

(257) *vyberte si tam lepší snění, jo?*
ni kan väl ta och välja ut en bättre dröm?
will you please select another dream?

(258) *mi pošeptala u gramofonu na Žofíně, půjdeme spolu na rande*
…viskade till mej vid grammofonen, vi tar och stämmer träff
She whispered into my ear at the grampohone at Žofín: we will have a date

Ta och verb-a with an atelic (or at least not unambiguously telic) event triggers perfective verb form in Czech translation.

(259) *Vi ska kanske ta och studera vännen Eriksson lite närmare*
Asi bychom si měli trochu zblízka posvítit na přítele Erikssona
We should perhaps take a closer look at our friend Eriksson.

Part IV

Implementation of Swe-VALLEX/PNL

Chapter 15

Preparatory Work

15.1 Organizing Lexical Sources and Tools

Lexicographers typically point out routine being one of the most distinctive features of their work. The aim of the present study is to leave the Swe-VALLEX/PNL lexicons just on the verge of becoming such an efficient routine. However, an efficient routine arises from a painstaking adjustment of the working conditions [4]. Without a comfortable access to data and a reasonable pre-sorting of concordances it would be hard to adequately reflect the linguistic issues discussed in the previous sections. Therefore this entire section is dedicated to the corpus acquisition and corpus adjustment.

Despite its technical focus, this section can be regarded as somewhat 'crypto-linguistic'. Surprisingly much linguistic reasoning is hidden between the lines of script files, which seek to make the subsequent lexicographical routine

- **comprehensive**
- **fast**

…and **reproducible** in the best possible way.

15.2 Preparing the Corpora

Using corpora in one or the other way[1] has become a standard in dictionary making. The following corpora were eligible to work with:

- PAROLE [121] – Swedish monolingual corpus with automatic morphosyntactic tagging, not lemmatized. Size: Approx. 19,000,000 tokens.

[1]We will not go into differences between 'corpus-based' and 'corpus-oriented' work.

- SUC [35] – Swedish monolingual corpus with manual morphosyntactic tagging and lemmatization. Size: 1,000,000 tokens.
- SYN 2005 [33] – Czech monolingual corpus with automatic morphological tagging and lemmatization. Size: more than 100,000,000 tokens.
- Intercorp [70] – a set of nearly 30 parallel corpora with Czech as the pivot language. It is still under construction, currently with approx. 25,000,000 tokens. The Czech-Swedish pair comprises about 2,000,000 tokens. No tagging or lemmatization is provided for this language pair.

At the beginning of this work in 2005, the parallel Swedish-Czech corpus Intercorp was very embryonic and too small to deliver any linguistic evidence. However, with its current 2,000,000 tokens it yields interesting hits, which inspire further explorations of the large monolingual corpora PAROLE and SYN2005.

The most important corpus to work with has been PAROLE from the very beginning. Initially it was only accessible online on the web (http://spraakbanken.gu.se/parole). Later it became available for downloading. SUC was possible to download right away, but it was too small compared to PAROLE and no browser was attached to it. Therefore PAROLE was preferred. The download permission 2003 and its kind hosting by the Institute of the Czech National Corpus (Faculty of Arts, Charles University) enabled searching PAROLE with the corpus manager BONITO [142]. Yet the real breakthrough in exploiting PAROLE came in 2004, thanks to the collaboration with Pavel Rychlý (Faculty of Informatics, Masaryk University, Brno), who declared himself ready to host PAROLE in Brno and to make it browsable with his new version of BONITO, in which the Word Sketch Engine [78] had recently been integrated.

15.3 Word Sketch Engine and Collocation Analysis

The Word Sketch Engine (sometimes just called 'Sketch Engine') is a tool for preliminary concordance sorting. The Word Sketch Engine generates **Word Sketches**. Word Sketches are 'one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour' [78]. They were first used in the production of the Macmillan English Dictionary [38] to speed up the lexicographer's work by preliminary sorting of concordances according to their collocational salience, which was examined with respect to various grammatical relations between the keyword and its potential collocates. To give an example from [78]: when the word *pray* is queried, the Sketch Engine will look for prepositions that it precedes and for words that most typically follow these prepositions in combination with *pray*. Moreover, it will look for adverbs that most typically occur together with *pray*, etc. It will return the results

in form of a table where the items (collocates) are clickable. Thanks to this feature, the user can have a quick look at only those concordances that contain the given collocate the user had clicked on in the table. Besides Word Sketches, the Sketch Engine provides the user with additional functions like Thesaurus and Sketch Differences. For more information see e.g. [78].

The Word Sketch Engine is not the very first collocation sorter invented. The huge half-billion corpus of modern written German, built at the Institute for German Language in Mannheim, Germany, has had a built-in automatic collocation analysis since 1995 [7], which has, however, never been explicitly offered for use with other corpora.

15.4 Adjusting the Word Sketch Engine for PAROLE

To be able to use the Sketch Engine with a particular corpus, the corpus must have a specific format. Each word must be on its own line, followed by its tag and its lemma. It is strongly recommended to have the corpus lemmatized since the statistical tools inside the Sketch Engine have been designed to operate with lemmas, not word forms. The Sketch Engine itself does not support lemmatizing.

When the corpus is prepared in an input format appropriate for the Sketch Engine, the Sketch Engine must acquire the knowledge of syntactic relations within sentences. This is to be done by computing 'gramrels', grammatical relations. The procedure will be briefly commented below.

15.4.1 Lemmatization

In 2005 and 2006, no existing Swedish-made lemmatizers were possible to obtain due to unclear legal conditions, nor were any electronic lexicons available[2]. The last resort was creating a make-do rule-based lemmatizer for the Sketch Engine. This was only enabled by the collaboration with Jan Pomikálek (FI MU, Brno), who wrote the initial lines of the sed script and the entire complementary Perl script. It was also him who performed the final evaluation of the results on SUC. This section presents the basic features of LEMPAS, our rule-based lemmatizer for the Swedish PAROLE corpus [27].

The linguistic task to be processed by the Sketch Engine only required the lemmatization of nouns and finite verbs. Besides, we added a fuzzy lemmatization of adjectives; i.e. we lemmatized only tokens with the following tags: NC.* (common nouns),

[2]A very precise statistical lemmatizer (94,72% accuracy) had existed [52], though unpublished, and thus undetectable. It could only have been found by Google search as 'tagger', but with no hint to the lemmatizing feature.

A.* (adjectives and participles) and V.* except V@S.* and V@M.* (verbs except imperatives and conjunctives). We also systematically ignored numbers (M.*), proper nouns (NP.*), pronouns (P.*) and adverbs (R.*).

LEMPAS comprises a sed script and a complementary Perl script. The sed script gathered related inflection forms, while the Perl script corrected the pre-lemmas to comply with headwords. The lemmatization is tag-dependent. We used a simple regular expressions syntax to build the basic lemmatization rules. This made the implementation straightforward enough to be performed by a linguist with very limited computer skills.

The Sed Script

The data of the PAROLE corpus to be processed with the Sketch Engine had the following structure: one token per line, followed by a tag separated by a tab. A lemma string was to be inserted in between. For example, the token *katterna* ('cats-the') would have looked like *katterna NCUPN@DS* and would be replaced by *katterna katt NCUPN@DS*. The "find-replace" sed-structure hosts linguistic rules that decompose the "word form" string into segments to be preserved, omitted or modified in the "lemma" string, which is newly created by the replacement. The next sections describe the linguistic rules in more detail.

Rules for Nouns

The script groups the rules approximately according to declension types as listed by Nylund and Holm [112]. Many declensions go across genders; therefore genders can only be read from tags.

Many declensions contain an additional rule that affects the indefinite singular. This rule chopped off the last character from stems ending with *-e*. As the rules for definite singulars and both plurals were unable to determine whether the lemma was supposed to end with *-e* or not, we decided to pretend that Swedish had no words ending with *-e*, even in derivation affixes like *-else*. Due to this rule, the word *stavelse* got the lemma *stavels* in all forms. Table 15.1 lists some examples of the basic rules for nouns.

The rules for genitives are identical with the rules for nominatives except for the *-s* added to the respective endings and G replacing N at the case position. We neglected words ending with *-s*, *-z*, and *-x*, which do not attach *-s* in the genitive. We assume that genitives of such words would mainly occur in proper nouns, which lemmatization ignores in general.

On the other hand, we paid attention to types of common nouns that do not fit the basic rules.

One of the important rule restrictions was applied to nouns whose plural forms and definite singular form in neuters follow a stem *l*, *r* or *n*, which is neither duplicate

Table 15.1: Examples of lemmatization rules for nouns

ending	**tag**	**rule**	**example**
–an	NCUSN@DS	–n deletion	flickan → flicka
–or	NCUPN@IS	–or deletion, a-addition	flickor → flicka
–ar, –er	NCUPN@IS	–ar, –er deletion	katter → katt, stolar → stol
–en	NCUSN@DS	–en deletion	katten → katt, stolen → stol
–[iouyöää]n	NCNPN@IS	–n deletion	hjärtan → hjärta, möten → möte
–[iouyöää]t	NCNSN@DS	–t deletion	hjärtat → hjärta, mötet → möte
–n	NCNPN@DS	–n deletion	husen → huse
–et	NCNSN@DS	–et deletion	huset → hus
–[iouyöää]n	NSUSN@DS	–n deletion	byrån → byrå

(*stället – ställe*) nor preceded by a vowel (*signaler – signal*): *muskler*, *mörkret*. These nouns have usually dropped their *e* in the indefinite singular: *muskel*, *mörker*. We have thereby ignored words that originally had no *e* in their stem. So far we have found and listed the counterexamples *moln*, *karl*, *kärl*, *sorl*, *porl*, *regn*, *ugn*, *agn*, *vagn*, *lögn*, *stygn*, *lugn* and *dygn*. We also listed the noun *morgon* (plural *morgnar*).

We add an *m* to lemmas with the singular definite ending *–mlen*, *–mlet*: *himlen – himmel*, *skramlet – skrammel*. We also add an *m* to lemmas with the *–ar* plural endings following an *m* that is neither duplicate (*dammar – damm*) nor preceded by a vowel (*kramar – kram*): *kamrar* → *kammare* (*somrar* → *sommar* is listed). This rule ignores nouns ending with *–mer* in the singular indefinite as the plural forms of these nouns (*glimmer*, *flimmer*, *bekymmer*) only occurred as tagging errors. Another m-rule chops off *m* in neuters in which the definite endings follow a double *m* (*hemmet* → *hem*, *programmet* → *program*). Only suggesting the *l*, *r*, *m*, *n* subtypes, this text does not list rules for all inflection forms, though they are present in the script.

Another problematic group of nouns were loan neuters with *–er* ending in plural, which can have two different suffixes in singular: *–ium* vs. *–eri*, e.g. *podier* → *podium* vs. *skafferier* → *skafferi*. Our rules were able to correctly lemmatize nouns ending with *–orier*, *–arier* and *–eer*. The types *mysterier* → *mysterium* and *podier* → *podium* could not be distinguished by rules and were therefore processed by the Perl script. There was a similar problem with plural uters ending with *-ier* (*serier* → *serie* vs. *harmonier* → *harmoni* vs. *historier* → *historia* vs. *irakier* → *irakier*). To a certain extent, these types have also been resolved by the Perl script.

Table 15.2: Examples of lemmatization rules for verbs

ending	**tag**	**rule**	**example**
–ar	V@IPAS	–r deletion	klarar → klara
–er	V@IPAS	–r deletion, –a insertion	läser → läsa
–[öäiouåy]r	V@IPAS	–r deletion	syr → sy, tror → tro

Rules for Verbs

Irregular, modal and auxiliary verb forms have been listed. Conjunctives and imperatives have been ignored by lists as well as by rules. Examples of rules for present active forms of regular verbs are given in Tab. 15.2.

The rules for deponential forms are almost identical with the rules for active forms, just an *–s* suffix follows the conjugated form. We only consider the *–s* suffix, and disregard the alternative *–es* suffix.

Rules for Adjectives

The tagsets of both PAROLE and SUC regard participles as a subset of adjectives. The present participle forms differ only in case (nominative vs. genitive) and the nominative is the lemma (e.g. *asylsökandes – asylsökande*, *oberoendes – oberoende*). Nominative forms were copied into the lemma and *–s* was deleted from the genitives.

The lemma of a perfect participle is the indefinite singular uter nominative (e.g. *avklarade – avklarad*) as is found in adjectives. Perfect participles of irregular verbs acquire different endings than those of regular verbs. The genitive is marked by the *–s* suffix. Rules are set for all genders (incl. masculine), both numbers, and for both definite and indefinite forms. They cover the following types: *klarad*, *berörd*, *köpt*, *ansedd*, *bruten*, *välkommen*.

Rules for adjectives include gradation, itemizing the commonest irregular forms, and seeking to cover systematic stem changes like the noun rules.

The Perl Script

As mentioned above, some of the rules only produce pre-lemmas rather than real lemmas. To fix this, we used a simple Perl script for postprocessing. The script operates in two steps. In the first step, the whole corpus is read and a set of dictionaries is built of words meeting certain conditions. Then the corpus is gone through again and some of the lemmas are modified according to the dictionaries.

The following strategy is used when building the dictionaries:

if word ends with *ending* **and** its tag equals to *tag* **then**

Table 15.3: Rules for building dictionaries

dict #	*ending*	*tag*	*delete-ending*	**example**
1	–e	NC.SN@IS	–e	möte, add möt
2	–ia	NCUSN@IS	–a	historia, add histor
3	–um	NCNSN@IS	–um	podium, add podi
4	–ier	NCUSN@IS	–er	irakier, add irak
5	–are	NCNSN@IS	–are	läkare, add läk

add the word to the dictionary with the *delete-ending* deleted

The values of the parameters for each of the dictionaries are listed in Tab. 15.3. The rules for modifying lemmas according to the dictionaries include:

if pre-lemma in $dict_1$ **and** tag matches *NCU[SP][NG]@[ID]S* **then**
lemma := pre-lemma + *–a*

if pre-lemma in $dict_2$ **and** tag matches *NCN[SP][NG]@[ID]S* **then**
lemma := pre-lemma + *–um*

if word ends with *–arna* **and** tag=*NCUPN@DS* **and** word base in $dict_3$ **then**
lemma := word base + *–are*

if word ends with *–arnas* **and** tag=*NCUPG@DS* **and** word base in $dict_3$ **then**
lemma := word base + *–are*

if pre-lemma in $dict_4$ **and** tag matches *NCU[SP][NG]@[ID]S* **then**
lemma := pre-lemma + *–er*

Example. The *e*-eliminating sed rule lemmatizes the singular indefinite nominative noun *möte* as *möt*, which unifies it with the inflected forms *mötet*, *mötes*, *mötets*, *möten*, *mötens*, *mötena*, *mötenas*, which all have been lemmatized as möt by the basic rules. The Perl script detects the difference between the word string and the lemma string in the indefinite nominative singular (*möte NCNSN@IS möt*) and includes *möt* in the dictionary as $case_1$. Then it searches all of the tokens for this string in the lemma and groups the relevant ones under the $case_1$-parameter. Finally, the lemmas of all inflection forms of the word *möte* get *–e* attached, i.e. the lemma is corrected in all inflection forms.

The Perl script can neither resolve homonymous forms nor orthographic variants. It simply selects the more frequent indefinite singular form to be the lemma. For example, the noun *herr* ('Mr.') will be lemmatized as *herre* due to the predominant form *herre* ('sir', 'Lord'). Initially, one single occurrence of the given word string ending with *-e* was enough to include the word in the dictionary. However, rare archaic

Table 15.4: Lemmatization results evaluation

	all words			unique words		
	correct	errors	accuracy	correct	errors	accuracy
common nouns	220814	13900	94.08 %	56939	5598	91.05 %
finite verbs	113452	22006	83.75 %	7707	1123	87.28 %
adjectives	56066	16670	77.08 %	10742	1426	88.28 %
present participles	5414	8	99.85 %	1462	7	99.52 %
perfect participles	11059	1431	88.54 %	4348	532	89.10 %
all	406805	54015	**88.28 %**	81198	8686	**90.34 %**

and colloquial word forms turned out to harm the lemmatization. For example, the extremely common noun *hus* ('house') was lemmatized incorrectly (with *-e*) due to its archaic inflection form *huse*, nowadays only used in the phrase *man ur huse* ('altogether')! To avoid this problem, the script was enhanced with a frequency comparison of *e*-ending vs. non-*e*-ending indefinite singular nominative word strings.

Besides that, forms lacking their singular indefinite counterpart in the corpus are never lemmatized correctly. This is often the case of occasional compounds as well as group names such as *irakier* and *indier*, which typically occur in the plural.

Results Evaluation

The manually lemmatized SUC was used for evaluating our lemmatizer. As already mentioned, we focused on lemmatization of some parts of speech only. When evaluating the results, we ignored any other word types. Out of the 115,228 words 89,912 were analyzed, i.e. 78%.

LEMPAS was run on SUC and its results were compared to the original lemmatization. In addition to overall results we report the number of the correctly and incorrectly lemmatized words for each of the following word groups: common nouns (NC.*), finite verbs (V@I.*), adjectives (AQ.*), present participles (AP.*) and perfect participles (AF.*). In Swedish, the lemma is always uniquely determined by the word and its POS-tag. Consequently, if a word appears in the corpus with the same POS-tag repeatedly it is always assigned the same lemma. Therefore we report the results both for all the words in the corpus, and for unique words. In the latter case the evaluation is done as if every word appeared in the corpus with the same POS-tag only once.

15.4.2 Lemmatization Results Related to Generating Word Sketches

So far, the Sketch Engine has mainly been run to query the corpus for potential light verbs in order to gain a list of predicate nouns, which typically become their collocates. As the next step, the predicate nouns yielded by the light verb queries are queried separately as collocation bases of verb collocates. Isolated lemmatization errors were to observe. Some of them may well be systematic errors. To give an example: all words ending with *–tecken*, which are (sometimes incorrectly!) tagged as indefinite neuters in **plural** (while they mostly are indefinite neuters in **singular**) will be incorrectly lemmatized as *–tecke* due to the rule concerning plural neuters ending with *–en*. The script relates them to the declension pattern *äpple, möte* etc., in which *–n* is supposed to be chopped off in indefinite plural to obtain the correct lemma form. Not so few errors arise due to tagging errors as well, which the script cannot affect. Perhaps the most serious source of lemmatization errors is the fact that the rules are case-sensitive. Words starting with a capital letter are often lemmatized separately (although the outcome may be identical). Nevertheless, the 90,34% accuracy seems to be good enough, and no further alterations of the lemmatizing scripts are foreseen for this particular task to be performed by the Sketch Engine.

15.4.3 Computing Grammatical Relations

The Sketch Engine enables the user to write a 'grammar' of predefined queries. The queries take the form of functions with a defined number of variables. They could be very simplistically paraphrased as follows: "I am specifying the features of a token and label it as X. When I query the corpus for a token that matches the features, find and list its collocates, whose features I am specifying under the label(s) X (and possibly Z)". For instance: the user wants to find typical direct objects of a verb. He defines the verb by means of the part-of-speech tagging. The label will make it clear that that particular token is going to be the one that will be typed in the query field when this query is performed. The user then provides the features of the direct object ; e.g. by part of speech and by its typical left or right distance from the verb to be queried. These pre-defined functions are called **gramrels** (i.e. 'grammatical relations'). There are several types of relations that can be formulated by the gramrels[3]:

- ***SYMMETRIC** evaluates queries also with the labels '1' and '2' swapped. This directive is active up to the next gramrel line.
- ***DUAL** is similar to *SYMMETRIC but it affect gramrels. It defines two gramrels from the same set of gramrel queries. Gramrel names are separated by a slash (/). All queries are evaluated for the first gramrel and then for the second gramrel with the '1' and '2' labels swapped.

[3]Their definitions have been quoted from the WSGEN.txt file of the Sketch-Engine documentation

- ***UNARY** says that the following gramrel is an unary relation. Only one label is used for unary gramrel queries.
- ***TRINARY** is used for trinary relations. These are translated into regular binary relations with different names. A name of a trinary gramrel should contain '%s' and the respective queries should contain a third label '3'. A value of the word sketch base attribute on the position labeled '3' is then substituted for '%s' in the gramrel name.

The definitions will be exemplified with a fragment of the gramrel for finding prepositional phrases that would typically modify a noun typed into the query line when searching the corpus. The first line of the example below specifies the type of the gramrel relation. The gramrel type determines the syntax of the gramrel. This particular gramrel is a TRINARY one.

```
*TRINARY
=noun_prep_noun_\%s
1:any_noun_nominative 3:any_prep [(tag="DH.*"|tag="DI.*"|\\
tag=poss_pro|tag=number|tag=any_adv|tag=any_noun_genitive|\\
tag=any_adj)]{0,3} 2:any_noun_nominative
```

The second line gives the name of this gramrel. The format of the name is obligatory for each respective gramrel type. The name of a

```
*TRINARY
```

gramrel must end with '%s' The third line contains a regular expression. The labels

```
1:, 2:
```

and

```
3:
```

introduce the three variables in this function. This particular gramrel will be triggered by a search for a word that matches the definition of

```
any_noun_nominative.
```

It will list:

- all nouns that typically act as **modifiers** of the noun typed into the query, which will be further sorted according to by which prepositions they are **introduced**
- all nouns that typically **govern** the noun typed into the query, which will be further sorted according to by which preposition they are **followed**.

The line actually presents the features of a potentially relevant sequence of tokens: a nominative (i.e. non-genitive) noun is followed by a preposition. The preposition can be followed by a determiner or a possessive pronoun or a numeral or an adverb or a noun in the genitive or an adjective within the interval of zero to three tokens. This interval must be again followed by a non-genitive noun. Hence, the regular expression captures all the following examples (the query is 'besvär', and some of the examples are made-up):

besvär **med** den italienska kommunismen besvär med sina barn besvär med dem som vistas där besvär **i**ögon besvär med sociala relationer besvär med sitt sociala liv besvär med sitt dåligt opererade knä besvär med det dåligt opererade knäet **till** besvär till stort besvär till oerhört stort besvär

The regular expression above is partly created using the part-of-speech tags from the PAROLE corpus, partly by macros. E.g., the elements

```
any_noun_nominative
```

is itself a macro defined before as

```
define('any_noun_nominative', '"N\ldotsN@.."').
```

The string

```
'"N\ldotsN@.."')
```

is part of the actual POS-tag for non-genitive nouns and personal pronouns from the PAROLE corpus. Dots stand for 'any character' on the given position.

15.4.4 Empirical Evaluation of Word Sketches

When writing the experimental lexicon entries, results returned by the Word Sketch Engine have been continuously checked against Svenskt Språkbruk [31] and sometimes also against Norstedts Stora Svenska Ordbok [3]. The Sketch Engine proved excellent at detecting morphosyntactic variations of common light verbs and predicate nouns, even in prepositional phrases. Nevertheless, the Word Sketch Engine missed quite a few structures captured by the printed dictionaries. It was especially weak in catching idioms, even when adjusted to taking into account all word combinations occurring more than once. The possible explanation is that some of the missed idioms were probably not present in the PAROLE corpus at all. Some could remain undetected because their frequency was low and they consisted of common words, which freely combine with many others. On the whole: the procedure proved useful in that it yielded sensible results with almost no noise. Its precision seems satisfactory.

Doubts can however be raised about its recall. Zinsmeister and Heid [168] observed a noticeable increase of recall in their collocation extraction when they performed it on statistically parsed German data. It is thus not unlikely that a shallow syntactic parsing would increase the recall on the Swedish PAROLE. In winter 2008, Språkbanken released the first version of Svensk Trädbank [111]. A Word Sketch grammar applied to this corpus with and without making use of the parsing, respectively, would help assess how much parsing could affect the recall, but this experiment goes beyond the scope of the present thesis. A small case study on the predicate noun rekord can be found in Chapter 17.

Quite expectedly, the syntactic criteria for collocation computing are not powerful enough to affect any semantic sorting of the listed collocates. Word Sketches can by no means substitute the manual Corpus Pattern Analysis. (see Chapter 7), at least not in the very complex verbs like the basic movement verbs *lägga, sätta, gå*, but not even in a verb like *bjuda*. A fully automatic sorting of meaning potentials in the sense of CPA would have required semantic markup in the corpus. e.g. with the verb *bjuda* the Word Sketch Engine could not distinguish between the following Lexical Sets:

1. *Mina närmaste vänner tänker jag inte bjuda på officiella middagar, deklarerade prinsen.*
2. *Därför stannade vi ombord och kaptenen bjöd på härligt kaffe och wienerbröd.*
3. *[han] Spelade apa och bjöd på vodka ur en helbutelj.*
4. *Mannen bjuder på cigarretter .*
5. *Skjuts skall vi inte bjuda på .*

- The first and the second sentence belong to the Lexical Set

  `[Host] invites [Guest] to a [Meal, Meal as Social Event]`

- The third and the fourth sentence belong to the Lexical Set

  `[Owner] offers [Recipient, Sharer] [Something Good].`

- The fifth and the sixth sentence belong to the Lexical Set

  `[Provider] offers [Client] [Performance, Service]`

Evidently, this granularity of sorting goes beyond the limits of a lemma-based collocation sorting. Nevertheless, collocate lists generated by the Sketch Engine definitely help to get the first idea of typical uses. It is particularly useful for sorting of concordances with the same lemmas but different morphosyntactic features.

15.5 Determining Entry Candidates

When determining the lexemes and collocations to be included into Swe-VALLEX/PNL, two ways were followed:

- excerpts from relevant studies [161], [159], [157], [158], [156], [155], [97], [71], [138], [34] and [37]
- extraction of LVC-candidates from the PAROLE corpus.

To gain an overall idea of the collocational potential of typical light verbs, several steps towards the extraction of LVC-like structures from the Swedish PAROLE corpus have been taken.

The very first attempt was made in 2003, with PAROLE only accessible on the web and not lemmatized. Nine corpus queries were formulated to vary the distance of the noun object from the lexical verb (see Fig. 15.1). Swedish is an SVO language, and PAROLE contains mainly newspaper and fiction texts, in which declarative sentences clearly predominate over interrogative sentences with verb-subject inversion. Direct objects are therefore typically located at the right-hand side of the verb in this corpus.

The concordances obtained were copied and pasted from the html-page into the Word processor to create subcorpora of verb-noun collocations. The number of concordances was limited by the capacity of the Clipboard in Word. Then each subcorpus was loaded into the concordancing tool WINCONCORD [98], which enabled some basic frequency sorting. No markup was included in the subcorpora due to virtually no option of a more complicated sorting than the alphabetical sorting of the text strings. Therefore the searches were deliberately limited to the infinitives of lexical verbs by narrowing the search to the sequences 'modal verb-lexical verb-noun in various distances from the lexical verb'.

For each subcorpus two frequency-based lists had been built:

1. the verb-noun pairs according to the frequency of the entire pair in the given subcorpus
2. the verb-noun pairs according to the frequency of the noun form in the given subcorpus.

The first list gave the overview of the most frequent verb-noun collocations. The second list suggested which noun forms (i.e. which values of number and definiteness) combined with the given verbs in alphabetical order, with frequency counts. The frequency lists obtained for each subcorpus were exported into Excel and manually checked for LVCs and other fixed collocations. Fig. 15.2 and Fig. 15.3 show the two types of lists. For more details see [24].

In 2006, the manually evaluated collocation lists were utilised as training data for an experiment with a statistical method of automatic collocation extraction, performed by Pecina and Schlesinger [29] on parsed Czech data. It consists of combining the already known statistical features of collocation extraction (Mutual Information, Student's T-Test, Log Likelihood and others) by logistic regression (see [122]).

In this experiment, the manually annotated Swedish concordances from 2003 were used to train the method for searching the – not parsed – Swedish PAROLE corpus for LVCs. The goals of [122] had been to develop a method combining multiple association meassures and to estimate its quality by *precision* and *recall* curves to see whether it could substitute manual collocation extraction. The byproduct of this experiment was again a list of LVC-candidates. The automatically obtained list had the advantage that it had been built from the entire corpus, unlike the manually obtained list, which was biased by the limitations of the original subcorpora (only sequences starting with a modal verb, incomplete export over the Clipboard). Both lists confirmed the significant potential of basic verbs[4] to act as light verbs.

[4]spatial verbs, verbs of physical action, verbs of motion

Query Design			Example	Literal translation of example
1.	modal/ auxiliary verb + lexical verb in infinite form	+ articleless noun in singular	*kan ta ansvar*	*"can take responsibility"*
2.		+ noun in singular with suffigated definite article	*borde ta ledningen*	*"ought to take the leading"*
3.		+ articleless noun in plural	*borde dra diskrimineringsslutsatser*	*"ought to draw discrimination conclusions"*
4.		+ noun in plural with definite article	*ska ta besluten*	*"will take the decisions"*
5.		+article + 1 or 0 deliberate position + any noun	*borde ta ett större ansvar*	*"ought to take a greater responsibility"*
6.		+ preposition + any noun	*bör komma till uttryck*	*"ought to come to expression"*
7.		+ preposition + preposition + any noun	*måste ta i på skarpen*	*"has to take in on the edge"* - idiom
8.		+ any adjective/participle + any noun	*får ge ytterligare besked*	*"has to give further information"*
9.		+ preposition+ article+adjective or nothing + noun	*måste gå över ett extra val*	*"has to go over one more choice/election"* (context not explored)

Figure 15.1: Creation of a subcorpus for the verb-noun collocation extraction – queries

noun	verb	occurrence
1. del	**ta**	102
2. hänsyn	**ta**	**57**
3. hjälp	**ha**	**55**
4. ansvar	**ta**	**52**
5. ställning	**ta**	**49**
6. rätt	**ha**	**39**
7. hjälp	**få**	**35**
8. tag	**få**	**33**
9. pengar	tjäna	33
10. verklighet	**bli**	**31**
11. upphov	**ge**	**28**
12. vakt	**slå**	**28**
13. svar	**ha**	**25**
14. nytta	**dra**	**24**
15. vara	**ta**	**24**
16. beslut	**fatta**	**22**
17. chansen	**få**	**22**
18. rum	**äga**	**20**
19. möjlighet	**få**	**20**
20. barn	ha	20
21. tag	**ta**	**20**
22. problem	bli	19
23. folk	få	19
24. tillgång	**få**	**19**
25. fred	**ha**	**19**
26. tid	**ta**	**19**
27. medlem	bli	18
28. medlemmar	bli	18
29. tag	**ha**	**18**
30. steget	**ta**	**18**

Figure 15.2: The first 30 positions of collocation frequency in the results of queries 1–4

beslag	lägga	6	6
beslut	avvakta	1	1
beslut	fatta	22	22
beslut	få	1	1
beslut	försvara	1	1
beslut	ha	1	1
beslut	överklaga	2	2
beslut	sinka	1	1
beslut	styra	1	1
beslut	vara	1	1
besluten	fatta	2	2
besluten	flytta	1	1
besluten	föregå	1	1
besluten	påverka	2	2
besluten	säga	1	1
besluten	ta	1	1
beslutet	ändra	3	3
beslutet	fatta	1	1

Figure 15.3: A sample from the alphabetical list of nouns and their collocate verbs

Chapter 16

Data Structure

16.1 Usefulness of Word Sketches

The customization of the Word Sketch Engine for the PAROLE corpus provides a basis for a lexicographical description of basic verbs and their typical noun collocates in the position of a direct object, a prepositional object, and a subject. It captures adjectival collocates of nouns, as well as the prepositions the noun requires when modified by a prepositional phrase, along with typical noun collocates in the position of prepositional phrase modifiers. In addition, typical verb collocates are listed for each noun, whose lemma is investigated. The Word Sketch provides lists of characteristic collocates within seconds. The collocation analysis results are quantified and linked to the concordances, from which the numbers result. Figures 16.1 and 16.2 show the Word Sketches for a verb and a noun, respectively.

16.2 Main Principles and Features

The structure of the proposed lexicon was motivated by the needs of an advanced Czech student of Swedish. There are numerous good monolingual Swedish lexicons (in the first place Svenskt Språkbruk [31], which do not only explain the meaning of lexemes, but also describe their behaviour in context and partly their morphosyntactic restrictions (e.g. *used only with negation*). However, even Svenskt Språkbruk pays little attention to the morphosyntactic variation and to the modifying options in phrasemes and light verb constructions.

In addition, no monolingual dictionary can anticipate all contrastive issues that arise for learners with different native-language backgrounds. A nice example is the Swedish triple *sätta-lägga-ställa* versus the English *put* (something somewhere), where the Czech equivalent *dát (give)* has the same problem as English, namely being too unspecific in comparison to Swedish. It is extremely difficult to create lexicon defini-

tions of these three respective Swedish verbs that would teach the non-native speaker to consistently choose the proper variant: the choice is based on the Swedish native conception of items as predominantly vertical vs. predominantly horizontal, or 'axis irrelevant', in connection with other aspects (whether the item must be fixed or whether it keeps its position by itself, etc.).

Certainly, the main issue with the most concrete, literal uses of basic verbs is not collocability, but cognitive conceptualization, which is specific to the respective language communities. Needless to say, the Word Sketch Engine is not the appropriate tool for this task, since it captures tokens, not concepts. The only way out seems to be sufficient exemplification (exceeding the usual number of examples allowed by the space limits of printed dictionaries), which would enable the non-native learner to create analogies.

The lavish exemplification of the *put*-like reading of *sätta* makes SweVallex resemble the clue page of a textbook exercise rather than a dictionary entry. The examples are simply chosen from a number of random concordances (in case of *sätta* some 2 000 of the total 9 000 concordances). Such concordances are preferred that appear surprising to the Czech speaker (e.g. *sätta en pil i*, since Czech requires a more specific verb than the equivalent of *put* (approximately *sting*), and so for a Czech speaker *put* is absolutely unpredictable in this context).

The lexicon is bilingual, with Czech being the target language. The Czech part includes just a minimal description of the Czech equivalents. This feature makes the lexicon more or less useless to a Swedish-speaking student of Czech. Creating a Swedish-Czech lexicon as a production-focused lexicon for Czechs can also seem as missing the point; apparently, the most straightforward way for the non-native Swedish text production would be using a reliable Czech-Swedish dictionary. However, production dictionaries 'atomize' the description of the source-language units according to their equivalents in the target language, such that the picture of the uses of one single Swedish word gets lost. This is also why advanced language students prefer using monolingual dictionaries of the source language instead of bilingual dictionaries: a good monolingual dictionary seems to help draw a 'mental map' of the given lexeme. This map is a blending of semantic features and collocation options.

What production-oriented bilingual as well as monolingual dictionaries can easily miss is a target-language-specific forewarning for collocational as well as cognitive mismatches within the given language pair. There is a need for a description system that would capture the language traps explicitly – at least those based on morphosyntax and on collocability. Such a system is tested by a Czech-related description of cognitively and collocationally difficult Swedish verbs (basic verbs), which are so frequent that nobody can avoid them, and yet they are not fully explained in the teaching materials.

SweVallex-PNL is machine-readable, and its structuring allows for an automatic extraction of a Czech-Swedish glossary. The Czech glossary obtained by the extraction

of the Czech equivalents of Swedish verb uses has the advantage of being fully Swedish-centered. If the lexicon was primarily designed as a Czech-Swedish dictionary, it would be Czech-centered: the mental map of each word would remain Czech, and the Swedish equivalents would be chosen in a way that would disambiguate the respective Czech-centered readings of the given Czech word *('how do I say X in Swedish?')*.

As a result, among all the potential Swedish equivalents such Swedish equivalents would be intuitively selected, whose collocational preferences are not much wider than those of the Czech source word, and the commonest verbs (which are the vaguest) would be in danger of being omitted.

On the other hand, creating an ex-post Czech glossary from a Swedish-Czech lexicon allows the learner to avoid what John Sinclair noticed long ago (see Section 1.1): learning rare words instead of using the less cognitive salient uses of the commonest words. A Swedish-Czech lexicon with a Czech glossary preserves the 'mental maps' of the Swedish words and can be used for learning more about one particular difficult (i.e. vaguely polysemous) verb, as well as it encourages the user to use these verbs in an idiomatic, native-speaker-like way.

The issue of sense disambiguation in bilingual dictionaries is very interesting, and the approach chosen varies from dictionary to dictionary. In each described word, there is a dilemma of whether the reading split is to be based primarily on differences in the collocational preferences in the source language, or rather on differences of the equivalents in the target language. SweVallex attempts at avoiding this dilemma by defining the respective readings by corpus patterns (cf. Chapter 7), enhanced with functors (cf. Chapter 6). The internal structure of the entries is described in Sections 16.3 and 16.4. As a result, the Czech equivalents of one Swedish reading are not necessarily synonymous, as Fig. 16.3 illustrates.

Basic verbs and their lexicalized uses touch the area of grammar as well as that of phraseology. Some uses are specific to the given verb (phrasemes), but other uses put the given verb in connection with other basic verbs, and they are possibly instances of a morphosyntactic or semantic/cognitive regularity, which is waiting to be explored[1]. The regularity possibly hides in the morphosyntactic categories applied in given contexts. One theory that relates grammar directly to cognition, is the Transitivity Hypothesis (see Section 5.3). The Transitivity Hypothesis has been taken into account in the description of light verb constructions in the proposed lexicon by special attention paid to the noun definiteness in the entry structure.

In sum, Swevallex-PNL was designed with respect to the following points:

1. describe and explain a given Swedish lexeme in detail like a monolingual dictionary,

[1] *Semantic* in this sense means *deliberately related to a linguistic interpretation put in contrast to other interpretations*, while *cognitive* relates to the way we perceive the world, which shapes the language, without us necessarily realizing how.

2. provide the morphosyntactic and collocational preferences for each reading in form of a corpus pattern (Chapter 7),
3. determine the underlying valency frame (Chapter 6) of each Swedish corpus pattern,
4. provide Czech equivalents and their patterns with valency frames,
5. list phrasemes and indicate their variability options,
6. pay special attention to light verb constructions and their morphosyntactic preferences with respect to the definiteness of predicate nouns,
7. inform about the options of modifier insertion in light verb constructions, and
8. provide enough examples from the corpus.

SweVallex as well as PNL are xml files with their respective document type definitions (DTD's) and CCS templates. The data was edited in the XMLMind editor [165]. The CCS templates, although they may resemble dictionary entries, have no greater ambition but to facilitate the navigation through the data during the editing, and thus, this is to be emphasized, **they are not meant as the final layout for the users**. Creating the final layout, e.g. for a CD or web release, has never been the purpose of this study, which is a purely linguistic one.

Sections 16.3 and 16.4 analyze and explain the structures of both the lexicon parts, respectively.

16.3 SweVallex

16.3.1 Macrostructure

SweVallex is the lexicon of verbs. Its structure is to the greatest extent possible derived from the structure of Vallex 2.5 [93], the Czech verb valency lexicon. The major deviations from the Vallex 2.5 DTD are motivated by the adaptation to Swedish and by including a second language and the Corpus Pattern Analysis.

The lexicon Swevallex consists of elements `lexeme_cluster` nested in the root element `swevallex_verbs`. Lexeme clusters bring together verbs (elements `lexeme`) that are related by word formation, e.g. *sätta, sätta sig, värdesätta, sätta på*.

```
<?xml version='1.0' encoding='UTF-8'?>
<!ELEMENT swevallex_verbs (lexeme_cluster+)>

<!ELEMENT lexeme_cluster (lexeme+)>
```

```
<!ATTLIST lexeme_cluster
              cluster_id ID #IMPLIED
>
```

Each element `lexeme` has its unique ID. Each element `lexeme` contains the elements `lexical_forms` and `patterns`. Here Swevallex starts to differ from Vallex 2.5. *Patterns* is an element of the same level as `lu_cluster` in Vallex 2.5, but its function is different. Swevallex has **patterns** (like *Corpus Patterns*, Chapter 7) instead of LU's (lexical units) introduced in Vallex 2.5. The element `lexeme` contains the actual lexicon entry.

```
<!ATTLIST lexeme
          lexeme_id ID #IMPLIED
>
<!ELEMENT lexeme (lexical_forms, patterns)>
```

16.3.2 Lemma

The element `lexical_forms` consists of a lemma (element `mlemma`), or a set of lemma variants (`mlemma_variants`). If the lemma is a homograph, it gets its homograph index. The past forms are listed for each lemma separately. Reflexive pronouns as well as particles are captured in the element `admorpheme`, which is optional and can be repeated. The element `admorpheme` has an obligatory attribute, which indicates its type. The values indicate whether the morpheme is a reflexive pronoun, or a particle. This solution was adopted due to the semantically relevant variability in their order – cf.: *ställa in sig* vs. *ställa sig in*.

```
<!ELEMENT lexical_forms ((mlemma|mlemma_variants), admorpheme*, constraints?) >
<!ELEMENT constraints (#PCDATA)>

<!ELEMENT mlemma (#PCDATA)>
<!ATTLIST mlemma
          homograph CDATA #IMPLIED
          preteritum CDATA #REQUIRED
          supinum CDATA #REQUIRED
>
<!ELEMENT mlemma_variants (mlemma+)>
<!ELEMENT admorpheme (#PCDATA)>
<!ATTLIST admorpheme
          type  (reflex|particle) #REQUIRED
>
```

16.3.3 Patterns

The element patterns consists of at least one element pattern. Apart from its unique ID, each element pattern carries the following information in form of attribute values: is it an idiom or not? Is the form of the verb constrained for this particular pattern in any way (e.g. does it only occur in imperative)?

```
<!ELEMENT patterns (pattern+)>
<!ATTLIST pattern
          idiom (0|1) #IMPLIED
          verb_form_constraints CDATA #IMPLIED
          pattern_id ID #IMPLIED
>
```

Each pattern consists of the following elements: proposition, czech, and example.

```
<!ELEMENT pattern (proposition, czech*, example*)>
```

The proposition is the Swedish corpus pattern. It has the form of a Swedish declarative sentence in the present tense (when possible), whose predicate is the lemma verb. Its inner participants and free modifications (cf. Chapter 6) are rendered by slots (element slot), integrated in the proposition (element pattern_text). Each piece of pattern_text has an attribute value according to whether it is the lemma verb or not. The proposition can finish with a (usually English) explaining gloss, which is called implicature (element implicature).

```
<!ELEMENT proposition (pattern_text|slot|implicature)*>
<!ELEMENT pattern_text (#PCDATA)>
<!ATTLIST pattern_text
          verb (1|0) "0"
>
<!ELEMENT implicature (#PCDATA)>
```

Fig. 16.4 shows the proposition *sätta fart på något* in the sense of starting a motor. Note that the word *fart*, which is regarded as a predicate noun, is not explicitly present in the data, but it is referred to via a reference to PNL. The CCS template (in the picture) visualizes only the ID of the given predicate noun. For more details on the description of predicate nouns see Section 16.4.

The Czech equivalents are also presented in form of corpus patterns with slots, pattern text, and implications. When all the equivalents presented have the same corpus pattern, they are all placed in a row of the pattern_text elements with the attribute value verb=1. When an equivalent requires a different pattern, a new Czech pattern is created. Each Czech corpus pattern is classified according to whether it is an idiom or not and whether it really is an equivalent, or just a gloss (used in case there is a lexical gap in Czech).

```
<!ELEMENT czech (pattern_text|slot|implicature)*>
<!ATTLIST czech
          match (equivalent|gloss) #REQUIRED
          idiom (1|0) ”0”
>
```

Each pattern finishes with examples taken from PAROLE or (extremely rarely) from Konkordanser or Google. Examples are elements with free text. Sometimes, examples are shortened, but not consequently. In light verb constructions it is often the case that the examples even include some context.

```
<!ELEMENT example (#PCDATA)>
```

16.3.4 Slot

A lot of linguistic information is hidden in the complex internal structure of the slots. The slots have attributes and a nested element called `occupation`, which is present at least once per slot.

```
<!ELEMENT slot (occupation+)>
```

16.3.5 Surface Form

The element `occupation` carries the information about the surface form of the given slot; i.e., about prepositions, lemma, number, definiteness and other restrictions (this is important with very lexicalized collocations). Occupation can also be represented by a deliberate number of references to PNL (the optional and repetitive empty element `pnl_ref` with the obligatory attribute `ref`). The elements `slot` as well as `occupation` are common for both the Swedish and the Czech patterns. Some of the internal elements of `occupation` are therefore Swedish-specific, while others are Czech-specific, and some are common.

```
<!ELEMENT occupation ((surface_form|cz_surface)*, lexical?, pnl_ref*)>
<!ELEMENT pnl_ref EMPTY>
<!ATTLIST pnl_ref ref IDREF #IMPLIED>
<!ELEMENT surface_form EMPTY>
<!ATTLIST surface_form
          form (på|om|i|till|efter|från|framför|ifrån|för|av|med
  |utan|över|genom|att|vid) #IMPLIED
          case (basic|genitive) ”basic”
>
```

```
<!ELEMENT cz_surface EMPTY>
<!ATTLIST cz_surface
          cz_form (bez|do|k|kolem|na|o|od|po|pro|před
  |s|u|v|vedle|z|za) #IMPLIED
          cz_case (1|2|3|4|6|7) #REQUIRED



>
<!ELEMENT lexical (#PCDATA)* >
<!--text: word forms. Everything else should be in the attributes-->
<!ATTLIST lexical
          lemma CDATA #IMPLIED
          number CDATA #IMPLIED
          article CDATA #IMPLIED
          other_constraint CDATA #IMPLIED



.
```

16.3.6 FGD-Information

The element `slot` has two obligatory attribute values: functor and its obligatoriness according to the valency theory of the Functional Generative Description (see Chapter 6 for reference).

```
<!ATTLIST slot
          functor   (ACT|PAT|ADDR|EFF|ORIG|ACMP|ADVS|AIM|APP|APPS|
ATT|BEN|CAUS|CPHR|CNCS|COMPL|COND|CONJ|
CONFR|CPR|CRIT|CSQ|CTERF|DENOM|DES|DIFF|
                  DIR1|DIR2|DIR3|DISJ|DPHR|ETHD|EXT|FPHR|GRAD|
HER|ID|INTF|INTT|LOC|MANN|MAT|MEANS|MOD|
NA|NORM|PAR|PARTL|PN|PREC|PRED|REAS|
                  REG|RESL|RESTR|RHEM|RSTR|SUBS|TFHL|TFRWH|
THL|THO|TOWH|TPAR|TSIN|TTILL|TWHEN|VOC|
VOCAT|SENT|DIR|OBST|RCMP)   #REQUIRED

          obligatoriness (obl|opt|typ) #REQUIRED
>
```

16.3.7 CPA-Information

The information related to the Corpus Pattern Analysis (CPA, see Chapter 7 for reference) is also contained in the slot. These attribute values are implied as the CPA is less formalized at this stage of the lexicon editing than the FGD-related part.

The attribute `sem_type` contains one or more instances from the current version of the ontology used in the Corpus Pattern Dictionary, which is being built by Hanks ([59]).

```
sem_type CDATA #IMPLIED
```

The attribute `lex_set` contains the lexical sets.

```
lex_set CDATA #IMPLIED
```

16.4 Predicate Noun Lexicon

16.4.1 Macrostructure

The Predicate Noun Lexicon (PNL) contains entries of nouns that occur as nominal components of light verb constructions. They are typically, but not necessarily, event nouns. Besides pure predicate nouns the lexicon also contains parts of phrasemes that exhibit morphosyntactic variability. This can be nominal components of phrasemes governed by a verb, as well as dependent parts of verbless phrasemes (e.g. *pris* ***på någons huvud***). Dependent parts of phrasemes governed by a noun have a simplified entry.

The root element of PNL is the element `predicate_noun_lexicon`, which consists of at least one element `pred_noun_entry` or at least one `phraseme_entry` in deliberate order.

```
<!ELEMENT predicate_noun_lexicon (pred_noun_entry+|phraseme_entry+)* >
```

16.4.2 Predicate Noun Lemma

The element `pred_noun_entry` displays the lemma, its possible homograph index, and the basic information about its genus and declension. As with the verb entries in SweVallex, variant lemmas (e.g. orthographic variants) are allowed.

```
<!ELEMENT lemma_variants (lemma)+>

<!ELEMENT lemma (#PCDATA)>
<!ATTLIST lemma
          lemma_id ID #IMPLIED
```

```
        homonym_index CDATA #IMPLIED
        genus (utrum|neutrum|NA|neutrum_utrum) #REQUIRED
        plural CDATA #REQUIRED
>
```

The introductory part of the entry is followed by up to three lists of typical adjectival and prepositional-group collocates of the given lemma, regardless the other context (elements `adjectives` and `pps`), and the most frequent compounds that occur with the given noun as the base (element `compounds`). Each item of the lists of collocates is surrounded with the nested element `collocate`.

```
<!ELEMENT adjectives (collocate+)>
<!ELEMENT compounds (collocate+)>
<!ELEMENT pps (collocate+)>
<!ELEMENT collocate (#PCDATA)>
```

16.4.3 Light Verb Unit

Like the verb entries were divided into patterns, the predicate noun entries are divided according to the combinations of the given predicate noun with a particular light verb (element `light_verb`).

```
<!ELEMENT pred_noun_entry ((lemma|lemma_variants),

adjectives?,compounds?,pps?,light_verb+)>
```

The light-verb unit consists of the optional element `czech`, which can have an unlimited number of instances, along with two optional elements that cannot be repeated: `definiteness` and `pred_noun_slots`.

```
<!ELEMENT light_verb (czech*, definiteness?, pred_noun_slots?)>
```

The element `light_verb` contains a lot of information in form of attribute values.

The lemma of the light verb occurring in the light verb construction described is to be filled in as the first attribute value.

```
<!ATTLIST light_verb
        lemma CDATA #REQUIRED
```

Each light verb construction in PNL has its unique ID:

```
 id_for_verbslot ID #REQUIRED
```

and it is classified by means of the Lexical Functions (see Chapter 8).

```
 basic_LF (Oper1|Oper2|Copul|Func|Labor1_2|Labor2_1|NA) #REQUIRED
          phasal_LF (Incep|Cont|Fin) #IMPLIED
          causative_LF (Caus|Perm|Liqu) #IMPLIED
          anti_LF (Anti) #IMPLIED
          prox_LF (Prox) #IMPLIED
```

In addition, three properties of the verb in its light-verb use are observed: telicity, punctuality, and volitionality (cf. Chapter 5.3):

```
 telicity (telic|atelic|NA) #IMPLIED
          punctuality (punctual|durative|NA) #IMPLIED
          volitionality (volitional|non-volitional|NA) #IMPLIED
>
```

The `NA` values stand for *non-applicable*, and they are selected when they depend on the context. The attribute `volitionality` describes whether or not the event denoted by the verb normally is a volitional action (regardless of the animacy and agentivity of the agent). The simplified entry for a dependent part of a phraseme does not contain the light-verb unit:

```
<!ELEMENT phraseme_entry ((lemma|lemma_variants),slot*)>
```

When the Czech equivalent is not given in the form of a corpus pattern within the verb entry in SweVallex, it is stated here. The Czech equivalents are obtained by a combination of introspection and searches in the Czech corpus SYN2005. They are nevertheless preferably captured in SweVallex. This element is much of an auxiliary element for editing noun entries that do not have their complements in SweVallex yet. As soon as they get a corresponding entry in SweVallex, the Czech equivalent gets the form of the corpus pattern and moves there.

```
<!ELEMENT czech (#PCDATA)>
```

16.4.4 Noun Definiteness, Modifier Insertion

Several parameters of noun definiteness are observed in the analysis of concordances of each light verb construction (cf. 5.3):

- noun with no determiner (element `bare_noun`)
- noun with the indefinite article (element `indef_art`)
- noun with the postpositive definite article (element `def_art_post`)
- noun with both the prepositive and the postpositive definite article (element `def_art_prepost`)

- noun determined by a genitive or by a possessive pronoun (element `posgen_determiner`)
- noun determined by other non-article determiner (element `other_determiner`)

When an option is clearly predominant or, conversely, extremely rare, it is indicated by a note. When some option does not occur at all in the concordances (or there are just few concordances and they are dubious), the entire element is omitted. Each option is documented by examples. The number of the examples is not necessarily proportional to the ratio of the given option in the concordances. On the contrary, more attention is paid to the less represented options: the examples tend to be longer in context in order to make it possible for the user to find out more about its motivation (e.g. markedness in the information structure, coreferential reasons, etc.). Hypotheses about the motivation of a rare pattern, when any, are formulated in the element `note`. The examples also contain implicit information about the option of the insertion of adjectival and prepositional modifiers (cf. Chapter 17 for more detail).

```
<!ELEMENT definiteness
(bare_noun?, indef_art?, def_art_post?,
def_art_prepost?,
posgen_determiner?, other_determiner?)>
<!ELEMENT example (#PCDATA)>
<!ELEMENT note (#PCDATA)>

<!ELEMENT bare_noun (example|note)*>
<!ELEMENT indef_art (example|note)*>
<!ELEMENT def_art_post (example|note)*>
<!ELEMENT def_art_prepost (example|note)*>
<!ELEMENT posgen_determiner (example|note)*>
<!ELEMENT other_determiner (example|note)*>
```

16.4.5 Slot

The last unit in the PNL entry is the slot. It has a similar structure as in SweVallex: the attributes `functor` and `obligatoriness` and the element `occupation`. Unlike in SweVallex, obligatoriness is not an obligatory attribute in PNL, as the complementations are regarded as optional by default. The attribute obligatoriness is primarily used to mark surface obligatoriness of modifiers in multi-word phrasemes; e.g. *på **rätt/fel** spår, pris på **någons** huvud.*

```
<!ELEMENT pred_noun_slots (slot*)>
```

```
<!ELEMENT slot (occupation*)>

<!ATTLIST slot
functor   (ACT|PAT|ADDR|EFF|ORIG|ACMP|ADVS|AIM|APP|APPS|
ATT|BEN|CAUS|CPHR|CNCS|COMPL|COND|CONJ|
CONFR|CPR|CRIT|CSQ|CTERF|DENOM|DES|DIFF|
            DIR1|DIR2|DIR3|DISJ|DPHR|ETHD|EXT|FPHR|GRAD|
HER|ID|INTF|INTT|LOC|MANN|MAT|MEANS|MOD|
NA|NORM|PAR|PARTL|PN|PREC|PRED|REAS|
            REG|RESL|RESTR|RHEM|RSTR|SUBS|TFHL|TFRWH|
THL|THO|TOWH|TPAR|TSIN|TTILL|TWHEN|VOC|
VOCAT|SENT|DIR|OBST|RCMP)   #REQUIRED
obligatoriness (obl|opt|typ) #IMPLIED
>


<!ELEMENT occupation (surface_form, lexical, cpa, example*, ref*)>

<!ELEMENT lexical (#PCDATA)>
<!ATTLIST lexical
          lemma CDATA #IMPLIED
          number CDATA #IMPLIED
          article CDATA #IMPLIED
          other_constraint CDATA #IMPLIED

>

<!ELEMENT ref EMPTY>
<!ATTLIST ref ref IDREF #IMPLIED>

<!ELEMENT cpa EMPTY>
<!ATTLIST cpa
          sem_type CDATA #IMPLIED
          lex_set CDATA #IMPLIED
          implicature CDATA #IMPLIED
>


<!ELEMENT surface_form EMPTY>
<!ATTLIST surface_form
          form (possgen|hos|på|om|i|till|från|för|av|med|
```

```
utan|över|genom|att|vid) #IMPLIED>
```

16.5 Linking

The SweVallex-PNL lexicon comprises two parts: SweVallex, which captures verbs and their patterns, and nouns and the valency frames they have in connection with the respective light verbs with which they combine. Apart from that, PNL captures all multi-word idioms, whose structure is too complex to be described by the SweVallex pattern system. References go currently from SweVallex to PNL (Fig. 16.5), or from one PNL light-verb frame to another PNL light-verb frame. Lemmas and patterns/light verb frames have their ID's in both lexicons, such that more relations among and within the entries can be displayed in the future.

Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff

Turn on clustering | More data | Less data | Save

change options

sätta parolesw freq = 9492

object_of	2025	5.2
stopp	110	47.15
fart	89	34.35
prägel	28	29.85
spår	62	29.83
punkt	48	27.17
världsrekord	21	26.77
fot	51	26.0
rekord	32	25.83
press	40	25.42
gräns	46	23.84
käpp	15	23.23
finger	21	22.15
likhetstecken	5	20.06
skylt	15	19.83
annons	13	19.11
sprätt	5	19.01
betyg	16	17.86
ära	12	17.86
lit	4	17.5
kryss	8	17.17
tand	13	16.42
värde	21	16.08
eld	12	15.0
nyckel	9	14.86

subject_of	822	3.2
Dollar	11	26.14
dollar/y	4	19.24
Sam	4	11.87
riksdag	6	9.94
Polis	4	9.74
regering	10	8.6
polis	6	7.38
åtgärd	4	6.08
plan	4	5.85
fall	4	4.3

supportverb_prep_predicatenoun_bakom	17	26.4
ord	4	13.57

supportverb_prep_predicatenoun_vid	77	18.6
skrivbord	10	27.64
bord	16	27.34
köksbord	7	25.57
ratt	7	24.83
piano	6	21.5
sida	4	8.59

supportverb_prep_predicatenoun_ur	26	16.9
spel	24	36.43

supportverb_prep_predicatenoun_på	570	15.2
huk	15	32.82
sängkant	12	28.38
radio	15	22.0
kaffevatt	4	20.87
prov	13	20.44
spel	19	19.55
stol	12	19.16
karta	9	18.96
plaen	9	18.56
säng	10	16.36
skolbänk	4	16.22
kaffe	8	15.44
plats	18	14.75
TV-program	5	14.65
bänk	6	14.31
tåg	8	14.24
häst	6	10.99
golv	4	8.84
huvud	6	8.45
söndag	4	7.78
debatt	4	7.15

supportverb_prep_predicatenoun_i	661	10.1
gång	109	31.37
rörelse	29	25.8
hjul	15	24.95
brand	15	23.75
säng	20	23.16
gungning	8	22.09
samband	23	21.78
soffa	12	21.23
hals	13	19.57
pant	5	18.88
lås	6	16.21
bil	19	15.83
verk	14	15.1
centrum	9	13.77
fokus	4	12.23
stol	6	12.17
fängelse	5	11.07
jord	6	10.03
mark	6	9.2
lära	4	8.75
trafik	4	8.62
golv	4	8.5
relation	4	8.28
huvud	6	8.04

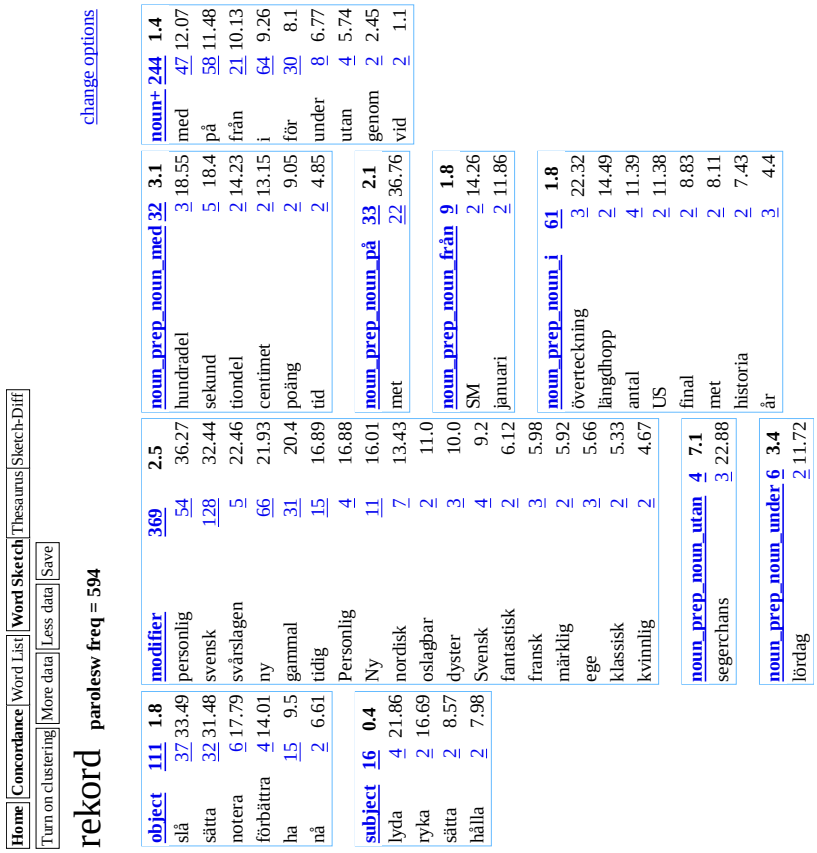
supportverb_prep_predicatenoun_för	119	7.4
munn	5	15.47

supportverb_prep_predicatenoun_under	28	6.1
press	4	15.61

verb+	1108	2.6
på	389	20.15
i	490	18.42
ur	21	12.82
åt	13	8.37
under	21	7.09
framför	4	4.82
till	51	4.44
vid	8	1.79
att	22	1.27
efter	6	1.25
av	31	0.29

supportverb_prep_predicatenoun_till	36	1.5
värld	5	12.98

Figure 16.1: The PAROLE Word Sketch for the verb *sätta*



Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff

Turn on clustering | More data | Less data | Save

rekord parolesw freq = 594

change options

object	111	1.8
slå	37	33.49
sätta	32	31.48
notera	6	17.79
förbättra	4	14.01
ha	15	9.5
nå	2	6.61

subject	16	0.4
lyda	4	21.86
ryka	2	16.69
sätta	2	8.57
hålla	2	7.98

modifier	369	2.5
personlig	54	36.27
svensk	128	32.44
svårslagen	5	22.46
ny	66	21.93
gammal	31	20.4
tidig	15	16.89
Personlig	4	16.88
Ny	11	16.01
nordisk	7	13.43
oslagbar	2	11.0
dyster	3	10.0
Svensk	4	9.2
fantastisk	2	6.12
fransk	3	5.98
märklig	2	5.92
ege	3	5.66
klassisk	2	5.33
kvinnlig	2	4.67

noun_prep_noun_utan	4	7.1
segerchans	3	22.88

noun_prep_noun_under	6	3.4
lördag	2	11.72

noun_prep_noun_med	32	3.1
hundradel	3	18.55
sekund	5	18.4
tiondel	2	14.23
centimet	2	13.15
poäng	2	9.05
tid	2	4.85

noun_prep_noun_på	33	2.1
met	22	36.76

noun_prep_noun_från	9	1.8
SM	2	14.26
januari	2	11.86

noun_prep_noun_i	61	1.8
överteckning	3	22.32
längdhopp	2	14.49
antal	4	11.39
US	2	11.38
final	2	8.83
met	2	8.11
historia	2	7.43
år	3	4.4

noun+	244	1.4
med	47	12.07
på	58	11.48
från	21	10.13
i	64	9.26
för	30	8.1
under	8	6.77
utan	4	5.74
genom	2	2.45
vid	2	1.1

Figure 16.2: The PAROLE Word Sketch for the noun *rekord*

[[Human, Device--]]ACT-obl ***sätter*** [[Physical Object --]]PAT-obl [[Location, Physical Object--]]DIR3-obl
(where it is meant to come, and the entity to be placed is not perceived as primarily vertical or primarily horizontal)
[[Human, Device--]]ACT-obl ***dá umístí usadí strčí zastrčí připevní přibije přilepí přišpendlí přišije přitiskne nasadí vloží přiloží zasune*** [[Physical Object--]]PAT-obl [[Location, Physical Object--]]DIR3-obl

Figure 16.3: Non-synonymous Czech equivalents

Idiom:0
[[Human--driver]]ACT-obl ***sätter*** [[--]]CPHR-obl ***fart*** pnl_ref:fart-saetta-5 [[Car, Motorbike, Truck, Boat, Device--]]PAT-obl **på**

Figure 16.4: Swedish pattern (proposition)

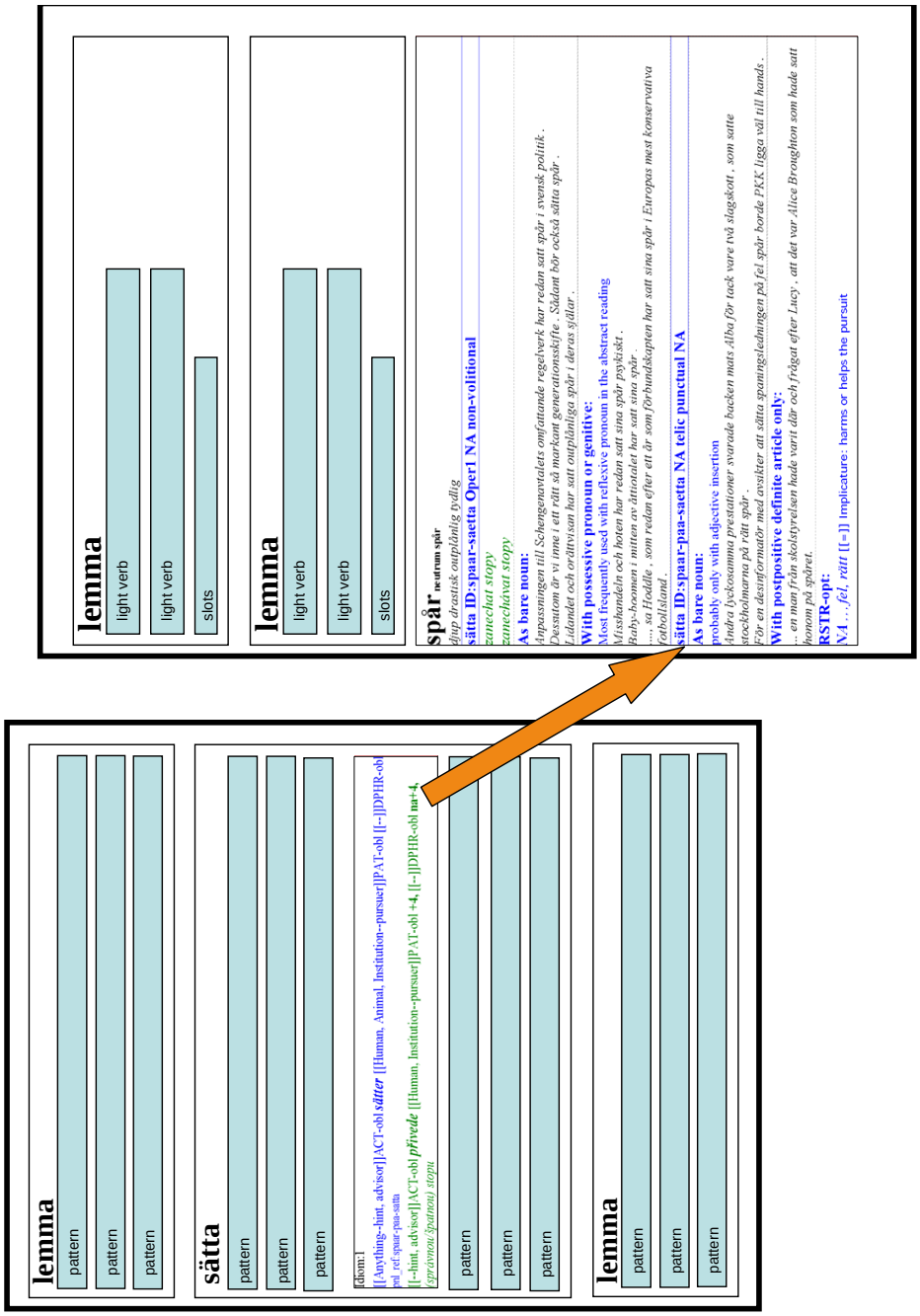


Figure 16.5: Reference from a pattern of *sätta* in SweVallex (left) to the relevant light-verb frame of *spår* in PNL (right)

Chapter 17

Discussion

17.1 Increasing Recall with Targeted Corpus Queries

This Section discusses a few issues that have arisen during the implementation of the initial ideas, and the way they have been dealt with, as well as suggestions of possible improvements of the technical basis that would increase the applicability of the resulting lexical resource.

What has to be first looked into is the expected low recall of the Word Sketches in unparsed corpora, along with the limitations of the lemmatizer. A small experiment was set up to find out how serious this problem is.

A preliminary comparison of the entries *rekord* and *slå* in [3] with the Word Sketch from PAROLE was encouraging: the entry did not contain anything more about the collocation *slå rekord* and the noun *rekord* itself, than was obvious from the Word Sketches, on the contrary. The Word Sketches of several more light verbs and predicate nouns were compared to dictionaries, and the conclusion has always been that the Word Sketch Engine is good enough at finding relevant collocation candidates. The Word Sketch Definitions (the corpus queries in its configuration) were designed to capture direct objects in sentences without subject-verb inversion, taking into account that a number of positions in between them can be occupied by modifiers and determiners. If the recall of the Word Sketch Engine was close to 100%, the sum of concordances yielded by corpus queries targeted at the respective determiner and modifier options should be roughly the same.

To make the experiment easy, the query was designed to capture a direct object (*rekord*) of *slå* that is not preceded by any adjective and not followed by any prepositional modifier:

- bare noun
- noun with the indefinite article

- noun with the definite article
- noun with the pronoun *den här*
- noun with a determiner except articles and *den här*
- noun determined with a noun in genitive or a possessive pronoun (each possibly preceded by another determiner)

The queries mostly avoided lemmas, but listed the alternative word forms to eliminate lemmatization errors. Further setup and observations are as follows:

The cooccurrence of the predicate noun *(rekord)* without any modifier with the light verb *(slå)* is computed for a 0–3 position span in between the two words. This helps capture even sentences with subject-verb inversion as well as negated sentences. The 0–3 positions may not be occupied by the indefinite article, a noun in genitive, a possessive pronoun and any other determiner. The noun *rekord* is accepted only in its non-suffixated form, which excludes the definite forms *rekordet* and *rekorden*. The noun may not be immediately followed by a preposition (to eliminate prepositional modifiers):

```
[word="slå"|word="slår"|word="slog"|word="slagit"]
[word!="ett"\&tag!="A.*"\&tag!="N...G.*"
\&tag!="PS.*"\&tag!="D.*"]
{0,3} [word=".*rekord"] [tag!="SPS"]
```

While the Word Sketch Engine yields 37 occurrences of the collocation *slå rekord*, the targeted query yields 35 occurrences (excluding one wrong hit, which came in due to a typing and subsequently a tagging error in PAROLE).

Fig. 17.1 shows the concordances of bare noun rekord without modifiers.

The query that yields all concordances with the noun with indefinite article and no modifiers looks like this:

```
[word="slå"|word="slår"|word="slog"|word="slagit"]
[word="ett"][]{0,1} [word=".*rekord"] [tag!="SPS"].
```

This query yielded 0 hits.

The query for the definite article with no modifiers takes into account the two possible forms of the definite nouns: either with the postpositive article *(rekordet)*, or with both the pronominal and the postpositive article *(det rekordet)*. This form is regularly used when the noun is modified by an adjective, but it can also be used as a demonstrative pronoun without any adjective. See below and Fig. 17.2:

```
[word="slå"|word="slår"|word="slog"|word="slagit"]
[word="det"|word="den"|word="de"]{0,1} [word="rekordet"
|word="rekorden"] [tag!="SPS"].
```

This query yielded 1 concordance with the double article and four concordances with the postpositive article.

The demonstrative pronoun *den här (det här, de här)* requires the postpositive definite article. Therefore it is treated by a separate query (Fig. 17.3):

```
[word="slå"|word="slår"|word="slog"|word="slagit"]
[word="det"|word="den"|word="de"]
[word="[dh]är"][word=".*rekord"|word="rekordet"|
word="rekorden"] [tag!="SPS"].
```

This query yielded one concordance.

The following query template for nouns without modifiers searches for determiners that have the form of a noun in genitive, or a possessive pronoun. These determiners can be preceded by other determiners (except *den här (det här, de här)*), and therefore one optional position is reserved for these determiners (Fig. 17.4):

```
[word="slå"|word="slår"|word="slog"|word="slagit"]
 [tag="D.*"]{0,1} [tag="N...G.*"|tag="PS.*"]
 {1,3} [word=".*rekord"] [tag!="SPS"].
```

This query yielded 2 concordances.

The last query (Fig. 17.5) captures the noun rekord without modifiers with a determiner other than articles, the demonstrative pronoun *den här*, and a genitive noun or a possessive pronoun:

```
[word="slå"|word="slår"|word="slog"|word="slagit"]
 [tag="D.*"] [word=".*rekord"] [tag!="SPS"].
```

This query yielded 9 concordances.

Alone the set of queries associated with the different morphosyntactic patterns of predicate nouns without any adjectival and prepositional modifiers yielded 52 concordances in total, which is more than 130% of the amount of all occurrences of the collocation *slå rekord* yielded by the Word Sketch. The possible causes of this significant difference are:

- lemmatization errors
- errors in the Word Sketch Definitions
- Word Sketch Definitions power limited by the missing parsing.

The lemmatization can be probably improved by replacing LEMPAS with Hajič's statistical tagger, or by combining the two tools. The Word Sketch Definitions, however, cannot probably be dramatically improved by correcting possible small errors, since each was tested as a regular corpus query, and yielded good results. Given the current Word Sketch Definitions are basically correct, the improvement would have to be sought in the parsing. For the first, it would enable capturing of structures that can impossibly be captured by queries on linear text: sentences with parenthesis, ellipsis, unusual adjective concatenations, etc. For the second, it would make the Word Sketch Definitions much simpler and easier to check.

Eventually, the current outcome is that Word Sketches, despite their limitations, are reliable in finding relevant collocation candidates, but they must be complemented with the usual corpus queries populated with concrete words from the list obtained by the Sketch Engine. Only queries targeted at concrete words yield numbers of concordances relevant for frequency counts and deeper corpus analysis.

17.2 Frequency Counts

The original ambition was to add frequency counts to the respective patterns or morphosyntactic options (cf. the motto of Chapter 7, taken from [55]). However, experiments with the manual frequency counting and manual editing of the frequency counts in the test entry showed that manual frequency counting would burden the lexicographer with an unrealistic amount of work per one single entry. On the other hand, the recall of the targeted corpus queries turned out to be excellent in comparison to the Word Sketch Engine, while high precision remained. Several verb-noun collocations were processed by the same set of queries (see Appendix 21), and the query results had a stable quality. The automation of applying the set of corpus queries to concrete verb-noun combinations would be very time-saving. It would not be a difficult task for a programmer to compile a script that would take these queries as templates and insert relevant values (word forms, lemmas) into the templates one after another according to a list obtained from the Word Sketch Engine. The query results could be visualized in the GUI. The lexicographer would only go through the concordances and remove the wrong ones, while the frequency counts would automatically change. The lexicographer should also be able to edit the queries in the GUI, as different types of light verb constructions might need slightly different corpus queries.

17.3 Irrealis, Negation, and Semantic Definiteness

For a deeper analysis of the light verb constructions from the point of the Transitivity Hypothesis it would be useful to observe in how many concordances of the respective groups (bare noun, indefinite article, etc.) the predication in question is negated or

meets the conditions for being irrealis (cf. Section 5.3), as well as to investigate the **semantic** definiteness of the predicate nouns in more detail; e.g. by observing their adjectival and prepositional modfiers. The corpus query templates (see Appendix 21) are designed to observe adjectives and prepositional groups. Naturally, the queries about prepositional modifiers have to be checked manually, since the numbers are biased by the fact that the corpus queries can impossibly ignore irrelevant prepositional groups. Irrelevant prepositional groups are such groups that do not modify the predicate noun but the light verb, and in addition they are free modifiers, which can be inserted virtually anywhere. Example 260 shows such an irrelevant prepositional modifier and Fig. 17.6 illustrates how the manual editing of query results can be done. The example in the figure was processed by saving a local copy of the query result presented by the html-based online GUI.

(260) *Jag ska försöka slå personligt målrekord i år.*

The prepositional modifier *i år* is a free adjunct of *slå*. The modifiers that are either not modifying the predicate noun or denote typical free adjuncts (e.g. temporal or spatial – *rekord under det gångna året, rekord i bolagen*), are not considered in queries that contain prepositional modifiers. Needless to say, they also bias queries about no prepositional modifiers by eliminating these irrelevant ones as well. Their number in all queries could be easily decreased by providing the entire corpus with the FGD-based tectogrammatical parsing.

17.4 Undetected Information

One very interesting issue in light verb constructions is the preferred form of syntactic negation, which is not the same for all light verb constructions, and a non-native speaker is basically unable to predict how many of them are acceptable, and how to select the one most appropriate for his communicational goal in the given context, if they are not synonymous. cf. Examples 261, 262, and 263 with 264, 265, and 266 or 267, 268, and 269.

(261) *Regeringen har ännu **inte** fattat **beslut** i någon riktning , säger Hjalmar Strömberg.*

(262) *Persson vill **inte** fatta **några beslut**, tror Sydsvenska Dagbladet.*

(263) *I dag fattas dock **inget beslut** om hur detta ska gå till.*

(264) *Butiken stängde hon för tre år sedan , men satte därmed **inte punkt** för sitt yrkesliv.*

(265) *? Butiken stängde hon för tre år sedan , men satte därmed **inte någon punkt** för sitt yrkesliv.*

(266) *? Butiken stängde hon för tre år sedan , men satte därmed **ingen punkt** för sitt yrkesliv.*

(267) *Den unge knektspolingen hade **inte** satt mig i samband med Draken.*

(268) *? Den unge knektspolingen hade **inte** satt mig i **någon samband** med Draken.*

(269) *? Den unge knektspolingen hade satt mig i **inget samband** med Draken.*

In addition, the preferred negation can take a lexical instead of a syntactic expression.

Nevertheless, the PAROLE data turned out to be too sparse for this kind of investigation, and so it was given up. Where possible, negated concordances are presented in the examples to give at least implicit information about negation. A manual identification of negated light verb constructions would be useful for further verification of the Transitivity Hypothesis (as the irrealis identification).

17.5 Different GUI – New Corpus Annotation

To make the routine on SweVallex-PNL more 'comprehensive, fast and reproducible', as mentioned in Section 15.1, more technical support is required. As a first step, the corpus needs to be parsed at least on the level of surface syntax, and the Word Sketches need to be rewritten with respect to the parsing. This will increase their recall. Next, the lexicon should be interlinked with the corpus, on which it is based. Instead of an example selection, the user ought to have the option of viewing all relevant concordances or a selection limited by his own choice. This would of course require a new editing interface, which would also enable the backtracking of the lexicographer's conclusions in case of doubts. However, such interface would not have to be written from scratch, as an interface for this kind of project already exists. In fact, SweVallex-PNL is making use of the possible predecessor of this GUI, thanks to the kindness of the NLP Lab at FI.

Parallel to the later stages of this work, a lexicographical project was launched at the Faculty of Informatics, Masaryk University in Brno, which combined the building of a new lexicon with a new corpus annotation – the Pattern Dictionary of English Verbs [59]. The GUI used for the editing of the Pattern Dictionary allows for querying a lemmatized corpus, creating Word Sketches, and for classifying the respective concordances according to a CPA annotation scheme. If this GUI was adjusted to the SweVallex-PNL annotation scheme, such linking between the corpus and the lexicon would be ensured.

This kind of automation would not only speed up the work, but also increase the precision and consistency even on a much larger piece of data than PAROLE is. The result of this work would not be only a machine-readable lexicon, but also a Swedish treebank with a multi-level annotation of predicates with very little additional efforts.

17.6 Parallel Data

SweVallex-PNL seems to be the first attempt at bilingual CPA patterns, though multilinguality has been experimented with in a number of formal approaches (e.g. [145], [49], [47], and [48], or [19]), not to speak of the intense discussion on bilingual entries among lexicographers in general (cf. e.g. [9], [18]).

The Czech part of the lexicon, unlike the Swedish part, is to a large extent based on introspection, although the Czech corpus SYN2005 has been consulted quite often. With this approach the quantitative information gets lost on whether a light verb construction is preferably translated with a light verb construction or not. To make the bilingual work really corpus-based, the editing work would have to proceed on the basis of a large parallel corpus. Such a corpus is not available yet, despite the effors made within the Intercorp project [70].

Home **Concordance**Word List
Word Sketch ThesaurusSketch-Diff

Corpus: **paroles**
Hits: **36**
conc description

parole.txt	. Till Ebba sa han : --- Nu har Henning	**slagit rekord igen**	. Men de grälade inte på mej utan lät mej
parole.txt	än toppåret 1988 . Omsättningen på småhus	**slog alltså rekord redan**	1988 , tre år innan priserna nådde toppen
parole.txt	ett par minuter , försäkrar han . När man	**slår världsrekord är**	, förutom förmågan att hålla bollen i luften
parole.txt	beundrade han för tålamodet . När Lundholm inte	**slår världsrekord har**	han uppvisning . Det handlar mer om att
parole.txt	år sedan 1977 , då " Stjärnornas krig "	**slog publikrekord ,**	premiärvisas denna sommar en rad filmer
parole.txt	drar dig till blåbärstuvorna , pengar , att	**slå rekord eller**	något annat ? - Jag var getapojke när jag
parole.txt	Gebrselassie . - Jag var bara tvungen att	**slå rekord ,**	fortsatte Gebrselassie och pekade på sina
parole.txt	jättevinster sedan år tillbaka . Börsen	**slår rekord .**	Samtidigt har vi inte ens längre råd att
parole.txt	är dagens huvudnyhet att Kanarieöarna har	**slagit turistrekord igen**	. Förra året bröts 8-miljonersvallen för
parole.txt	det ursprungliga . Julmarknaden på Skansen	**slår rekord varje**	år . Luciasånger framförs av barn i allmogemilj
parole.txt	lopp . Min målsättning med mina lopp är att	**slå världsrekord ,**	sa hon till nyhetsbyrån Reuter efteråt
parole.txt	SOX-omsättning : 88,5 miljoner kronor . Vårruset	**slog rekord -**	i år igen Årets Vårrus i Slottsskogen lockade
parole.txt	finaler . Däremot kommer alla simmare som	**slår världsrekord att**	testas . Dopingtesterna flygs till Paris
parole.txt	och därmed hundratals biografchefer . I år	**slog festivalen deltagarrekord .**	2000 kom . Tacka Kristin Lavransdotter
parole.txt	031-62 41_17 - Ja , det stämmer , vi har	**slagit försäljningsrekord ,**	ganska rejält , två lördagar i rad , säger
parole.txt	Varför inte . Han känns lika bra som då han	**slog svenst rekord och**	vann SM på Åby häromveckan , påstår ägaren
parole.txt	förslag i kongressen . Tysk arbetslöshet	**slår rekord "**	Återhämtningen inte tillräcklig - det finns
parole.txt	osäker framtid till mötes . Arbetslösheten	**slår rekord .**	Tillväxttakten stagnerar . Vreden över
parole.txt	medaljer i guld och tre i silver och han	**slog 22 världsrekord .**	Efter Amsterdam la sig Nurmi i hårdträning
parole.txt	reklamationsnämnden när det gäller datorer	**slog rekord förra**	året . ARN fick då ta emot 158 anmälningar
parole.txt	plus lika mycket i bonus , om han eller hon	**slår världsrekord .**	- Tvåmilsloppet i Hengelo är mitt stora
parole.txt	knappast klaga PARIS : " Jag ska nog försöka	**slå världsrekord ,**	" sa Wilson Kipketer till Malin Ewerlöf
parole.txt	specialeffekter och uppföljare självklara , och	**slog rader av kassarekord .**	Tjugo år senare gör reprisen om bravaden
parole.txt	79_72_22-7 så kommer den på posten . Cavendish	**slår rekord Auktionen**	på de 60 par som bjudits in till årets
parole.txt	SKARA : Det är 15 år sedan Andreas Andersson	**slog rekord och**	gjorde tio mål i en och samma match för
parole.txt	4_000 kr . Sista sidan del 2 Schlager-EM	**slog spelrekord Melodifestivalen**	slog omsättningsrekordet på Tipstjänsts
parole.txt	inköp av verk fortsätter Pompidoucentret att	**slå rekord ,**	berättar Paristidningar , som alla hyllade
parole.txt	pågår den 28 --- 31 maj . Jazzen i Haag	**slår rekord Europas**	största och mest cirkusbetonade jazzfestival
parole.txt	höjdhoppet avgörs den 6 september . Patrik har nu	**slagit världsrekord inomhus**	två gånger . Europarekord har han satt
parole.txt	och Havsbandet . Fri entré . Sverigefinnar	**slår världsrekord !**	Finnstafetten , världens deltagarmässigt
parole.txt	kan mäta sig med Sverige . Etikettböckerna	**slår försäljningsrekord ,**	men de oskrivna konventionerna är minst
parole.txt	av tekniska problem - final i dag Wedlund	**slog rekord Inte**	nog med att friidrottsstjärnorna och publiken
parole.txt	personbästa på 3,70 för två år sedan . I fjol	**slog hon världsrekord tio**	gånger om , men hennes 4,22 från i höstas
parole.txt	räkna med ? Häromdagen , just när högtrycket	**slog hundraårsrekord ,**	minst , var vi ute vid Elfvik på Lidingö
parole.txt	såg han hur alla ville ha mer . Girigheten	**slog rekord .**	Bankerna erbjöd sina anställda 16 procent
parole.txt	sysselsättning " . De vinstrika företag som i dag	**slår börsrekord skulle**	helt enkelt tvingas behålla alla anställda

Figure 17.1: The concordances yielded by the query `[word="slå"|word="slår"|word="slog"|word="slagit"] [word!="ett"&tag!="A.*"&tag!="N...G.*" &tag!="PS.*" &tag!="D.*"]0,3 [word=".*rekord"] [tag!="SPS"].`

Home | **Concordance** | Word List | **Word Sketch** | Thesaurus | Sketch-Diff
View options | Sample | Filter | Sort | Frequency | Collocation | Save

Corpus: **parolesw**
Hits: **1**
conc description

parole.txt medeldistans . - Med lite bättre väder hade vi **slagit det rekordet ,** säger Erik Berglöf . Säkert har han rätt

Figure 17.2: The concordances yielded by the query `[word="slå"|word="slår"|word="slog"|word="slagit"] [word="det"|word="den"|word="de"] [word=".*rekord"|word="rekordet"|word="rekorden"] [tag!="SPS"]`

Home | **Concordance** | Word List | **Word Sketch** | Thesaurus | Sketch-Diff
View options | Sample | Filter | Sort | Frequency | Collocation | Save

Corpus: **parolesw**
Hits: **1**
conc description

parole.txt drömrekordet . - Jag var inte ensam om att **slå det här rekordet .** Alla som var här i kväll var med mig .

Figure 17.3: The concordances yielded by the query `[word="slå"|word="slår"|word="slog"|word="slagit"] [word="det"|word="den"|word="de"] [word="[dh]är"] [word=".*rekord"|word="rekordet"|word="rekorden"] [tag!="SPS"]`

Home | **Concordance** | Word List | **Word Sketch** | Thesaurus | Sketch-Diff
View options | Sample | Filter | Sort | Frequency | Collocation | Save

Corpus: **parolesw**
Hits: **2**
conc description

parole.txt trots det faktum att filmen nästan hade **slagit veckans kassarekord .** Oberoende distributörer som intervjuas
parole.txt Jacquet och under hans ledning har landslaget **slagit Platiniepokens rekord .** I genrepet i lördags besegrades Tyskland

Figure 17.4: The concordances yielded by the query `[word="slå"|word="slår"|word="slog"|word="slagit"] [tag="D.*"]0,1 [tag="N...G.*"|tag="PS.*"]1,3 [word=".*rekord"] [tag!="SPS"]`

Home | **Concordance** | Word List | **Word Sketch** | Thesaurus | Sketch-Diff

View options | Sample | Filter | Sort | Frequency | Collocation | Save

Corpus: **parolesw**
Hits: **9**
conc description

parole.txt	Och ändå blev de alltid kassafilmer som	**slog alla rekord .**	Schamyl skaffade sig Gustafsvik , ett slottsliknande
parole.txt	Svär inte , ber hon . --- Den åttan jag har	**slår alla rekord .**	Det är en samling snorungar . Helst skulle
parole.txt	mobiltelefoner . Försäljningsstatistiken har	**slagit alla rekord och**	det har skrivits otaliga artiklar om fenomenet
parole.txt	fått vederbörligt beröm för , men det här	**slår alla rekord .**	Det finns inga gränser för hur löjliga
parole.txt	rapporterna som bekräftar att turistandet	**slår alla rekord .**	Utländska turister reser hit en masse ,
parole.txt	under åren , men detta torde väl i så fall	**slå alla rekord .**	Det är så orimligt att det knappast kan
parole.txt	HÖGRE , FORTARE , TÄTARE Den svenska skogen	**slår alla rekord Den**	svenska skogen växer så det knakar . Ett
parole.txt	Men som Janne säger , själv tycks han nu	**slå alla rekord .**	FOTNOT : Det är inte första gången det
parole.txt	början hatad av sina grannar har hon redan	**slagit alla rekord .**	Med 74 miljoner besökare på tio år , 24_000

Figure 17.5: The concordances yielded by the query `[word="slå"|word="slår"|word="slog"| word="slagit"] [tag="D.*"] [word=".*rekord"] [tag!="SPS"]`

Home | **Concordance** | Word | List
Word Sketch | Thesaurus | Sketch | -Diff

Corpus: **parolesw**
Hits: **35**
conc description

nyhetsbyråer har rapporterat att Oslobodjenje	**slagit världsrekord i**	antal sidor som någon dagstidning tryckts
offervilja . Nästan alla hjälporganisationer	**slog insamlingsrekord under**	1994 . Katastrofen i Rwanda bidrog starkt
har Johansens " Se på mig " den här veckan	**slagit rekord med**	sina 15 veckor på Svensktoppens förstaplats
från Bergamo . Laget har under säsongen	**slagit rekord i**	hur få meter motståndarlagets offensiva
. Det har det egentligen varit sedan han	**slog distriktsrekord med**	1,83 som 14-åring . Redan som 16-åring
skäl är grupptryck . Aktieägarvänligheten	**slår rekord i**	bolagen . Sedan Nordbanken i höstas har
fria spännvidd på ungefär 500 meter att	**slå världsrekord för**	järnvägsbroar . Vid ingången av 1800-talet
flera mil ut mot havet . Ångermanlands kust	**slår skandinaviskt rekord i**	landhöjning , 300 meter har landet stigit
km/tim . När fransmannen Philippe Goitschel	**slog världsrekord i**	Les Arcs 21 april 1993 gick det ännu mycket
Lady . En sagolik radiosommar Ernst Günther	**slog rekord i**	långsamt tal . Och stockholmarna åkte till
vände brant uppåt 1993 . Såväl 1993 som 1994	**slog småföretagen rekord i**	lönsamhet . Tonie fick napp till slut Tonie
Norge är det uppåt värre . Auktionsfirmorna	**slår rekord på**	löpande band . Firman Kjell Germeten sålde
som efter att ha hamnat i vardagsutbudet	**slog tittarrekord efter**	tittarrekord under våren . Dessutom blir
som efter att ha hamnat i vardagsutbudet	**slog tittarrekord efter tittarrekord under**	våren . Dessutom blir det en himla massa
framtiden . Björn Carapi var närmast att	**slå rekord i**	Falkenberg med sin De Tomaso Pantera .
Trek-filmen First Contact på bio som i USA . Där	**slog filmen kassarekord i**	slutet av förra året . UIP planerade nämligen
den nya serien vet jag inte , eftersom den	**slår rekord i**	långrandig och nästan obegriplig avannonsering
som Greenpeace tagit fram och som nyligen	**slog rekord genom**	att köra 189 mil på 40 liter bensin . Men
den ekonomiska utvecklingen . Sverige har	**slagit världsrekord i**	sanering av de offentliga finanserna och
chef Göran B Sundström , hade hoppats kunna	**slå världsrekord med**	Smilet . Hade hon fått leva tills hon fyllt
Pokal . Hon har överlägsna meriter . Har	**slagit världsrekord i**	USA och tjänat en bra bit över tre miljoner
bana är inte så snabb ännu , men Rex Rodney	**slog ändå banrekord med**	13,7 . Och det syntes redan på sista långsidan
toppform i de senaste starterna . Först när han	**slog världsrekord på**	Solvalla över medeldistans i slutet av
Låneberg , keramiker från Halmstad , har	**slagit värdsrekord i**	konsten att utvinna ett stort fat ur en
förklarade läkaren . Said Aouita har i år	**slagit världsrekord på**	2_000 meter och 5_000 meter och ansågs
meter har kanadensaren Ben Johnson lovat att	**slå världsrekord om**	de yttre förutsättningarna är goda . ---
Patrik obestridligen världsetta inomhus ,	**slog världsrekord med**	2,38 i Berlin i februari . Han vann då
Amanda Carriero , USA . Amanda Carriero	**slog i långdtävlingen världsrekord med**	124,73 meter . Över huvud taget höll längdgrenen
nämnas att exporten av hyvlade trävaror	**slog rekord under**	det gångna året , säger Strängh . --- Bara
vunnit ett SM i rull --- stolsslalom och	**slagit rekord i**	100 meter rullstol , känns det snopet att
vanligt blev värst utsatta . Elförbrukningen	**slog rekord i**	Sydsverige . Snöovädret spred sig den 10
förväntningarna höga , inte minst sedan han	**slagit världsrekord på**	100 yards , kort bana , i våras . Nu fick
mjukt och glömdes bort . Allt medan börsen	**slog kursrekord på**	löpande band , räntan traskade stillsamt
sydafrikanerna mest är brottsligheten . Sydafrika	**slår världsrekord i**	kriminalitet . - Brottsligheten har ökat
fotboll i England , då inte minst massmedierna	**slog rekord i**	självgod efterklokhet . Felet med World

Figure 17.6: The number of relevant prepositional modifiers is in fact lower than calculated by the GUI, after a manual inspection. Concordances highlighted with grey are to be disregarded, as the prepositional modifiers are free adjuncts. The yellow-highlighted concordances were classified as irrealis.

Chapter 18

Conclusion

This thesis has aimed at exploring the approaches to the so-called basic verbs (commonest lexical verbs) that are relevant for a Czech learner of Swedish. Basic verbs are the commonest lexical verbs, which typically denote motion, position, or physical control (e.g., *stand, set, get, go, give, hold*). They are subject to various semantic shifts, through which they exchange their literal meaning for the ability to express some general cognitive categories – an ability that is typical of auxiliary and modal verbs, rather than of lexical verbs. Quite often, the basic verbs combine with (mostly) abstract nouns in the so-called light verb constructions.

Secondary meanings of lexical verbs are used probably in all languages, and there is a significant overlap among languages in respect to which verbs behave this way and what they express. Therefore, they usually do not pose significant understanding problems for foreign speakers. On the other hand, for a correct and idiomatic text production in the target language, the secondary uses of common verbs in the target language have to be explicitly learned, as they are unpredictable.

The secondary uses of basic verbs are a tricky issue both for a foreign learner and for the teacher. They often modify the information structure of a different predicate like auxiliary verbs do, but their distribution may be limited to a certain group of collocates, unlike the genuine auxiliaries – which is why they can be easily ignored by grammar textbooks as well as by dictionaries. They are even neglected by many native speakers among teachers; probably due to the significant disproportion between the cognitive salience and the social salience of the uses of these verbs. Their cognitively salient uses (i.e. uses that we intuitively associate with these particular verbs) do not necessarily dominate over their socially salient uses (i.e. the most frequent ones in the corpus).

One major problem with the basic verbs is that they are so frequent that their concordances in any reasonably large corpus cannot be checked manually and need some automatic pre-sorting. Hence, a substantial part of the work described in this thesis comprised the configuration of the Word Sketch Engine, a tool for collocation

analysis, for Swedish. As the Word Sketch Engine requires the corpus to be lemmatised and no existing Swedish lemmatiser was available at the moment due to legal reasons, a make-do lemmatiser had to be improvised.

Despite all limitations of the corpus setup achieved, the analysis of the actual use of some basic verbs revealed a few astonishing findings. Even though some of these verbs have been well described in grammars (e.g. *komma att, hålla på*), the corpus analysis uncovered the importance of their revisiting (see Chapters 11 and 12).

To give the corpus analysis a consistent form, the data was explored from several points of view:

1. grammaticalization
2. underlying syntax (FGD)
3. Corpus Pattern Analysis
4. Lexical Functions (in predicate nouns light verb constructions)
5. Transitivity Hypothesis (in predicate nouns light verb constructions)

Considering the grammaticalization enables a certain kind of 'thinking beyond phraseology'. This means a systematic reflection of the fact that formulations that appear particularly useful have the tendency to spread beyond their semantically compatible collocates, or they abandon their original morphosyntactic behaviour and develop a new pattern instead, in order to express a more abstract category than their original use. The interesting thing about this is that this process does not finish by crossing the border between 'concrete' and 'abstract', but that the shifts can even continue within the abstract domain (see Section 3.5). The findings about *hålla på* and *komma att*, for instance, are based on this approach.

The underlying syntax (Tectogrammatical Representation) is a useful tool for describing the valency behaviour of frame-evoking words. In combination with Corpus Pattern Analysis it is able to describe the arguments and typical adjuncts with respect to their thematic roles as well as the semantic implicatures evoked by the given pattern.

Two additional theoretical approaches were employed in the description of predicate nouns: the Lexical Functions and the Transitivity Hypothesis. The Lexical Functions are a well-approved instrument for capturing the semantic relation between a collocational base and its collocate. It is particularly useful in a multilingual description of light verb constructions, since, given that there are equivalent light verb constructions in the source language and the target language, the predicate nouns are usually equivalent, and it is the light verbs that differ across the languages in the given semantic relation to the noun base.

The Transitivity Hypothesis, however, is a speculation, whose validity could not be proved in this thesis. Nevertheless, pursuing it does not add up any work, and

it could explain the varying predicate noun definiteness in light verb constructions. The noun definiteness, along with the ability of the predicate noun to be modified by other elements, is a feature that has been ignored in the currently available dictionaries, although it is vital for the idiomatic use of light verb constructions as well as phrasemes. The proposed description pays special attention to this issue, and would do so even without considering the Transitivity Hypothesis at all.

The proposed description of the Swedish basic verbs has the shape of a bilingual machine-readable, corpus-based Swedish-Czech lexicon, which consists of two interlinked XML components: SweVallex, the lexicon of verbs, and Predicate Noun Lexicon (PNL), the lexicon of predicate nouns. The structure of both the lexicon components was tested and refined on a handful of sample entries, which are attached as an appendix.

The lexicon is now ready for the start of routine lexicographical work, with the exception of automatic frequency counts and manual markup of irrealis and negation. Yet moving the annotation scheme into the new CPA GUI, as proposed in the discussion, is not expected to be an issue, and it would mean a substantial enhancement of the final product.

Bibliography

[1] Adam Meyers and Reeves, R. and Macleod, C. and Szekely, R. and Zielinska, V. and Young, B. and Grishman, R. The NomBank Project: An Interim Report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 – May 7 2004. Association for Computational Linguistics. 9.4

[2] Sture Allén. *Tiotusen i topp*. Almqvist & Wiksell, Stockholm, 1972. 10.1

[3] Sture Allén et al. *Norstedts stora svenska ordbok*. Norstedts, 1995. 1.2, 15.4.4, 17.1

[4] Sue Atkins. Tools for Computer-Aided Lexicography: The Hector Project. *Acta Linguistics Hungarica*, 41:5–72, 1993. 15.1

[5] Michael Barlow. Paraconc-beta, version 1.0 – build 269. software, 2004. 7

[6] Irène Baron and Michael Herslund. Support Verb Constructions as Predicate Formation. In *The Structure of the Lexicon in Functional Grammar*. John Benjamins, Amsterdam/Philadelphia, 1998. 1, 4.2, 4.6, 3

[7] Cyril Belica. Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Institut für Deutsche Sprache, Mannheim, Germany, 1995. 15.3

[8] Bernd Heine and Ulrike Claudi and Friederike Hünnemeyer. *Grammaticalization. A Conceptual Framework*. University of Chicago Press, 2001. 3.1, 4, 3.5, 3.5.1, 13, 14, 3.5.2, 17, 3.5.2, 19

[9] Kristín Bjarnadóttir. Verbal Syntax in an Electronic Bilingual Icelandic Dictionary: A Preliminary Study. *LexicoNordica. Tidsskrift om leksikografi i Norden utgitt av Nordisk forening for leksikografi i samarbeid med Nordisk språksekretariat*, 8:5–24, 2001. 17.6

[10] Tavs Bjerre. Event Structure and Support Verb Constructions. In *Proceedings of the ESSLLI Student Session 1999*, 1999. 1, 4.4, 9

[11] Frede Boje. Hvor finder man *finde anvendelse*? In Ásta Svavarsdóttir, Guðrún Kvaran, and Jón Hilmar Jónsson, editors, *Nordiske Studier i Leksikografi Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995*, volume 3 of *Skrifter utgitt av Nordiske forening for leksikografi*, pages 51–68, Reykjavík, 1995. 1, 4.3

[12] I.A. Bolshakov, A.F. Gelbukh, and S.N. Haro Galicia. Electronic Dictionaries: For both Humans and Computers. In *International Forum on Information and Documentation*. Federation Internationale de Documentation (FID), 1999. 1.3

[13] Anna Braasch. Formalised Representation of Collocations in a Danish Computational Lexicon. In *The ninth EURALEX International Congress Proceedings*, volume II., pages 475–488. Stuttgart, 2000. 9.8

[14] Anna Braasch and Sussi Olsen. Towards a Strategy for a Representation of Collocations – Extending the Danish PAROLE Lexicon. In *Second International Conference on Language Resources and Evaluation, Proceedings. vol. 2, Athen*, pages 1009–1064, 2000. 1, 9.8

[15] Anna Braasch and Sussi Olsen. Formalised Representation of Collocations in a Danish Computational Lexicon. In *The Ninth EURALEX International Congress, Proceedings, Vol. II, Stutztgart*, pages 475–488, 2000. 1

[16] Miriam Butt. The Light Verb Jungle. *Harvard Working Papers in Linguistics*, 9 (Papers from the Harvard/Dudley House Light Verb Workshop), 2003. URL http://ling.uni-konstanz.de/pages/home/butt,quoted2007-01-19. 4.1, 4.3, 4.4, 4.8

[17] Joan Bybee. *Morphology: a study of the relation between meaning and form*, volume 9 of *Typological Studies in Language*. John Benjamins, Amsterdam/Philadelphia, 1985. 3.2, 3, 3.2, 5, 3.3, 3.3.5, 7, 3.3.5, 9, 3.4, 3.5.1, 3

[18] Lise Cantell. Målspråkets verbkonstruktioner i en tvåspråkig produktionsordbok. *LexicoNordica. Tidsskrift om leksikografi i Norden utgitt av Nordisk forening for leksikografi i samarbeid med Nordisk språksekretariat*, 2:19–32, 1995. 17.6

[19] Tommaso Caselli, Nancy Ide, and Roberto Bartolini. A Bilingual Corpus of Inter-linked Events. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. URL http://www.lrec-conf.org/proceedings/lrec2008/. 17.6

[20] František Čermák, Jiří Hronek, Jaroslav Machač, et al. *Slovník české frazeologie a idiomatiky – výrazy slovesné*, volume 3, chapter Introduction, pages 26–28. Academia, Praha, 1994. 9.3, 9.3, 9.3, 4

[21] František Čermák. Víceslovná pojmenování typu verbum – substantivum v češtině (příspěvek k syntagmatice tzv. abstrakt). *Slovo a slovesnost*, 4(35): 287–306, 1974. 4.6

[22] František Čermák. Abstract Nouns Collocations: Their Nature in a Parallel English-Czech Corpus. In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Barnbrook, Danielsson & Mahlberg, Birmingham, 1995. 4.2

[23] František Čermák. Syntagmatika slovníku: typy lexikálních kolokací. In Z. Hladká and P. Karlík, editors, *Čeština – Univerzália a specifika*, volume 3, pages 223–232. Brno, 2001. 2.2

[24] Silvie Cinková. Extraction of Swedish Verb-Noun Collocations from a Large Msd-Annotated Corpus. *The Prague Bulletin of Mathematical Linguistics 82*, pages 99–102, 2004. 15.5

[25] Silvie Cinková. From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *LREC 2006 Proceedings*. Genoa, Italy, May 2006. 9.5.3

[26] Silvie Cinková and Veronika Kolářová. *Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank*, pages 113–139. Veda Bratislava, Slovakia, 2005. ISBN 80-224-0880-8. 4.6

[27] Silvie Cinková and Jan Pomikálek. LEMPAS: A make-do lemmatizer for the swedish PAROLE-corpus. *Prague Bulletin of Mathematical Linguistics*, 86: 47–54, 2006. ISSN 0032-6585. 15.4.1

[28] Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Annotation of English on the Tectogrammatical Level. Technical Report 35, UFAL MFF UK, 2006. 6.3.2

[29] Silvie Cinková, Petr Podveský, Pavel Pecina, and Pavel Schlesinger. Semi-automatic Building of Swedish Collocation Lexicon. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1890–1893, Paris, France, 2006. ISBN 2-9517408-2-4. 15.5

[30] Silvie Cinková, Josef Toman, Jan Hajič, Kristýna Čermáková, Václav Klimeš, Lucie Mladová, Jana Šindlerová, Kristýna Tomšů, and Zdeněk Žabokrtský. Tectogrammatical Annotation of the Wall Street Journal. *Prague Bulletin of Mathematical Linguistics*, 92, 2009. ISSN 0032-6585. 1

[31] Ulla Clausén et al. *Svenskt Språkbruk – ordbok över konstruktioner och fraser*. Norstedts Ordbok and Svenska Språknämnden, 2003. 9.7, 15.4.4, 16.2

[32] Jan Cuřín, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank Version 1.0, 2004. 9.5.3

[33] Czech National Corpus. Czech National Corpus – SYN2005. Institute of the Czech National Corpus, Praha, 2005. URL http://www.korpus.cz>. 15.2

[34] Ela Dura. *Substantiv och stödverb*, volume 18 of *Meddelanden från Institutionen för Svenska Språket*. Göteborgs universitet, 1997. 1, 4.1, 4.5, 5.2, 5.7, 15.5

[35] Eva Ejerhed, Gunnel Källgren, and Benny Brodda. Stockholm-Umeå Corpus Version 2.0. Stockholm University, Dep. of Linguistics and Umeå Uni-versity, Dep. of Linguistics, 2006. 15.2

[36] Lena Ekberg. *Gå till anfall och falla i sömn. En strukturell och funktionell beskrivning av abstrakta övergångsfaser*, volume A 43 of *Lundastudier i nordisk språkvetenskap*. Lund University Press, Lund, 1987. 1, 4.5

[37] Lena Ekberg. Verbet *ta* i metaforisk och grammatikaliserad användning. *Språk och Stil*, 3:105–139, 3. 4, 15.5

[38] Michael Rundell et al. *Macmillan English Dictionary for Advanced Learners*. Macmillan Publishers Ltd., 2002. 1, 15.3

[39] Stefan Evert and Hannah Kermes. Experiments on Candidate Data for Collocation Extraction. In *Companion Volume to the Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*, pages 83–86, Budapest, Hungary, 2003. 2.2

[40] Stefan Evert and Birgitte Krenn. Exploratory Collocation Extraction. In *PHRASEOLOGY 2005 The many faces of Phraseology*. Université catholique de Louvain (Belgium), 13-15 October 2005. 2.2

[41] C. Fellbaum. *WordNet: An Electronical Lexical Database*. The MIT Press, Cambridge, MA, 1998. 9.4

[42] Katalin Fenyvesi-Jobbágy. Non-literal and non-metaphorical uses of Danish komme 'come': A case study. *Jezikoslovlje*, 2003. 4

[43] Charles J. Fillmore. Frame Semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing, Seoul, South Corea, 1982. 9.6

[44] Charles J. Fillmore, Christopher R. Johnson, and M. L. R. Petruck. Background to FrameNet. *FrameNet and Frame Semantics. International Journal of Lexicography – Special Issue*, 16:235–250, 2003. 4.4, 9.6

[45] John Rupert Firth. *Papers in Linguistics 1934-1951*. London: Oxford University Press, 1957. 2.2

[46] John Rupert Firth. *Papers in Linguistics 1934-1951*, chapter Modes of Meaning, pages 190–215. London: Oxford University Press, 1957. 5.2

[47] Thierry Fontenelle. Co-occurence Knowledge, Support verbs and Machine Readable Dictionaries. In *Papers in Computational Lexicography, COMPLEX'92, Budapest*, 1992. 4.2, 17.6

[48] Thierry Fontenelle. Using a Bilingual Computerized Dictionary to Retrieve Support Verbs and Combinatorial Information. *Acta Linguistica Hungarica*, 41(1-4):109–121, 1993. 17.6

[49] Thierry Fontenelle. Using a Bilingual Dictionary to Create Semantic Networks. *International Journal of Lexicography*, 10(4):274–303, 1997. 17.6

[50] Stefan Thomas Gries. Dispersions and Adjusted Frequencies in Corpora. *International Journal of Corpus Linguistics*, 13(4):403–437, 2008. 7.3

[51] Heide Günther and Sabine Pape. Funktionsverbgefüge als Problem der Beschreibung komplexer Verben in der Valenztheorie. In Helmut Schumacher, editor, *Untersuchungen zur Verbvalenz: eine Dokumentation über die Arbeit an einem deutschen Valenzlexikon*, Forschungsberichte/Institut für deutsche Sprache Mannheim, pages 92–128. Narr, Tübingen, 1976. 4.1

[52] Jan Hajič. SE030107x – A statistical tagger/lemmatizer based on the SUC corpus. Personal communication 2006-03-19, 2002. 2

[53] Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68. „ Vaxjö University

Press, November 14–15, 2003 2003. ISBN 91-7636-394-5. GA405/03/0913, LN00A063. 6.3.1, 9.5.1

[54] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. Prague Dependency Treebank 2.0. Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, Jul 2006, 2006. URL http://ufal.mff.cuni.cz/pdt2.0/. 6.1, 1, 6.3.5

[55] Patrick Hanks. The Probable and the Possible: Lexicography in the Age of Internet. In Lee Sangsup, editor, *Asialex 2001 Proceedings, Seoul Korea*, 2001. 1, 17.2

[56] Patrick Hanks. Lexicography. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, chapter I. – Fundamentals, pages 48–70. Oxford University Press, Oxford, 2003. 3.1

[57] Patrick Hanks. *Norms and Exploitations: Corpus, Computing, and Cognition in Lexical Analysis*. MIT Press, forthcoming. Manuscript, obtained 2003 from the author. 1.1, 1.3, 3.1, 3.5.2, 4.2, 7.1, 1, 7.2, 7.3

[58] Patrick Hanks and Elisabetta Jezek. Shimmering Lexical Sets. In *Euralex XIII 2008 Proceedings*, Pompeu Fabra University, Barcelona, 2008. 7.2

[59] Patrick Hanks and James Pustejovsky. A Pattern Dictionary for Natural Language Processing. *Revue Francaise de linguistique appliquée*, 10(2), 2005. 7.2, 7.3, 16.3.7, 17.5

[60] Patrick Hanks, Anne Urbschat, and Elke Gehweiler. German Light Verb Constructions in Corpora and Dictionaries. *International Journal of Lexicography – Special Issue: Corpus-Based Studies of German Idioms and Light Verbs*, 19(4): 439–458, 2006. 2, 2.2, 4.1, 4.7, 4.7, 4.8, 5.7

[61] Erik Hansen. *Stå, sidde, ligge*. *Mål & Mæle*, 1(2):26–32, 1974. 4

[62] Ulrich Heid. Towards a Corpus-based Dictionary of German Noun-verb Collocations. In Thierry Fontenelle, Philippe Hilligsmann, Archibald Michiels, AndréMoulin, and Siegfried Theissen, editors, *Actes EURALEX'98 Proceedings*, volume 1, pages 301–312, Liège, 1998. Universitéde Liège, Départements d'anglais et de néerlandai. 4

[63] Gerhard Helbig and Joachim Buscha. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Verlag Enzyklopädie, Leipzig, 1996. 1.2, 4.1, 4.3, 4.7, 4.7, 4.7, 4.7

[64] Dana Hlaváčková, Aleš Horák, and Vladimír Kadlec. *TSD 2006*, chapter Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. Springer Verlag, 2006. 9.9

[65] Michael Hoey. *Lexical Priming. A new theory of words and language*. Routledge, 2005. 2.2

[66] Paul Hopper. Emergent Grammar. *Berkeley Linguistics Conference (BLS)*, 13:139–157, 1987. URL http://eserver.org/home/hopper/emergence.html. 1.1, 3.2, 3.4, 3.5.2, 10.2

[67] Hopper, Paul J. and Thompson, Sandra A. Transitivity in Grammar and Discourse. *Language*, 56(2):251–299, 1980. 9, 5.1, 5.1, 5.3, 3, 4, 5.3.10, 5.4, 5.5, 5.6

[68] Susan Hunston and Gill Francis. *Pattern Grammar: a Corpus-driven Approach to the Lexical Grammar of English*. John Benjamins Publishing Company, 200. 7.3

[69] Igor Mel'čuk, Alexander Žolkovskij. *Tolkovo-kombinatornyj slovar'sovremennovo russkovo jazyka*. Wiener Slawistischer Almanach, Sonderband 14, 1984. 8.1, 9.2.4

[70] Intercorp. Intercorp – projekt paralelních korpusů Filozofické fakulty Univerzity Karlovy v Praze. www.korpus.cz-intercorp, quoted 2009. 15.2, 17.6

[71] Ulrike Jakobsson. Familjelika betydelser hos STÅ, SITTA och LIGGA. En analys ur den kognitiva semantikens perspektiv. Technical report, Lunds universitet. Institutionen för nordiska språk, Lund, 1996. 4, 2, 15.5

[72] M. Jelínek. O verbonominálních spojeních ve spisovné češtině. In *Přednášky a besedy z XXXVI. běhu LŠSS*, pages 37–51. MU Brno, 2003. 4.1, 4.3, 4.7, 15, 16

[73] Torben Juel Jensen. Kan man 'ligge' i et mentalt rum? In *Nydanske studier & almen kommunikationsteori. Artikler om partikler.* , pages 73–100. Københavns Universitet. Institut for Nordisk Filologi, København, 2000. 4

[74] Otto Jespersen. *A Modern English Grammar on Historical Principles*, volume 6. London: George Allen & Unwin & Copenhagen: Ejnar Munksgaard., 1954. 2, 4.1

[75] Joan Bybee and Revere Perkins and William Pagliuca. *The Evolution of Grammar. Tense, aspect, and modality in the languages of the world*. The University of

Chicago Press, Chicago & London, 1994. 3.2, 2, 3.2, 3.3, 3.4, 3.5, 13, 15, 10.2, 1

[76] Christopher Johnson and Charles Fillmore. The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure, 2000. URL citeseer.ist.psu.edu/johnson00framenet.html. 9.6

[77] Sylvain Kahane. The Meaning-Text Theory. In *Dependency and Valency. An International Handbook on Contemporary Research*. de Gruyter, Berlin, 2003. 8.1, 9.2

[78] Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress. Lorient, France*, pages 105–116. Universite de Bretagne-Sud, 2004. 15.2, 15.3

[79] Karin Kipper, Martha Palmer, and Owen Rambow. Extending PropBank with VerbNet Semantic Predicates. In *Workshop on Applied Interlinguas, held in conjunction with AMTA-2002*. Tiburon, CA, October 2002. 9.4

[80] Veronika Kolářová. *Valence deverbativních substantiv v češtině (PhD thesis)*. PhD thesis, ÚFAL MFF UK, Prague, 2005. 4.6

[81] Veronika Kolářová. Valency of Deverbal Nouns in Czech. *Prague Bulletin of Mathematical Linguistics*, 86:5–20, 2006. ISSN 0032-6585. 6.3.1

[82] George Lakoff. *Women, Fire and Dangerous Things. What categories reveal about the mind*. Chicago University Press, Chicago, 1987. 3.1, 4, 3.5.1

[83] George Lakoff and Mark Johnson. *Metaphors We Live By*. Chicago University Press, 1980. 4

[84] Mats Larsson. translation of B. Hrabal's text "Taneční hodiny pro starší a pokročilé". personal communication, September 18 2006. 3, 6

[85] Beth Levin. *English Verb Classes and Alternations*. University of Chicago Press, 1993. 9.1, 9.4

[86] Ann Lindvall. *Transitivity in Discourse. A Comparison of Greek, Polish and Swedish.*, volume 37 of *Travaux de l'Institut de Linguistique de Lund*. Lund University Press, 1998. 5.1, 5.1, 5.1, 5.6

[87] Ann Lindvall. Swedish verb particles and Polish aspect marking. In Arthur Holmer, Jan-Olof Svantesson, and Åke Viberg, editors, *Proceedings of the 18th Scandinavian Conference of Linguistics, Lund 18-20 May 2000*, volume 2, 2001. 5.6

[88] Markéta Lopatková. Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *Prague Bulletin of Mathematical Linguistics*, 79–80:37–60, 2003. 9.5.2

[89] Markéta Lopatková and Jarmila Panevová. *Recent Developments of the Theory of Valency in the Light of the Prague Dependency Treebank*, pages 83–92. Veda Bratislava, Slovakia, 2005. ISBN 80-224-0880-8. 6.3.1, 6.3.5

[90] Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, and Václava Benešová. VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report TR-2003-18, UFAL/CKL MFF UK, Prague, 2003. 1.3, 9.9

[91] Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, and Václava Benešová. VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report TR-2003-18, UFAL/CKL MFF UK, Prague, 2003. 6.3.6

[92] Markéta Lopatková, Zdeněk Žabokrtský, and Karolina Skwarska. Valency Lexicon of Czech Verbs. In *LREC 2006 Proceedings*. Genoa, Italy, May 2006. 9.5.2

[93] Markéta Lopatková, Zdeněk Žabokrtský, Václava Kettnerová, Karolina Skwarska, Eduard Bejček, Klára Hrstková, Michaela Nová, and Miroslav Tichý. VALLEX 2.5 – Valency Lexicon of Czech Verbs, version 2.5. Software prototype, 2007. 16.3.1

[94] Markéta Lopatková, Zdeněk Žabokrtský, and Václava Kettnerová. *Valenční slovník českých sloves*. Nakladatelství Karolinum, Praha, 2008. ISBN 978-80-246-1467-0. 6.3.6

[95] Eva Macháčková. Analytické predikáty. Substantivní názvy dějů a statických situací ve spojení s funkčními slovesy. *Jazykovědné aktuality*, 3, 4(10):122–176, 1983. 4.6, 12

[96] C. Macleod. Lexical Annotation for Multi-word Entries Contatining Nominalizations. In *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002); Las Palmas, Canary Islands, Spain*, pages 943–948, 2002. 9.4

[97] Sven-Göran Malmgren. *Begå eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet 1800-2000*. Rapporter från ORDAT. Göteborgs universitet. Institutionen för svenska språket., Göteborg, 2002. 4.2, 4, 15.5

[98] Zdeněk Martínek. Winconcord. software, 1996. Developed by Zdeněk Martínek at the University of West Bohemia, Pilsen, Czech Republic, in close Collaboration with Les Siegrist from the Technische Hochschule Darmstadt, Germany. Version 2.0 (July, 1996). 15.5

[99] Igor Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, 1988. 8.1, 9.2.3

[100] Igor Mel'čuk. *Dictionnaire explicatif et combinatoire du français contemporain*. Les Presses de l'Université de Montréal, 1984. 8.1, 9.2.4

[101] Igor A. Mel'čuk. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–105. John Benjamins, Amsterdam/Philadelphia, 1996. 8.2, 8.3, 8.3.1, 8.3.2, 8.3.3

[102] Igor A. Mel'čuk and Alain Polguère. A Formal Lexicon in the Meaning-Text Theory (or How to Do Lexica with Words). *Computational Linguistics*, 13 (3–4):261–275, 1987. 9.2, 9.2.1, 9.2.2

[103] Adam Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. Annotating Noun Argument Structure for NomBank. In *Proceedings of LREC-2004*, Lisbon, Portugal, 2004. 9.4

[104] Adam Meyers, Ruth Reeves, and Catherine MacLeod. *NomBank v 1.0*. Linguistic Data Consortium, Philadelphia, 2008. 6.3.1, 6.3.7

[105] Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, and Lucie Kučová. Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical Report TR-2005-28, ÚFAL MFF UK, Prague, Prague, 2005. 4.6, 4.6, 4.6

[106] Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep., 2006. 6.3.1, 6.3.6

[107] P. M. Mitchell et al. Building a Large Annotated Corpus of English: The PennTreebank. *Computational Linguistics*, 1993. 9.4

[108] Morton Benson and Evelyn Benson and Ilson, Robert F. *BBI Dictionary of English Word Combinations*. John Benjamins, 1997. 2.2, 8.1, 9.1

[109] Alexander Nakhimovsky. A Case of Aspectual Polysemy, with Implications for Lexical Functions. In Leo Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 169–179. John Benjamins, Amsterdam/Philadelphia, 1996. 5.3.3

[110] Sanni Nimb and Bolette Sandford Pedersen. Treating Metaphoric Senses in a Danish Computational Lexicon – Different Cases of Regular Polysemy. In *EURALEX 2000, Proceedings,*, pages 679–692. Stuttgart, 2000. 9.8

[111] Joakim Nivre, Jens Nilsson, and Johan Hall. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, May 2006. URL http://www.msi.vxu.se/users/nivre/papers/talbanken05.pdf. 15.4.4

[112] Elisabeth Nylund and Britta Holm. *Deskriptiv svensk grammatik*. Stockholm, 1993. 1.2, 15.4.1

[113] Karel Pala and Pavel Ševeček. Valence českých sloves. In *Sborník prací FFBU*, 1997. 9.9

[114] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 2005. URL http://www.cs.rochester.edu/~gildea/palmer-propbank-cl.pdf. 9.4

[115] Martha Palmer, Paul Kingsbury, Olga Babko-Malaya, Scott Cotton, and Benjamin Snyder. Proposition Bank I. LDC2004T14, ISBN: 1-58563-304-6, Sep 01, 2004. 9.4

[116] Jarmila Panevová. On Verbal Frames in Functional Generative Description I. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974. 6.3.1, 6.3.7

[117] Jarmila Panevová. On Verbal Frames in Functional Generative Description II. *Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975. 6.3.4, 3, 6.3.6

[118] Jarmila Panevová. Valency Frames and the Meaning of the Sentence. In Philip A. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*, pages 223–243. John Benjamins Publishing Company, Amsterdam, Philadelphia, 1994. 6.3.1

[119] Jarmila Panevová. Poznámky k valenci podstatných jmen. In *Čeština – univerzália a specifika 2. Sborník konference ve Šlapanicích u Brna, 17.-19.11.1999 (ed. Zdeňka Hladká, Petr Karlík)*, pages 173–180. Masarykova Univerzita v Brně, ISBN 80-210-2262-0, 2000. 6.3.1, 6.3.7, 6.3.7

[120] Jarmila Panevová. Sloveso: centrum věty, valence: centrální pojem syntaxe. In *Proceedings of Aktuálne otázky súčasnej syntaxe*, pages 73–77. Budmerice, Slovakia,, Nov. 7-8 2002. 9.5.2

[121] PAROLE. The PAROLE Corpus at The Swedish Language Bank. University of Gothenburg, 1997. URL http://spraakbanken.gu.se/parole. 15.2

[122] Pavel Pecina and Pavel Schlesinger. Combining Association Measures for Collocation Extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Poster Sessions*, Sydney, Australia, July 2006. 15.5

[123] Bolette Sandford Pedersen. Den danske SIMPLE-ordbog – en semantisk, ontologibaseret ordbog. In *Datalingvistisk Forenings Rrsmøde 1999, Proceedings.*, pages 71–80. København, 1999. 9.8

[124] Bolette Sandford Pedersen and Sanni Nimb. Semantic Encoding of Danish Verbs in SIMPLE – Adapting a Verb-framed Model to a Satellite-framed Language. In *Proceedings from 2nd Conference on Language Resources and Evaluation, LREC.*, pages 1405–1412. Athens, 2000. 9.8

[125] Susanne Nøhr Pedersen. The Treatment of Support Verbs and Predicative Nouns in Danish. In Jörgen Pind and Eiríkur Rögnvaldsson, editors, *Papers from the Seventh Scandinavian Conference of Computational Linguistics*, pages 208–218, Reykjavík, 1990. Institute of Lexicography. 4.6

[126] Ingemar Persson. *Das System der kausativen Funktionsverbgefüge. Eine semantisch-syntaktische Analyse einiger verwandter Konstruktionen*. Liber, Malmö, 1975. PhD thesis. 4.1, 4.2

[127] Ingemar Persson. Das kausative Funktionsverbgefüge (FVG) und dessen Darstellung in der Grammatik und im Wörterbuch. *Deutsche Sprache*, 20: 153–171, 1992. 4.1, 4.2, 4.5

[128] Sven Pihlström. *Hålla på* och *hålla på och*. *Språkvård*, 2:8–10, 1988. 4, 12.2, 2, 13

[129] Petr Piťha. On the Case Frames of Nouns. *Prague Studies in Mathematical Linguistics*, 1981. 6.3.1, 6.3.7

[130] Peter von Polenz. Funktionsverben im heutigen Deutsch. *Sprache in der rationalisierten Welt. Wirkendes Wort*, Beiheft 5, 1963. 4.1, 4.7

[131] Alain Polguère. Towards a theoretically motivated general public dictionary of semantic derivations and collocations for French. In *Proceedings of EURALEX 2000*, pages 517–527, 2000. 9.2

[132] projektet OSA – Svenska Akademiens Ordbok i databasform. Svenska akademiens ordbok, quoted 2006-09-17. URL URL:<http://g3.spraakdata.gu.se/saob/index.html>. 2

[133] PropBank Annotation. PropBank Annotation Guidelines. Version 3., February 22 2002. URL http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf. [quoted 2009-02-23]. 9.4

[134] James Pustejovsky. The Syntax of Event Structure. *Cognition*, 41:47–81, 1991. 4.4

[135] James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995. 9.8

[136] James Pustejovsky, Catherine Havasi, Jessica Littman, Anna Rumshisky, and Marc Verhagen. Towards a Generative Lexical Resource: The Brandeis Semantic Ontology. In *Proceedings of the Fifth Language Resource and Evaluation Conference*, 2006. 7.2

[137] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Longman, London ; New York :, 1985. ISBN 0582517346 0582517346 0582517346 0582517346 0582517346. 3

[138] Mikael Reuter. Lägg ribban högt. *Reuters Ruta*, Forskningscentralen för de inhemska språken 1986. URL http://www.kotus.fi/svenska/reuter/Kotimaistenkieltentutkimuskeskus.quoted2003-04-16. 4, 15.5

[139] Veronika Řezníčková. Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora. In Petya Simov, Kiril//Osenova, editor, *Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003)*, pages 88–97, Lancaster, 2003. UCREL, Lancaster University. ISBN 1-86220-134-X. 4.6

[140] Annely Rothkegel. *Feste Syntagmen. Grundlagen, Strukturbeschreibung und automatische Analyse*. Linguistische Arbeiten. Niemeyer, Tübingen, 1973. 4.1, 4.2, 4.3

[141] Ruth Feil. Funktionsverber i det danske sprog. In *Nordiske Studier i Leksikografi. Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni*

1995, volume 3 of *Skrifter utgitt av Nordiske forening for leksikografi*, pages 137–148, Reykjavík, 1995. 1

[142] Pavel Rychlý and Pavel Smrž. Manatee, Bonito and Word Sketches for Czech. In *Proceedings of the Second International Conference on Corpus Linguisitcs*, pages 124–131, 2004. ISBN 5-288-03531-8. URL http://www.fit.vutbr.cz/research/view_pub.php?id=7700. 15.2

[143] Jan Schroten. Light Verb Constructions in Bilingual Dictionaries. In *From Lexicology to Lexicography*, pages 83–94. University Utrecht. Utrecht Institute of Linguistics OTS., Utrecht, 2002. 4.2, 4.3, 8

[144] Jiří Semecký and Silvie Cinková. Constructing an English Valency Lexicon. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 94–97, Sydney, Australia, July 2006. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W06/W06-0612. 9.5.3

[145] Gilles Sérasset. A Generic Collaborative Platform for Multilingual Lexical Database Development. In *COLING 2004, Multilingual Linguistic Resources*, pages 73–79, Geneva, Switzerland, Aug. 2004. 9.2.4, 17.6

[146] Petr Sgall. *Generativní popis jazyka a česká deklinace*. Prague:Academia, 1967. 6.1

[147] Petr Sgall. Underlying Structure of Sentence and its Relation to Semantics. In T. Reuther, editor, *Wiener Slawisticher Almanach. Sonderband 33*, pages 273–282. Institut für Slavische Philologie, Universität München, 1992. 6.1

[148] Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht:Reidel Publishing Company and Prague:Academia, 1986. 6.1

[149] John Sinclair. *Corpus, Concordance, Collocation*. Oxford University Press, 1993. 3rd edition, 1st edition 1991. 1.1, 2.2, 3.1, 4.2, 10.1

[150] Otakar Smrž, Viktor Bielický, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. Prague Arabic Dependency Treebank: A Word on the Million Words. In *Proceedings of the Workshop on Arabic and Local Languages (LREC 2008)*, pages 16–23, Marrakech, Morocco, 2008. 1

[151] Språkbanken. Språkbanken vid Göteborgs universitet, since 1975. URL http://spraakbanken.gu.se. 3

[152] U. Teleman, S. Hellberg, and E. Andersson. *Svenska Akademiens grammatik*. Svenska Akademien/Norstedts, Stockholm, 1999. 1.2, 10.2, 2, 3, 4, 6, 8, 11.3.3, 12.2, 12.4, 12.6, 12.6, 13

[153] Maria Toporowska-Gronostaj and Karen Warmenius. SWEDISH SIMPLE – Lexicon Documentation. Technical report, Språkdata, University of Gothenburg, 2000. URL http://www.ub.es/gilcub/SIMPLE/reports/simple/D312_{}GOTrev.html. [quoted 2003-03-10, S.C.]. 9.8

[154] Harald Trost. Morphology. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, chapter I. – Fundamentals, pages 25–47. Oxford University Press, Oxford, 2003. 6

[155] Åke Viberg. Fysiska kontaktverb i svenskan. En skiss. *Svenskans beskrivning*, 14:174–185, 1984. 15.5

[156] Åke Viberg. Svenskans lexikala profil. *Svenskans beskrivning*, 17:391–408, 1990. 10.1, 15.5

[157] Åke Viberg. The Meanings of Swedish *dra pull*: a Case Study of Lexical Polysemy. In *EURALEX'96 Proceedings*, volume 1, pages 293–308, Göteborg, 1996. Dept. of Swedish, Göteborg University. 15.5

[158] Åke Viberg. Cross-linguistic lexicology. The case of English *go* and Swedish *gå*. In *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*, volume 88 of *Lund Studies in English*, pages 157–186, Göteborg, 1996. Dept. of Swedish, Göteborg University. 15.5

[159] Åke Viberg. Polysemy and Differentiation in the Lexicon. Verbs of Physical Contact in Swedish. In Jens Alwood and Peter Gärdenfors, editors, *Cognitive Semantics. Meaning and Cognition*. John Benjamins, Amsterdam, 1998. 15.5

[160] Åke Viberg. *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday*, volume 19, chapter What One Verb can Do: The Swedish Verb *göra* in a Cross-linguistic Perspective, pages 243–257. The Linguistic Association of Finland, Turku, special supplement to sky journal of linguistics edition, 2006. 10.1

[161] Åke Viberg. Polysemy and Disambiguation Cues across Languages. The Case of Swedish *få* and English *get*. In S. Granger and B. Altenberg, editors, *Lexis in Contrast*, pages 119–150. John Benjamins, 2002. 10.1, 13, 15.5

[162] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, 1998. 9.9

[163] Leo Wanner, editor. *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Language Companion Series (SLCS)*, Amsterdam-Philadelphia, 1996. John Benjamins. 8.1

[164] Alison Wray. *Formulaic Language and the Lexicon*. Cambridge University Press, 2002. 5.2

[165] XMLmind. XMLmind XML Editor Personal Edition 3.7.0. free software version, www.xmlmind.com/xmleditor/, 2000–2007. 16.2

[166] Zdeněk Žabokrtský. *Valency Lexicon of Czech Verbs (PhD thesis)*. PhD thesis, Charles University, Prague, Czech Rep., 2005. 6.2, 9.5.2

[167] Zdeněk Žabokrtský and Markéta Lopatková. Valency Frames of Czech Verbs in VALLEX 1.0. In Adam Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 70–77, Boston, 2004. Association for Computational Linguistics. 6.3.6

[168] Heike Zinsmeister and Ulrich Heid. Collocations of Complex Words: Implications for the Acquisition with a Stochastic Grammar. In *Proceedings of the International Workshop on Computational Approaches to Collocations Vienna*, Vienna, Austria, 22., 23. July 2002 2002. Austrian Research Institute for Artificial Intelligence (öFAI). URL http://www.ofai.at/~brigitte.krenn/colloc02/index.html. 15.4.4

Part V

Appendices

Chapter 19

Data sample

sätta

Idiom:1

[[Human, Entity, Event--thwarter]]ACT-obl ***sätter*** [[Physical Object--obstacle]]DPHR-obl *käpp i hjulet/hjulen en käpp i hjulet/hjulen käppar i hjulen inte någon käpp i hjulet/hjulen* [[Human, Entity, Event, Process, Plan--victim]]PAT-obl **för**
(för att förhindra dess funktion.)
(PAT is conceived as the correct function of a wheel system in a machine or a vehicle, and the wheels are to be stopped by placing a stick in a wheel.)

[[Human, Entity, Event--thwarter]]ACT-obl **+1,** ***překazí kazí*** [[Action, Activity-- effort]]PAT-obl **+4,** [[Human, Entity, Event--victim]]ADDR-obl **+3,**
[[Human--thwarter]]ACT-obl **+1,** [[Institution, Social System, Process, Plan--destructive, arogant machinery]]PAT-obl **+3,** ***sype*** [[--]]DPHR-obl *písek do soukolí*

[[Human--thwarter]]ACT-obl **+1,** ***sype*** [[--]]DPHR-obl *písek do soukolí něčeho*

Då gäller det att föråldrade system inte sätter käppar i hjulen för entreprenörer och entusiaster.
Det nya bidragssystemet där universitet och högskolor skall slussa igenom allt fler studenter sätter också käppar i hjulet för inträdesprov .
Personalbrist har tidigare satt käppar i hjulet för just det här behandlingshemmet.

Idiom:0

[[Human, Horse, Animal--]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta [[Route--]]DIR3-obl
(walking, running)

[[Human, Horse, Animal--]]ACT-obl ***vyrazí vydá se vykročí vyjde jde*** [[Route--]]DIR3-obl

[[Human, Horse, Animal--]]ACT-obl ***dá se*** [[--]]CPHR-obl **do+2,** ***pohyb sg***

[[Human, Horse, Animal--]]ACT-obl ***nabere*** [[--]]CPHR-obl **+4,** ***rychlost sg***

Han satte fart ner mot båtarna .
Så satte jag fart ner mot söder och kom fram till porten i det hus där Sulan bor .
De satte full fart direkt och passerade banan på 10 sekunder .
Det var något magiskt med det där sekundsnabba svindlande ögonblicket när våra blickar möttes , innan de smäckra djuren satte fart över gärdet eller in i skogen .

Idiom:0

[[Human--]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta [[Route--]]DIR-obl
(hurry up to continue the journey)

[[Human--]]ACT-obl ***vyjde jde vyrazí vydá se vykročí*** [[Route--]]DIR1-obl

Nej , nu måste vi sätta fart . Vi ses på torsdag .

Idiom:1

[[--]]ACT-obl ***sätter*** [[--]]DPHR-obl *fart på hjulen fart på hjulet*

[[--]]ACT-obl ***dá*** [[--]]PAT-obl **+3,** [[--]]CPHR-obl *impuls k činnosti*
(ACT makes a process start or speed up)

[[--]]ACT-obl ***přinese oživení*** [[--]]PAT-obl

[[Human, Institution--]]ACT-obl ***se rozhoupe*** [[Activity, Action--]]PAT-obl **k+3,**
(ACT starts an activity)

Sänkta inkomstskatter - för medelinkomsttagare från 22 till 18 procent - parade med lättnader för småföretagarna skall enligt PP sätta fart på hjulen.
Är det rutin eller erfarenhet man pratar om ? Är det inte på tiden att vi sätter fart på hjulet ? Jag vet att det på riks- och distriktsnivå inom många idrotter väljs in allt fler kvinnor och ungdomar .

Idiom:0

[[Vehicle, Human--]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta [[Route--]]DIR3-obl

[[Vehicle, Human--driver]]ACT-obl ***se rozjede nabere rychlost vyrazí*** [[Route--]]DIR3-obl

Philip Kimberly såg på när Penelopes tåg satte fart mot London .

Idiom:0

[[Human--]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta [[Activity, Action--]]PAT-obl **med**
(energetically)

[[Human--]]ACT-obl ***se dá se pustí se vrhne*** [[Activity, Action--]]PAT-obl **do+2,**

Sedan satte hon full fart med disk och avdukning , bäddning och tvätt.

Idiom:

[[Process,Activity--]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta

[[Process,Activity--]]ACT-obl **+1,** ***nabere*** [[--]]DPHR-obl **+4,** ***obrátky pl***

[[Process,Activity--]]ACT-obl **+1,** ***se rozjede***

Produktionen satte fart och landsbygden var inte längre så isolerad .

Idiom:0

[[Human--driver]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta [[Car, Motorbike, Truck, Boat, Device--]]PAT-obl **på**

[[Human--driver]]ACT-obl **+4,** ***nastartuje rozjede*** [[Car, Motorbike, Truck, Boat, Device--motor]]PAT-obl **+4,**

man pulsar genom snön till sin Lada med vev i handen , nynnande 'Pråmdragarnas sång' , och med en enda kraftfull Absolut Ren Smirnoff sätter fart på den fyrcylindriga förbränningsmotorn (nästan samma som i Volvo) och sedan bjuder grannens förklemade och degenererade åkdon på startström ...

Idiom:0

[[Human--impulse]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta [[Human, Institution--]]PAT-obl **på**
(, så att PAT sätter i gång med en (implicit) aktivitet.)

[[Human--]]ACT-obl ***sebrat se a jít něco dělat***
(and start ot take up again what has to be done)

[[Human, Entity, Event--]]ACT-obl ***uvede*** [[--]]CPHR-obl **do+2,** ***pohyb sg*** [[Human, Institution--]]PAT-obl **+4,**

[[Human, Entity, Event--]]ACT-obl ***probudí povzbudí vybudí přiměje podnítí*** [[Human, Institution--]]PAT-obl **+4,** [[Action, Activity--]]EFF-obl **k+3,** *aby*

Beskedet från ÖCB satte fart på kommunen som inte räknat med att komma igång med arbetet förrän 1997 .
Men nu måste jag se till att jag sätter fart på mig och åker tillbaka till stan.

Idiom:0

[[Human, Entity, Event--]]ACT-obl ***sätter*** [[--]]DPHR-obl ***sg bare noun*** pnl_ref:fart-saetta [[Event, Process, Activity, Action--]]PAT-obl **på**
(, så att den sätter i gång.)

[[Human, Entity, Event--]]ACT-obl ***vyvolá rozvíří odstartuje rozjede*** [[Event, Process, Activity, Action--]]PAT-obl **+4,**

[[Human, Entity, Event--]]ACT-obl ***uvede*** [[--]]CPHR-obl **do+2,** ***pohyb sg*** [[Event, Process, Activity, Action--]]PAT-obl **+4,**

En oregelbunden öppning på New York börsen satte fart på köpintresset på Stockholmsbörsen i slutskedet av måndagens handel .
Tragedin har också satt ny fart på vapendebatten i Storbritannien.
Lägre räntor sätter nämligen fart på ekonomin och skapar jobb - vilket drar upp inflationen .

Idiom:0

[[--racer]]ACT-obl ***sätter*** [[--]]CPHR-obl pnl_ref:rekord-saetta

Idiom:0

[[Human--]]ACT-obl ***sätter*** [[Body Part, Artifact--cover]]PAT-obl [[Body Part, Artifact--]]ADDR-obl **för**
(för att täcka den)

[[Human--]]ACT-obl [[Human--]]ADDR-opt **+3,** ***dá*** [[Body Part--cover]]PAT-obl **+4,** [[Body Part,

Artifact--]]DIR3-obl **před+4, na+4,**
[[Human--]]ACT-obl ***zakryje přikryje*** [[Human--]]ADDR-opt **+3,** [[Body Part, Artifact--]]PAT-obl **+4,** [[Body Part, Artifact--cover]]EFF-obl **+7,**
" Åh , nej ! " Gwen satte en hand för munnen och bleknade .
Hon satte ett finger för mun och lyssnade uppåt taket .
Tora satte handen för munnen men det syntes på ögonen att hon skrattade .
Han sätter handen för luren .
Med ett lågt utrop satte han händerna för ansiktet.
Hotell och restauranger satte vindskivor för fönstren.

Idiom:0
[[Human--]]ACT-obl ***sätter*** [[Inanimate--]]PAT-obl [[Location--]]DIR3-obl **på** ***plats sg no possessive pron insertion allowed***
[[Human--]]ACT-obl ***dá*** [[Inanimate--]]PAT-obl [[Location--]]DIR3-obl **na+4,** ***místo sg typically modified by 'své'***
Jag satte cykelkorgen på plats och låste upp cykeln.

Idiom:0
[[Human--]]ACT-obl ***sätter*** [[Human, Animal--]]PAT-obl ***typically reflexive*** [[Location, Physical Object--]]DIR3-opt
[[Human--]]ACT-obl ***posadí usadí*** [[Human, Animal--]]PAT-obl ***typically reflexive*** [[Location, Physical Object--]]DIR3-opt
Modern sätter den lilla flickan på en stol vid sidan om Charles.
Han får sätta kärringen i traktorn .
Getrud sätter sig på stolen intill skrivbordet .
Jag kunde inte bara sätta mig på tåget och resa bort ifrån alltihop.
Han satte sig på sin plats i dubbelbänken och beredde sig på allting.
Hon satte sig på platsen bredvid honom och slätade till sin korta kjol .
Så fick jag sätta mig på toaletten.

Idiom:1
[[Human--]]ACT-obl ***sätter*** [[Human--]]PAT-obl ***never reflexive*** [[--]]DPHR-obl ***sg no no adjective****på plats* [[Speech Act, Action--]]MEANS-typ **med genom**
(After the given action/speech act, the counterpart is not able to fight back immediately.)
[[Human--]]ACT-obl ***usadí*** [[Human--]]PAT-obl [[Speech Act, Action--]]MEANS-typ
Äntligen , nu har vi satt dem på plats ! Nu vågar man vara socialdemokrat igen , jublade Spöri i mobiltelefonen.
Därmed satte Ingrid samtliga sina landslagstjejer på plats under de avslutande åtta serierna .

Idiom:0
[[Human, Device--]]ACT-obl ***sätter*** [[Physical Object --]]PAT-obl [[Location, Physical Object--]]DIR3-obl
(where it is meant to come, and the entity to be placed is not perceived as primarily vertical or primarily horizontal)
[[Human, Device--]]ACT-obl ***dá umístí usadí strčí zastrčí připevní přibije přilepí přišpendlí přišije přitiskne nasadí vloží přiloží zasune*** [[Physical Object--]]PAT-obl [[Location, Physical Object--]]DIR3-obl
hatten på termosflaskan
locket på tuben
nyckeln i låset
kaffekopp på bordet
klämmor i håret
handen bakom Andros rygg
händerna mot höfterna
extra glans i ögat eller på kinden

pilsnern i halsen
en guldkrona på hennes huvud
och så satte hon pekfingret under hakan på Franzon
en prydnad på blusen
sina tänder i skalet
Han satte pappformen med pommes-friten till munnen , stjälpte huvudet bakåt och hällde i sig de sista smulorna och saltkornen .
Grabben satte tummen över öppningen , skakade flaskan och räckte den till Reine .
fötterna på pedalerna
korten i album
rosor i vas
fyr på bilen
hatten på huvudet
ribban så högt att man måste vara världsmästare för att komma över
Hon satte örat till dörren och tyckte att hon hörde ett hasande ljud inne i rummet.

Idiom:
[[Human, Animal, Device--]]ACT-obl ***sätter*** [[Body Part, Weapon--edge, claws, teeth, nails]]PAT-obl [[Location--]]DIR3-obl **i**
(where it is meant to come, and the entity to be placed is not perceived as primarily vertical or primarily horizontal)
[[Human, Animal, Device--]]ACT-obl ***zatne zabodne zapíchne*** [[Body Part, Weapon--edge, claws, teeth, nails]]PAT-obl [[Location, Physical Object--]]DIR3-obl **do+2,**
Jag satte tänderna i hans axel och bet till så hårt jag kunde .
... kan man ibland komma ovanför björnar och sätta ett spjut eller en pil i dem innan de ser en .

Idiom:0
[[Human, Institution, Military--authority]]ACT-obl ***sätter*** [[Human--guilty]]PAT-obl [[Location, Institution--prison]]DIR1-obl **i**
[[Human, Institution--authority]]ACT-obl ***vsadí*** [[Human--guilty]]PAT-obl [[Location, Institution--prison]]DIR1-obl **do+4,**
Åtta biskopar plus ett stort antal präster sattes i fängelse .
Polisen lyfter bort en fyllerist från tunnelbanan och sätter honom i en cell på polisstationen .

Idiom:0
[[Human, Institution, Concept, Document, Social System, Medium--]]ACT-obl ***sätter*** [[Concept--]]PAT-obl [[Concept--]]DIR3-obl **i**
[[Human, Institution, Concept, Document, Social System, Medium--]]ACT-obl ***staví*** [[Concept--]]PAT-obl [[Concept--]]DIR3-obl **do+4,**
människans behov i centrum
arbetsinsatsen i proportion till indrivna bötespengar .
För alla verksamma inom rättssystemet borde det vara naturligt att sätta människan i fokus .

Idiom:0
[[Human--]]ACT-obl ***sätter*** [[Human--]]DIR1-obl **från** ***sig probably just reflexive*** **ifrån** ***sig probably just reflexive*** [[Artifact, Food, Container, Beverage--]]PAT-obl
(och gör därmed klart at han inte vill ha mera)
[[Human--]]ACT-obl ***odstrčí*** [[Artifact, Food, Container, Beverage--]]PAT-obl **+4,**
Hon satte ifrån sej brickan på bordet och gick fram mot ena väggen .

Idiom:0
[[Human--]]ACT-obl ***sätter*** [[Device--adjustable control]]PAT-obl [[Mathematical Value, Plan, Service--]]EFF-opt **på**
[[Human--]]ACT-obl ***nastaví nařídí*** [[Device--adjustable control]]PAT-obl **+4,** [[Mathematical Value, Plan, Service--]]EFF-opt **na+4,**

Han stiger upp igen och sätter väckarklockan på ringning
skrek Boel " Jättebra " och satte duschen på högsta fräs och lät vattnet rinna över sitt huvud .

Idiom:0

[[--]]ACT-obl ***sätter*** [[Artifact--imprint, trace, footprint]]PAT-obl pnl_ref: [[Location--]]DIR3-obl

[[--]]ACT-obl ***zanechá*** [[Picture, Property--imprint, trace, footprint, fingerprint]]PAT-obl **+4, *stopa, otisk*** [[Location--]]LOC-obl

sin lilla hund som satte jordiga tassavtryck på den rosarutiga dräktjackan .
...vad det var som hade satt sina spår runt liket .
Tårarna rann nerför kinderna och satte spår i make-upen .

Idiom:0

[[Eventuality, Event--]]ACT-obl ***sätter*** [[--trace]]CPHR-obl pnl_ref:spaar-saetta [[Physical Object, State, Property--]]PAT-opt

[[--]]ACT-obl ***zanechá*** [[Picture, Property--imprint, trace, footprint, fingerprint]]CPHR-obl **+4, *stopa*** [[Physical Object, State, Property--]]PAT-opt

Idiom:1

[[Anything--hint, advisor]]ACT-obl ***sätter*** [[Human, Animal, Institution--pursuer]]PAT-obl [[--]]DPHR-obl pnl_ref:spaar-paa-satta

[[--hint, advisor]]ACT-obl ***přivede*** [[Human, Institution--pursuer]]PAT-obl **+4,** [[--]]DPHR-obl **na+4,** *(správnou/špatnou) stopu*

Idiom:0

[[Physical Object, Event--]]ACT-obl ***sätter*** [[Cognitive State--]]CPHR-obl [[Human, Human Group, Animal--]]PAT-obl **i**

[[Physical Object, Event--]]ACT-obl ***vyvolá vzbudí probudí*** [[Cognitive State--]]CPHR-obl **+4,** [[Human, Human Group, Animal--]]PAT-obl **v+6,**

Han har förstås lagt dit dom för att sätta skräck i mej .
vederlägga pastor Öhrström och sätta kurage i hans döttrar.

Idiom:

[[Human, Human Role--teacher, judge]]ACT-obl ***sätter*** [[Information--score]]CPHR-obl [[Human, Activity, Action--pupil]]PAT-obl **på**

[[Human, Human Role--teacher, judge]]ACT-obl ***vystaví vystavuje dá dává*** [[Information--score]]CPHR-obl **+4, *vysvědčení*** [[Human--pupil]]PAT-obl **+3,**

[[Human, Human Role--teacher, judge]]ACT-obl ***udělí uděluje*** [[Information--score]]CPHR-obl **+4, *známka*** [[Human--pupil]]PAT-obl **+3,**

[[Human, Human Role--teacher, judge]]ACT-obl ***známkuje oznámkuje*** [[Human--pupil, performance]]PAT-obl **+4,**

censorerna var också till för att sätta betyg på lärarna

Idiom:

[[Human, Institution, State--]]ACT-obl ***sätter*** [[Concept, Rule, Mathematical Value--]]PAT-obl pnl_ref:pris-sätta-1

[[Human, Institution, State--]]ACT-obl ***určí určuje stanoví vymezí*** [[Concept, Rule, Mathematical Value--]]PAT-obl

Och dels vill de väl kunna reglera lönerna på det sättet , eftersom de själva sätter priserna på materialen .
Trots rörlig växelkurs kan riksbanken inte självständigt sätta räntorna .
Ändå visar exemplen tydligt hur svårt det kan vara för enskilda fondkommissionärsfirmor att sätta gränser .
Detta kan sätta en mental gräns för fortsatt integration .

Idiom:1

[[Human, Institution--authority]]ACT-obl ***sätter*** [[--]]PAT-obl pnl_ref:pris-sätta-2

[[--authority]]ACT-obl ***vypíše*** [[--]]PAT-obl **+4, *odměna sg na/za něčí hlavu***

Till och med den liberala amerikanska rättvisan tröttnade så småningom på detta näringsfång , dömde honom till döden och satte ett pris på hans huvud .

Idiom:0

[[Human--criminal]]ACT-obl ***sätter*** [[--]]DPHR-??? [[Physical Object--]]PAT-obl **på**
(always an evil and harmful deed)

[[Human--criminal]]ACT-obl ***zapálí podpálí*** [[Physical Object--]]PAT-obl

[[Human--criminal]]ACT-obl ***založí*** [[--]]CPHR-obl **+4,** ***požár něčeho, někde***

Jonna är rädd för att Johannes vill sätta eld på huset .

Idiom:1

[[--]]ACT-obl ***sätter*** [[--]]DPHR-obl ***no*** *krokben* [[Human--victim]]PAT-obl **för**
(to make PAT stumble)

[[--]]ACT-obl ***nastaví*** [[--]]DPHR-obl **+4,** ***noha sg*** [[Human--victim]]PAT-obl **+3,**

Plötsligt var det någon som satte krokben för honom och han föll handlöst till marken , yr och mörbultad .

Idiom:0

[[--]]ACT-obl ***sätter*** [[Physical object--]]PAT-??? ***fälla*** [[Animal, Human--prey]]ADDR-obl

[[Human--hunter]]ACT-obl ***straží nastraží*** [[Physical object--]]PAT-??? **+4,** ***past*** [[Animal, Human--prey]]ADDR-obl **na+4,**

Den enkla sanningen är den att de är för riskabla att jaga om man inte sätter fällor för dem .

Idiom:0

[[--]]ACT-obl ***sätter*** [[Information--punctuation]]PAT-obl ***kryss, frågetecken, parentes, utropstecken, punkt*** [[--]]DIR3-obl

[[--]]ACT-obl ***udělá napíše*** [[Information--punctuation]]PAT-obl **+4,** ***křížek, otazník, závorka, vykřičník, tečka*** [[--]]DIR3-obl

I höst får du sätta kryss framför din favoritpolitiker . EU-valet den 17 september ...

Idiom:1

[[--]]ACT-obl ***sätter*** [[Information--punctuation]]DPHR-obl ***frågetecken*** [[Information, Concept--]]PAT-obl **för framför**

[[--]]ACT-obl ***udělá napíše*** [[Information--punctuation]]PAT-obl **+4,** ***otazník*** [[Information, Concept--]]PAT-obl **u+2, k+3,**

[[--]]ACT-obl ***zpochybní zpochybňuje*** [[Information, Concept--]]PAT-obl **+4,**

Ett annat företag på listan man skulle kunna sätta ett frågetecken framför är WM-data .
Men professor Radetzki sätter ett frågetecken för de alternativa energislagens möjlighet att bli konkurrenskraftiga gentemot olja , gas och kol längre fram på 2000-talet .

Idiom:1

[[--]]ACT-obl ***sätter*** [[Information--punctuation]]DPHR-obl ***utropstecken*** [[Information, Concept--]]PAT-obl **efter**

[[--]]ACT-obl ***udělá napíše*** [[Information--punctuation]]PAT-obl **+4,** ***vykřičník*** [[Information, Concept--]]PAT-obl **u+2, k+3, za+4,**

[[--]]ACT-obl ***zdůrazní zdůrazňuje*** [[Information, Concept--]]PAT-obl **+4,**

Det är därför jag sätter ett utropstecken på Folkrörelsealliansens förslag till budget 2006 och planperioden.

Idiom:1

[[--]]ACT-obl ***sätter*** [[Information--punctuation]]DPHR-obl [[Information, Concept--]]PAT-obl

[[--]]ACT-obl ***udělá napíše*** [[Information--punctuation]]PAT-obl **+4,** ***vykřičník*** [[Information, Concept--]]PAT-obl **u+2, k+3, za+4,**

[[--]]ACT-obl ***zdůrazní zdůrazňuje*** [[Information, Concept--]]PAT-obl **+4,**

Det är därför jag sätter ett utropstecken på Folkrörelsealliansens förslag till budget 2006 och planperioden.

Idiom:1

[[--]]ACT-obl ***sätter*** [[--]]DPHR-obl pnl_ref:saetta-punkt-1 [[Activity, Process--]]PAT-obl **för**

[[--]]ACT-obl ***udělá*** [[--]]DPHR-??? **+4,** ***tečka sg*** [[Activity, Process--]]PAT-obl **za+7,**

Idiom:1
[[--]]ACT-obl ***sätter*** [[--]]DPHR-obl pnl_ref:saetta-finger-1 [[--something not obvious]]PAT-obl **på**
(to define or explain it)
[[--]]ACT-obl ***ukáže*** [[--]]DPHR-??? **+7,** ***prst sg*** [[Activity, Process--]]PAT-obl **na+4,**
Den är så enkel att ett par av mina favoritförfattare , Carl-Henning Wijkmark och Tobias Berggren , på endast ett par minuter i sina respektive Sommarprogram i radions P 1 på ett klargörande sätt lyckades sätta fingret på vad det handlar om .

Idiom:0
[[--]]ACT-obl ***sätter*** [[Animate--offspring]]PAT-??? [[--]]DPHR-obl ***värld sg def*** *till världen*
[[--]]ACT-obl ***přivádí přivede*** [[Animate--offspring]]PAT-??? **+4,** [[--]]DPHR-obl ***svět sg*** *na svět*
Andra djur som vargar , schimpanser , örnar och myror sätter många ungar till världen när de har chansen .

Idiom:0
[[--]]ACT-obl ***sätter***
[[--]]ACT-obl
sätta potatis

Idiom:1
[[--]]ACT-obl ***sätter***
[[--]]ACT-obl
Men dess förslag sätter i själva verket strålkastarljuset på ett långt mycket större svenskt problem...

Idiom:1
[[--]]ACT-obl ***sätter***
[[--]]ACT-obl
Och nu skulle han inte sätta sin fot i den här hålan förrän till rättegången om han kunde slippa .

sätta sig

Idiom:0
[[Human--]]ACT-obl ***sätter sig*** [[--]]DPHR-obl **på** ***huk sg no*** [[--]]DIR3-opt
[[Human--]]ACT-obl ***se posadí si sedne*** [[--]]DPHR-obl *na bobek* [[--]]DIR3-opt

Idiom:0
[[--]]ACT-obl ***sätter sig***
Det är alldeles för mycket mjöl i pannkakorna . Dom sätter sej i tänderna .
Lukten hade satt sig i kläderna . usadit se, ulpivat
Det är plusgrader , isen suger - yxan sätter sig djupt.

Idiom:0
[[Human--]]ACT-obl ***sätter sig*** [[--]]CPHR-obl

PNL - Predicate Noun Lexicon

rekord neutrum rekord

personlig svensk tidigare svårslagen oslagbar dyster alla
världsrekord publikrekord tittarrekord turistrekord kassarekord insamlingsrekord försäljningsreport hundraårsrekord
utan segerchans på löpande band

slå ID:rekord-slaa Oper1 Liqu telic punctual volitional

překonat rekord
překonávat rekordy
zlomit rekord
lámat rekordy
pokořit rekord
trhnout rekord
překročit rekord

With postpositive definite article only:

Målet är att slå svenska rekordet 1.45 ,92.
Världsrekordet håller sovjetiska geologer som nått tolv kilometers djup på Kolahalvön förra året . Nu ska västtyska forskare slå rekordet .
...är tre hundradelar snabbare än Strenius veckogamla notering . - Ändå gjorde jag inget perfekt lopp . Det finns mer att hämta , berättar Sveriges sprinterkung på telefon från Cottbus . - Men det satt skönt att slå rekordet .

With both pre- and postpositive definite article:

Vinnartiden 1 . 12,7 är bara två tiondelar från Sugarcane Hanovers världsrekord över medeldistans . - Med lite bättre väder hade vi slagit det rekordet , säger Erik Berglöf .
Så sent som förra lördagen satte han svenskt rekord i spjutkastning för sin årsklass . - I regn och ruskväder nere i Lund . Den som siktar på att slå det nya rekordet måste kasta 28,26 .
En häpnadsväckande segermarginal för Isabelle , som slog det tidigare BOC-rekordet för denna sträcka med drygt två dygn.

With possessive pronoun or genitive:

Detta trots det faktum att filmen nästan hade slagit veckans kassarekord.
Under hans ledning har landslaget slagit Platiniepokens rekord.
Bestsellern Golf har nu passerat tillverkningssiffran 16 miljoner , har därmed slagit T-Fordens gamla rekord , men har fortfarande en bit kvar till ur-Folkans världsrekord på 22 miljoner bilar.
...som häromdagen slog sitt personrekord i blåbärsplockning i Luleskogarna.
Agneta inledde med att i försöken slå sitt eget svenska rekord med 21/100-delar, noterade 1.02 ,00.

With other no-article determiner:

kassafilmer som slog alla rekord.
Den har nyss haft premiär och slagit alla tänkbara rekord.
Fredagsnatten slog alla rekord med omkring 90 omhändertagna personer och en fullbelagd arrestlokal .

ACT-:

possgen *, , , also reflexive* [[Human, Animal, Artifact=racer]] Implicature:
Bestsellern Golf har nu passerat tillverkningssiffran 16 miljoner , har därmed slagit T-Fordens gamla rekord , men har fortfarande en bit kvar till ur-Folkans världsrekord på 22 miljoner bilar .

PAT-:

i [[Activity, Property=discipline]] Implicature:
i längdhopp

i omdömeslöshet

på *, , , contains a numeric expression* [[Activity=subdiscipline]] Implicature: specified with amount/manner of performance

på 400 meter häck

EFF-:

med [[Mathematical Value=performance]] Implicature:

Produktionen under april slog nytt rekord med en genomsnittlig dagsproduktion av 3,1 miljoner fat olja.

genom [[Activity, Action=performance]] Implicature: the way in which the record was set

nyligen slog rekord genom att köra 189 mil på 40 liter bensin.

REG-typ:

för *, , , plural* [[Vehicle, Human, Horse=racers]] Implicature: group defined by a feature, e.g. age limit or weight

för enskrovsbåtar över Atlanten från väst till öst

för *, , , mostly singular* [[Property,Activity,Route=discipline]] Implicature:

för snödjup

för sträckan New York-San Francisco

för *, , , mostly singular* [[Vehicle, Human, Institution, Horse, Artifact=performer]] Implicature: performances of one single entity are compared

även öresundsbron kommer med sin fria spännvidd på ungefär 500 meter att slå världsrekord för järnvägsbroar .

TFRWH-typ:

från [[Action,Time Point=occasion]] Implicature: on which a relevant record was set

från VM i Rom förra året

vid [[Action=occasion]] Implicature: on which a relevant record was set

vid en tävling i östtyska Cottbus

för [[Time Period=time since when the latest relevant record was set]] Implicature:

hennes rekord för fyra år sen

TFHL-typ:

för *, , , noun phrase* [[Time Period=]] Implicature: since the former record or for a given period in which the performances were observed

för i år

possgen *, , , noun genitive* [[=Time Period]] Implicature: time span since the former record or for a given period in which the performances were observed

veckans kassarekord

sätta ID:rekord-saetta Oper1 telic punctual volitional

dosáhnout rekordu

vytvořit rekord

udělat rekord

ustanovit rekord

As bare noun:

commonest definiteness

Den karismatiske höjdhopparen tog EM-guld i Split i Kroatien 1990 , satte juniorvärldsrekord och vann junior-VM samma år.

Där satte hon nytt Julklappsrekord och nytt personligt rekord med 10,41 .

varje simmare som sätter personligt rekord ärligen har förtjänat resan till Strasbourg

Geir Karlstad satte världsrekord på distansen vid tävlingar i Heerenveen.

I fjol satte malmöborna rekord för Guinness egen bok med 84_000 kräftor.

Bäst på Chart Hills-banan denna dag var dock skotskan Dale Reid , som satte nytt banrekord med 66 slag .
Uddevallatjejen Louise Karlsson satte nytt rekord på 200 m medley .
Basketlaget Detroit Pistons har chansen att sätta nytt publikrekord för en säsong i proffsligan NBA .

With indefinite article:

När han väl var i luften verkade ribban ligga flera meter under honom , han satte ett oslagbart världsrekord !
Det var det år då priserna steg med 5000 procent och Argentina satte ett svårslaget världsrekord i hyperinflation .
Gebrselassie satte ett fantastiskt världsrekord på 5_000 meter vid DN Games i Globen den 20 februari

With possessive pronoun or genitive:

På Stadion , där Patrik satte sitt världsrekord (2.42) för mindre än två månader sedan , gick han nu in redan på 2.16 .
Henrik Sjöberg från Skara var målvakt när Andreas Andersson satte sitt personliga målrekord .

With other no-article determiner:

Om jag sätter något rekord så kommer det snart någon och slår det .

ACT-:

possgen [[Human, Animal, Artifact=racer]] Implicature:

PAT-:

i [[Activity, Property=discipline]] Implicature:
i fallskärmshopp
i höjdhopp inomhus
i spjutkastning
i antalet utvisningar för en målvakt
i hyperinflation

på *, , , contains a numeric value or a quantifier* [[Activity=subdiscipline]] Implicature: specified with amount/manner of performance
på 200 meter bröst
på 300 meter
på distansen

EFF-:

med *, , , contains a numeric value or a quantifier* [[Mathematical value=performance]] Implicature:
Han satte nytt svenskt rekord med 1.48 ,89 min
Det året satte firma Steiff rekord med 974_000 tillverkade nallar.

genom [[Activity, Action=performance]] Implicature:
Patrik Stenlund satte gårdagens andra svenska inomhusrekord genom att klara 5,70 m i stavhopp i Skellefteå .

notera ID:rekord-notera Oper1 telic punctual volitional

připsat si rekord
zaznamenat rekord

As bare noun:

commonest definiteness
Astraaktierna noterade kursrekord : Astra A steg 7:50 kronor till 247:50 .
Beijers samtliga aktieslag noterade igår nya kursrekord .
Det Skåne-baserade utvecklingsbolaget Actives båda aktieslag noterade kursrekord på torsdagen
Tre gånger noterade han världsrekord på den klassiska bandistansen 5 engelska mil

With indefinite article:

an gjorde det på 1_500 meter där han noterade ett praktfullt nytt svenskt rekord med tiden 15.17 ,01 .

With postpositive definite article only:

Under våren tränades han dock upp rejält , och första säsongen travade han in 194_115 kr och noterade

rekordet 1.30 ,8 vid seger .

With possessive pronoun or genitive:

VM-fotbollen noterade tidernas publikrekord med 68604 åskådare i snitt på 52 matcher .

ACT-:

possgen *, , , not realized* [[Human, Animal, Asset=racer]] Implicature:

EFF-:

, , , numeric expression [[Mathematical Value=performance]] Implicature:

noterade rekordet 1.30 ,8

med [[Mathematical Value=performance]] Implicature:

VM-fotbollen noterade tidernas publikrekord med 68604 åskådare i snitt på 52 matcher .

PAT-:

på [[Activity, Route=discipline]] Implicature: specified by amount/way of performance required

hon hade noterat ett nytt svenskt rekord på 5_000 meter

stopp neutrum stopp

sätta ID:stopp-saetta Oper1 Caus telic punctual volitional

zastavit něco

učinit něčemu přítrž

As bare noun:

Den inflytelserika militären har hittills satt stopp för en säkerhetspolitisk kursändring för Turkiets del .

Hagman spelade i AIK , Djurgården och Spånga innan en knäskada satte stopp för karriären .

Han säger sig vara tillräckligt disciplinerad för att sätta stopp vid i genomsnitt 40 timmars arbete per vecka .

Vi är inte oroliga för att regeringen sätter stopp för Telias planerade Guld-TV-kanal .

När Kroatien utropades som egen stat dröjde det inte länge förrän kriget satte stopp för all turism under flera år .

PAT-:

för [[Activity, Action, Plan, Process=]] Implicature:

fart utrum farter

full ny

sätta ID:fart-saetta-1 NA

RSTR-opt:

, , , full [[=]] Implicature:

, , , ny [[=]] Implicature:

namn neutrum namn

sätta ID:namn-saetta Oper1

pojmenovat

As bare noun:

Hon kunde inte sätta namn på det hon kände .

APP-:

på [[Entity, Eventuality, Event, Property=name-bearer]] Implicature: refer to with its correct name

respekt utrum none

respekt ID:respekt-i-saetta-sig Labor2_1 Incep Caus

As bare noun:

Öberg är skickligast i Sverige på att kunna hålla kontroll på sina matcher , utan att behöva sätta sig i respekt genom att utdöma en massa tidiga utvisningar .
Jag försöker att sätta mig i respekt hos Ronaldo på ett tidigt stadium .

ACT-:

hos [[Human, Institution,Animal=]] Implicature:

MEANS-typ:

genom [[Activity, Action, Property=]] Implicature:

spår neutrum spår

djup drastisk outplånlig tydlig

sätta ID:spaar-saetta Oper1 NA non-volitional

zanechat stopy
zanechávat stopy

As bare noun:

Anpassningen till Schengenavtalets omfattande regelverk har redan satt spår i svensk politik .
Dessutom är vi inne i ett rätt så markant generationsskifte . Sådant bör också sätta spår .
Lidandet och orättvisan har satt outplånliga spår i deras själar .

With possessive pronoun or genitive:

Most frequently used with reflexive pronoun in the abstract reading
Misshandeln och hoten har redan satt sina spår psykiskt .
Baby-boomen i mitten av åttiotalet har satt sina spår .
...., sa Hoddle , som redan efter ett år som förbundskapten har satt sina spår i Europas mest konservativa fotbollsland .

sätta ID:spaar-paa-saetta NA telic punctual NA

As bare noun:

probably only with adjective insertion
Andra lyckosamma prestationer svarade backen mats Alba för tack vare två slagskott , som satte stockholmarna på rätt spår .
För en desinformatör med avsikter att sätta spaningsledningen på fel spår borde PKK ligga väl till hands .

With postpositive definite article only:

... en man från skolstyrelsen hade varit där och frågat efter Lucy , att det var Alice Broughton som hade satt honom på spåret.

RSTR-opt:

, , , fel, rätt [[=]] Implicature: harms or helps the pursuit

pris 1 neutrum_utrum pris, priser

sätta ID:pris-sätta-1 NA NA volitional

As bare noun:

rare, probably just plural neutrum
Det finns många olika sätt att sätta pris och att ta betalt och där hoppas han på en större spridning än på Minitel .

With indefinite article:

Margita Björklund menar att det inte går att sätta ett pris på Energibolaget
Vi satte ett pris , mäklaren ordnade visning och vi fick tre intresserade spekulanter .

With postpositive definite article only:

rare
dessa människor ville nu själva sätta priset på sina arbetstimmar

APP-:

på , , , [[=]] Implicature:

sätta ID:pris-sätta-2 NA telic volitional

DPHR-:

, , , [[=]] Implicature: ref: phr-pahuvud

på huvud NA NA

APP-obl:

, , , *reflexive pronoun* [[=]] Implicature: ref:

punkt utrum punkter

sätta ID:punkt-saetta-1 NA

As bare noun:

commonest definiteness

Inom en halvtimme , när nattmörkret satte punkt för blodbadet , hade allt beklanskt motstånd krossats .
Kalla kriget satte punkt för en epok i världshistorien .
Ingvar Carlsson har satt punkt för sin gärning som statsminister och partiledare

With indefinite article:

probably triggered by adjective modifier

Manhem satte en fin punkt för sin absolut bästa säsong någonsin i elitserien .

Chapter 20

A Sample of the Lemmatizer Script

```
#lempas2 adding 'aeiouyöää'
#The following 3 lines resolve plural indefinites of uters and neuters ending with 'lerna', 'larna', 'narna', 'nerna' and 'rarna':
s/^\([^\t]*[^laeiouyöää]\)\(l[ae]rna\)\t\(NC[UN]PN@DS\)$/\1\2\t\1el\t\3/;
s/^\([^\t]*[^naeiouyöää]\)\(n[ae]rna\)\t\(NC[UN]PN@DS\)$/\1\2\t\1en\t\3/;
#s/^\([^\t]*[^raeiouyöää]\)\(rarna\)\t\(NC[UN]PN@DS\)$/\1\2\t\1er\t\3/;

#Nouns, 2nd and 3rd declensions (both '-a' and '-e' declensions) plural indefinite: katter-katt, stolar-stol. Tag: NCUPN@IS substantiv utrum pluralis obestamd nominativ.
s/^\([^\t]*\)\([ae]r\)\t\(NCUPN@IS\)$/\1\2\t\1\t\3/;

#Nouns, 2nd and 3rd declensions (both '-a' and '-e' declensions) plural definite: katterna-katt, stolarna-stol. Tag: NCUPN@DS substantiv utrum pluralis bestamd nominativ
s/^\([^\t]*\)\([ae]rna\)\t\(NCUPN@DS\)$/\1\2\t\1\t\3/;

#DEFINITE SINGULAR UTERS
#The following 2 lines resolve singular definite of uters ending with 'ln', 'rn':
s/^\([^\t]*\)\(ln\)\t\(NCUSN@DS\)$/\1\2\t\1l\t\3/;
s/^\([^\t]*\)\(rn\)\t\(NCUSN@DS\)$/\1\2\t\1r\t\3/;

#-slen
s/^\([^\t]*\)\(slen\)\t\(NCUSN@DS\)$/\1\2\t\1sle\t\3/;

#-klen
s/^\([^\t]*\)\(klen\)\t\(NCUSN@DS\)$/\1\2\t\1kel\t\3/;

#-plen
s/^\([^\t]*\)\(plen\)\t\(NCUSN@DS\)$/\1\2\t\1pel\t\3/;

#-tlen
s/^\([^\t]*\)\(tlen\)\t\(NCUSN@DS\)$/\1\2\t\1tel\t\3/;

#-mlen
s/^\([^\t]*\)\(mlen\)\t\(NCUSN@DS\)$/\1\2\t\1mmel\t\3/;


#The general rule: Nouns, 2nd and 3rd declensions (both '-a' and '-e' declensions) singular definite: katten-katt, stolen-stol. Tag: NCUSN@DS substantiv utrum singularis bestamd nominativ
s/^\([^\t]*\)\(en\)\t\(NCUSN@DS\)$/\1\2\t\1\t\3/;


#Nouns, 4th declension: singular definite: äpplet-äpple hjärtat-hjärta, knät-knäTag: NCNSN@DS substantiv neutrum singularis best. nominativ

s/^\([^\t]*[^nm]\)\(net\)\t\(NCNSN@DS\)$/\1\2\t\1en\t\3/;
s/^\([^\t]*[aåäo]\)\(t\)\t\(NCNSN@DS\)$/\1\2\t\1\t\3/;

#Nouns, 5th declension: plural indefinite: hjärtan-hjärta,...Tag: NCNPN@IS substantiv neutrum pluralis obest. nominativ

s/^\([^\t]*[eaåäouyö]\)\(n\)\t\(NCNPN@IS\)$/\1\2\t\1\t\3/;

#Nouns, 5th declension: plural definite: hjärtana-hjärta, Tag: NCNPN@DS substantiv neutrum pluralis best. nominativ
s/^\([^\t]*[eaåäouyö]\)\(na\)\t\(NCNPN@DS\)$/\1\2\t\1\t\3/;
#lempas2
```

```
#Nouns, 5th declension genitive
s/^\([^\t]*[eaääouyö]\)\(ns\)\t\(NCNPG@IS\)$/\1\2\t\1\t\3/;
s/^\([^\t]*[eaääouyö]\)\(nas\)\t\(NCNPG@DS\)$/\1\2\t\1\t\3/;

#lempas6
#type radion singular definite nominative and genitive
s/^\([^\t]*[iuyöääo]\)\(n\)\t\(NCUSN@DS\)$/\1\2\t\1\t\3/;
s/^\([^\t]*[iuyöääo]\)\(ns\)\t\(NCUSG@DS\)$/\1\2\t\1\t\3/;

#lempas2: inserted aeiouyöää and n to 'nen'
#Nouns, 6th declension, plural definite of words ending with 'len', 'nen, 'ren':
s/^\([^\t]*[^laeiouyöää]\)\(len\)\t\(NCNPN@DS\)$/\1\2\t\1el\t\3/;
s/^\([^\t]*[^raeiouyöää]\)\(ren\)\t\(NCNPN@DS\)$/\1\2\t\1er\t\3/;
s/^\([^\t]*[^rnaeiouyöää]\)\(nen\)\t\(NCNPN@DS\)$/\1\2\t\1en\t\3/;
#Nouns, 6th declension: plural definite of words ending with 'men':
s/^\([^\t]*m\)\(men\)\t\(NCNPN@DS\)$/\1\2\t\1\t\3/;

#General rule - Nouns, 6th declension: plural definite: husen-hus, Tag: NCNPN@DS substantiv neutrum pluralis best. nominativ
s/^\([^\t]*\)\(en\)\t\(NCNPN@DS\)$/\1\2\t\1\t\3/;

#lempas2 Nouns 6th declension singular definite 'mlet'
#-mlet
s/^\([^\t]*\)\(mlet\)\t\(NCNSN@DS\)$/\1\2\t\1mmel\t\3/;

#Nouns, 6th declension: singular definite of words ending with 'let'
s/^\([^\t]*[^laeiouyöää]\)\(let\)\t\(NCNSN@DS\)$/\1\2\t\1el\t\3/;
#vattnet
s/^\([^\t]*[^vatt]\)\(net\)\t\(NCNSN@DS\)$/\1\2\t\1en\t\3/;

#fönstret
s/^\([^\t]*[^raeiouyöää]\)\(ret\)\t\(NCNSN@DS\)$/\1\2\t\1er\t\3/;

#Nouns, 6th declension: singular definite of words ending with 'met':
s/^\([^\t]*m\)\(met\)\t\(NCNSN@DS\)$/\1\2\t\1\t\3/;

#Nouns, 6th declension: singular definite of 'regular' cases:
s/^\([^\t]*\)\(et\)\t\(NCNSN@DS\)$/\1\2\t\1\t\3/;

#Nouns, 6th declension: elimination of 'e-'words: Like the uters, neuters ending with 'e' have their lemma without 'e'.
s/^\([^\t]*\)\(e\)\t\(NCNSN@IS\)$/\1\2\t\1\t\3/;


#the general rule
s/^\([^\t]*\)\(et\)\t\(NCNSN@DS\)$/\1\2\t\1\t\3/;
s/^\([^\t]*\)\(en\)\t\(NCNPN@DS\)$/\1\2\t\1\t\3/;

#LOAN WORDS ENDING WITH S
s/^\([^\t]*\)\([aieuo]s\)\t\(NCNPN@IS\)$/\1\2\t\1\t\3/;

```

Chapter 21

Targeted Corpus Query Templates for PAROLE

```
CQL Queries for PAROLE

VERB-NOUN
[word="hade"|word="ha"|word="har"|word="haft"] []{0,4} [lemma="rekord"]

WITHOUT ADJ OR PP MODIFIERS

BARE NOUN
[word="ha"|word="har"|word="hade"|word="haft"]
[word!="ett"&tag!="A.*"&tag!="N...G.*"&tag!="PS.*" &tag!="D.*"]{0,3}
[word=".*rekord"] [tag!="SPS"]

INDEFINITE ARTICLE
[word="ha"|word="har"|word="hade"|word="haft"] [word="ett"] [word=".*rekord"]
[tag!="SPS"]

DEFINITE ARTICLE
[word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]{0,
1} [word="rekordet"|word="rekorden"] [tag!="SPS"]

DEN HÄR
[word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]
[word="[dh]är"][word=".*rekord"|word="rekordet"|word="rekorden"] [tag!="SPS"]

POSSESSIVE or GENITIVE DETERMINER
[word="ha"|word="har"|word="hade"|word="haft"] [tag="D.*"]{0,1}
[tag="N...G.*"|tag="PS.*"]{1,3} [word=".*rekord"] [tag!="SPS"]

OTHER DETERMINER
[word="ha"|word="har"|word="hade"|word="haft"] [tag="D.*"] [word=".*rekord"]
[tag!="SPS"]

ADJECTIVE MODIFIER ONLY
BARE NOUN
[word="ha"|word="har"|word="hade"|word="haft"]
[word!="ett"&tag!="N...G.*"&tag!="PS.*" &tag!="D.*"]{0,3} [tag="A.*"]{1,3}
[word=".*rekord"] [tag!="SPS"]

INDEFINITE ARTICLE
[word="ha"|word="har"|word="hade"|word="haft"] [word="ett"] [tag="A.*"]{1,3}
[word=".*rekord"] [tag!="SPS"]

DEFINITE ARTICLE
[word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]{0,
1} [tag="A.*"]{1,3} [word=".*rekordet"|word=".*rekorden"] [tag!="SPS"]

DEN HÄR
[word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]
[word="[dh]är"][tag="A.*"]{1,3}[word=".*rekordet"|word=".*rekorden"] [tag!="SPS"]


POSSESSIVE or GENITIVE DETERMINER
[word="ha"|word="har"|word="hade"|word="haft"] [tag="D.*"]{0,1}
[tag="N...G.*"|tag="PS.*"]{1,3} [tag="A.*"]{1,3} [word=".*rekord"] [tag!="SPS"]

OTHER DETERMINER
[word="ha"|word="har"|word="hade"|word="haft"]
[tag="D.*"&word!="en"&word!="ett"&word!="den"&word!="det"&word!="de"] [tag="A.*"]{1,
3} [word=".*rekord"] [tag!="SPS"]
```

```
45:
46:
47: PREPOSITIONAL MODIFIER ONLY
48: BARE NOUN
49: [word="ha"|word="har"|word="hade"|word="haft"]
    [word!="ett"&tag!="A.*"&tag!="N...G.*"&tag!="PS.*" &tag!="D.*"]{0,3}
    [word=".*rekord"] [tag="SPS"]
50:
51: INDEFINITE ARTICLE
52: [word="ha"|word="har"|word="hade"|word="haft"] [word="ett"] [word=".*rekord"]
    [tag="SPS"]
53:
54: DEFINITE ARTICLE
55: [word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]{0,
    1} [word=".*rekordet"|word=".*rekorden"] [tag="SPS"]
56:
57: DEN HÄR
58: [word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]
    [word="[dh]är"][word=".*rekordet"|word=".*rekorden"] [tag="SPS"]
59:
60: POSSESSIVE or GENITIVE DETERMINER
61: [word="ha"|word="har"|word="hade"|word="haft"] [tag="D.*"]{0,1}
    [tag="N...G.*"|tag="PS.*"]{1,3} [word=".*rekord"] [tag="SPS"]
62:
63: OTHER DETERMINER
64: [word="ha"|word="har"|word="hade"|word="haft"]
    [tag="D.*"&word!="en"&word!="ett"&word!="den"&word!="det"&word!="de"]
    [word=".*rekord"] [tag="SPS"]
65:
66: ADJ+PP MODIFIER
67: BARE NOUN
68: [word="ha"|word="har"|word="hade"|word="haft"]
    [word!="ett"&tag!="N...G.*"&tag!="PS.*" &tag!="D.*"]{0,3} [tag="A.*"]{1,3}
    [word=".*rekord"] [tag="SPS"]
69:
70: INDEFINITE ARTICLE
71: [word="ha"|word="har"|word="hade"|word="haft"] [word="ett"] [tag="A.*"]{1,3}
    [word=".*rekord"] [tag="SPS"]
72:
73: DEFINITE ARTICLE
74: [word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]{0,
    1} [tag="A.*"]{1,3} [word=".*rekordet"|word=".*rekorden"] [tag="SPS"]
75:
76: DEN HÄR
77: [word="ha"|word="har"|word="hade"|word="haft"] [word="det"|word="den"|word="de"]
    [word="[dh]är"][tag="A.*"]{1,3}[word="rekordet"|word="rekorden"] [tag="SPS"]
78:
79:
80: POSSESSIVE or GENITIVE DETERMINER
81: [word="ha"|word="har"|word="hade"|word="haft"] [tag="D.*"]{0,1}
    [tag="N...G.*"|tag="PS.*"]{1,3} [tag="A.*"]{1,3} [word=".*rekord"] [tag="SPS"]
82:
83: OTHER DETERMINER
84: [word="ha"|word="har"|word="hade"|word="haft"]
    [tag="D.*"&word!="en"&word!="ett"&word!="den"&word!="det"&word!="de"] [tag="A.*"]{1,
    3} [word=".*rekord"] [tag="SPS"]
85:
86: PREPOSITIONAL_SLOTS_GENERAL
```

```
[word="ha"|word="har"|word="hade"|word="haft"] []{0,4} [lemma=".*rekord"] [tag="SPS"]

PREPOSITIONAL_SLOTS_FILL-IN
[word="ha"|word="har"|word="hade"|word="haft"] []{0,4} [lemma=".*rekord"]
[tag="SPS"&word=""]

LIGHT VERB SUBJECT
[lemma="rekord"] []{0,4}[word="lyda"|word="lyder"|word="löd"|word="ludid"]
```