UNIVERSITÉ DE FRANCHE-COMTÉ
ÉCOLE DOCTORALE « LANGAGES, ESPACES, TEMPS, SOCIÉTÉS »

Thèse en vue de l'obtention du titre de docteur en

SCIENCES DU LANGAGE

COMPUTATIONAL SYNTAX OF HUNGARIAN: FROM PHRASE CHUNKING TO VERB SUBCATEGORIZATION

Présentée et soutenue publiquement le 12 Juin 2012 par

Kata Gábor

Président :	Professeur Valérie Spaeth
Rapporteurs :	Professeur Cécile Fabre
	Professeur Denis Maurel
Examinateurs :	Dr. Ágnes Sándor
	Dr. Tamás Váradi
Directeur de thèse :	Professeur Max Silberztein

Mis en page avec la classe Phdlasl

Acknowledgements

I wish to thank my supervisor, Max Silberztein, for accepting to supervise my work and for his valuable remarks on earlier versions of this thesis. I am grateful to Cécile Fabre and to Denis Maurel for agreeing to evaluate my work, as well as to Ágnes Sándor, Tamás Váradi and Valérie Spaeth for participating in my thesis committee.

I would like to thank my collegues and friends at the Linguistics Institute of the HAS, especially Judit Kuti, Péter Vajda, Csaba Oravecz, Bálint Sass, Viktor Nagy, Eszter Simon and Ágnes Mészáros for the inspiring and friendly atmosphere. I am grateful to Thierry Poibeau and Cédric Messiant for welcoming me at the LIPN research group in 2008, and for their contribution and comments on the classification of French verbs. My work has benefitted from discussions with András Komlósy and from comments of several anonymous reviewers from Hungary. I am indebted to Christine Fay-Varnier and Claire Gardent for inviting me to Nancy and for the comments I received on my work at these occasions. I am also grateful to Anna Korhonen and Olivier Ferret for their valuable questions at ACL 2007.

I cannot thank Enikő Héja enough, not only for our fruitful collaboration on several of the topics in this thesis, but also for her interest in my work and for our numerous enriching discussions at work or over a beer. Her orginal insights, scientific curiosity, ambition and humour have been a great inspiration to me.

On a personal note, I am very grateful to Anna, Eszter, Ági, Dani, Dorka for their friendship and for the encouragement. I wish to thank Jef for his love and patience, especially during the last weeks. Finally, I dedicate this thesis to my mother, Éva.

Contents

Abstract

We present the creation of two resources for Hungarian NLP applications: a rule-based shallow parser and a database of verbal subcategorization frames. Hungarian, as a non-configurational language with a rich morphology, presents specific challenges for NLP at the level of morphological and syntactic processing. While efficient and precise morphological analyzers are already available, Hungarian is less resourced with respect to syntactic analysis. Our work aimed at overcoming this problem by providing resources for syntactic processing. Hungarian language is characterized by a rich morphology and a non-configurational encoding of grammatical functions. These features imply that the syntactic processing of Hungarian has to rely on morphological features rather than on constituent order. The broader interest of our undertaking is to propose representations and methods that are adapted to these specific characteristics, and at the same time in line with state of the art research methodologies. More concretely, we attempt to adapt current results in argument realization and lexical semantics to the task of labeling sentence constituents according to their syntactic function and semantic role in Hungarian. Syntax and semantics are not completely independent modules in linguistic analysis and language processing: it has been known for decades that semantic properties of words affect their syntactic distribution. Within the syntax-semantics interface, the field of argument realization deals with the (partial or complete) prediction of verbal subcategorization from semantic properties. Research on verbal lexical semantics and semantically motivated mapping has been concentrating on predicting the syntactic realization of arguments, taking for granted (either explicitly or implicitly) that the distinction between arguments and adjuncts is known, and that adjuncts' syntactic realization is governed by productive syntactic rules, not lexical properties. However, besides the correlation between verbal aspect or actionsart and time adverbs ((Vendler, 1957, see) or (Kiefer, 1992) on Hungarian), the distribution of adjuncts among verbs or verb classes did not receive significant attention, especially within the lexical semantics framework. We claim that contrary to the widely shared presumption, adjuncts are often not fully productive. We therefore propose a gradual notion of productivity, defined in relation to Levin-type lexical semantic verb classes Levin (1993); Levin and Hovav (2005). The definition we propose for the argument-adjunct dichotomy is based on evidence from Hungarian and

exploits the idea that lexical semantics not only influences complement structure but is the key to the argument-adjunct distinction and the realization of adjuncts.

We argue that current definitions of the complement-adjunct dichotomy present contradictions, due to the fact that the semantic definitions and the corresponding syntactic tests do not designate the same set of constituents as complements. We also show by a counter-example that adjuncts, similarly to complements, can provoke event type shift at the semantic level, and hence change the syntactic distribution of the verb phrase they are attached to.

We target a lexical representation which makes use of semantically motivated and syntactically relevant features, while meeting the necessary criteria towards computational databases for natural language processing. This is a particularly challenging task, since neither configuration-based complement definitions, nor language-specific ones cannot be used satisfactorily for an extensive account on Hungarian complement structures. The methodology we propose is an adaption of Levin's verb classification (1993) to a non-configurational language where syntactic complement tests systematically fail.

The first part of the research concentrated on creating a shallow parser/phrase chunker for Hungarian. Syntactic parsing is the task of recognizing an input sentence and assigning a syntactic structure to it. The targeted annotation structure we assign to sentences is flat. Top-level sentence constituents are recognized and annotated according to their category. The output of the shallow parser is conceived as an input for potential further processing, aiming to annotate dependency relations between the predicate and its complements and adjuncts. The two basic features that determined our choice of the language processing framework were robustness, i.e. the flexibility to deal with unrecognized sequences in the input sentence, and expressive power, i.e. a formalism that allows linguistically motivated and transparent representation. The shallow parser is implemented in NooJ (Silberztein, 2004) as a cascade of grammars. The shallow parser was evaluated on an extract from a manually annotated Hungarian treebank. We also describe the construction of an automatically annotated 10.000.000 words corpus using this shallow parser.

The first step was to define the coding guidelines and to construct an extensive database of verbal complement frames. Subsequently, we undertook two experiments

on enhancing this lexicon with lexical semantic information in order to capture syntactically and semantically relevant generalizations about predicate classes. The first approach we tested consisted in a manual definition of verb classes based on the distribution of adjuncts and on semantic role assignment. We soughed the answer to the questions: which semantic components underlie natural classes of predicates in Hungarian? What are the syntactic implications specific to these classes? And, more generally, what is the nature of class-specific alternations in Hungarian? In the final stage of the research, we investigated the potential of automatic lexical acquisition to induce semantic verb classes from corpora. We performed an unsupervised clustering based on distributional data to obtain a semantically relevant classification of Hungarian verbs. We also tested the clustering method on French. The goal of the experiment was to confirm the hypothesis that semantic similarity underlies distributional similarity, and that full, generalized complementation patterns can be used efficiently to model the syntactic contexts of verbs. The major stake of this experiment was to support manual predicate classification and to facilitate the exploration of syntactically relevant facets of verbal meaning in Hungarian.

Keywords: syntax, syntax-semantics interface, verb classes, subcategorization, lexical semantics

Chapter I

Introduction

I.1 Context

Natural language technology deals with the automatic processing of natural languages from an application-oriented perspective. Computational linguistics on the other hand is a research field focusing on the methods and possibilities of formal - symbolic or statistical - modeling of natural languages. Computational linguistics, similarly to theoretical linguistics, is a highly modularized discipline: the levels of analysis include segmentation, morphological analysis, disambiguation, syntactic and semantic processing. Language technology applications differ with respect to the level of linguistic processing they require. Systems aiming at intelligent text processing (such as e.g. information or knowledge extraction, machine translation, question answering) deal with complex tasks requiring a certain understanding of natural language texts. These applications rely on a high level of language analysis, especially on syntax and semantics.

Whereas robust and accurate tools already exist for the lower levels of analysis of Hungarian, it can be considered as a relatively less resourced language with respect to syntactic and semantic processing. The work reported in present thesis aims at providing resources for the syntactic processing of Hungarian in the form of a phrase chunker and an extensive lexical database of verbal subcategorization frames.

Hungarian language is characterized by a rich morphology and a non-configurational encoding of grammatical functions. These features imply that the syntactic processing of Hungarian has to rely on morphological features rather than on constituent order. The

broader interest of our undertaking is to propose representations and methods that are adapted to these specific characteristics, and at the same time are in line with state of the art research problems and methodologies. More concretely, we attempt to adapt current results in argument realization and lexical semantics to the task of labeling sentence constituents according to their syntactic function and semantic role in Hungarian.

Syntax and semantics are not completely independent modules in linguistic analysis and language processing: it has been known for decades that semantic properties of words affect their syntactic distribution. Within the syntax-semantics interface, the field of argument realization deals with the (partial or complete) prediction of verbal subcategorization from semantic properties. In the present thesis, we target a lexical representation which makes use of semantically motivated and syntactically relevant features, while meeting the necessary criteria towards computational databases for natural language processing. In order to accomplish this objective, we need a method to separate lexical properties of words from phenomena pertaining to syntax. We also need definitions and tests to differentiate between complements and adjuncts, presuming that complement structure has to be lexically coded, while adjuncts are productive and, therefore, predictable. This is a particularly challenging task, since neither configuration-based complement definitions nor language-specific ones can be used satisfactorily for an extensive account on Hungarian complement structures.

I.2 Research Aims

The first part of the research presented here concentrated on creating a shallow parser/phrase chunker for Hungarian. Syntactic parsing is the task of recognizing an input sentence and assigning a syntactic structure to it. The targeted annotation structure we assign to sentences is flat. Top-level sentence constituents are recognized and annotated according to their category. The output of the shallow parser is concieved as an input for potential further processing, aiming to annotate dependency relations between the predicate and its complements and adjuncts. The two basic features that determined our choice of the language processing framework were robustness, i.e. the flexibility to deal with unrecognized sequences in the input sentence, and expressive power,

i.e. a formalism that allows linguistically motivated and transparent representation. The shallow parser is implemented in NooJ (Silberztein, 2004) as a cascade of grammars.

The second research objective was to propose a lexical representation for complement structure in Hungarian. The representation had to deal with a wide range of phenomena that misfit the traditional complement-adjunct dichotomy (partially productive structures, shifts in the parallelism between syntactic predictability and semantic transparency). It also had to provide a basis for distinguishing lexically encoded, idiosyncratic phenomena, to be included in a verbal subcategorization database, from adjuncts, to be processed by syntactic grammars. We resorted to results from recent research on argument realization and chose a framework which meets our criteria and is adaptable to a non-configurational language. We used Levin's semantic classification (Levin, 1993) as a model. Applying the basic notions pertaining to this predicate classification, i.e. semantic meaning components and diathesis alternations, as well as the methodology to explore and describe the behavior of predicates using this representation, had to be revised in the light of data from a non-configurational language.

After the theoretical reflections about the subcategorization properties of Hungarian verbs, the next research aim was to implement the model to create lexical resources for Hungarian verbs. The first step was to define the coding guidelines and to construct an extensive database of verbal complement frames. Subsequently, we undertook two experiments on enhancing this lexicon with lexical semantic information in order to capture syntactically and semantically relevant generalizations about predicate classes. The first approach we tested consisted in a manual definition of verb classes based on the distribution of adjuncts and on semantic role assignment. We sought the answer to the questions: which semantic components underlie natural classes of predicates in Hungarian? What are the syntactic implications specific to these classes? And, more generally, what is the nature of class-specific alternations in Hungarian? In the final stage of the research, we investigated the potential of automatic lexical acquisition to induce semantic verb classes from corpora. We performed an unsupervised clustering based on distributional data to obtain a semantically relevant classification of Hungarian verbs. We also tested the clustering method on French. The goal of the experiment was to confirm the hypothesis that semantic similarity underlies distributional similarity, and that full, generalized com-

plementation patterns can be used efficiently to model the syntactic contexts of verbs. The major stake of this experiment was to support manual predicate classification and to facilitate the exploration of syntactically relevant facets of verbal meaning in Hungarian.

I.3 Thesis Plan

Chapter II presents the work related to the thematics of shallow parsing. In II/1, we precise the scope of the work, the problems to be addressed, and describe the state of the art in symbolic and statistical shallow parsing and phrase chunking. We subsequently discuss the available language processing tools and methodologies, and motivate our choice of the language processing software used to implement our phrase chunker. In II/2, we demonstrate the structure of our shallow parser and describe the cascaded grammars in details. The evaluation of our shallow parser for the task of NP chunking is related at the end of II/2. In section II/3, we present a shallow parsed corpus, annotated automatically with NooJ, using our grammars.

Chapter III deals with the theoretical issues of representing verbal subcategorization in a lexicon. In III/1, we present current research on the syntax-semantic interface with respect to verbal argument structures, and explain why Hungarian as a non-configurational language constitutes a challenge for argument realization theories. We evoke the definition of the Semantic Basis Hypothesis and sum up the major presumptions shared across argument realization studies. In III/2 , we examine argument definitions and tests. We show how configuration-based argument tests systematically fail for Hungarian. We further demonstrate that adjuncts can also induce event type shift and change the complementation pattern of a predicate they are added to. Finally, we argue that there is an internal incoherence in argument definitions building on a supposed parallelism between semantic specificity and syntactic complement status, which is also reflected in current challenges for research on thematic role assignment. In III/3, we propose to build a lexical representation based on the model of Levin. We suggest that the model can be adapted to Hungarian by putting in the center semantic compositionality on the semantic level, and case suffixes on the morphosyntactic level.

Chapters IV–VI describe the implications of our model when put into practice. In IV,

we present the construction of a verbal subcategorization lexicon in detail. The coding guidelines, the main characteristics of the database, as well as the problems encountered are discussed. After presenting the creation of the lexicon, we report on two experiments aiming to enhance it by lexical semantic features. Chapter V deals with the manual exploration of semantic verb classes and the corresponding alternations. The methodology is based on a systematic study of case suffixes and the semantic roles they encode in the context of predicates belonging to different semantic verb classes. A case study is presented, aiming to exhaustively describe the semantic roles encoded by the instrumental case suffix in Hungarian, and the resulting verb classification. In chapter VI, we present an unsupervised learning experiment aiming to automatize the classification of predicates based on syntactic contexts extracted from corpora. In VII/1, we present the different NLP applications our resources were integrated to. In VII/2, we sum up the work covered in the thesis and conclude the important contributions. Finally, in VII/3, we designate directions for further research.

Chapter II

Formal Description of Sentence Constituents. Shallow Parsing.

II.1 Building a Shallow Parser for Hungarian

This chapter describes the construction of a phrase chunker and shallow parser for Hungarian, implemented as a series of cascaded grammars in the corpus processing software NooJ. The basic components are local grammars which perform a phrase chunking function, i.e. recognize and bracket constituents (such as basic NPs, adjectival phrases and postpositional phrases), and produce a flat structure. The core of the chunker is the two-level NP grammar: the basic grammar is non-recursive while the extended one is recursive. Beside phrase chunking, the shallow parser is enhanced with additional grammars for clause boundary detection and for an approximation of the basic dependency structure between constituents within the same clause.

The shallow parser was designated to provide any piece of information about the sentence and its top-level constituents necessary for the recognition of argument structure. It is integrated into the Hungarian NooJ module, which consists of a series of grammars and lexical resources supporting syntactic and/or semantic annotation of Hungarian texts. The annotation can be performed entirely in NooJ, and the results can be exported in XML. The secondary goal was to provide a data-driven formal description of sentence constituents (especially noun phrases) in Hungarian, which can also be re-implemented in other corpus processing tools or NLP applications.

Section II/1 defines the goals of this work and its context. Section II/2 describes the structure of the shallow parser. The last part of this chapter, II/3 presents a 10 million words Hungarian corpus, syntactically annotated by our shallow parser. This is currently the biggest syntactically annotated Hungarian corpus, and the first one which is analyzed fully automatically.

II.1.1 Parsing: Types of Formal Grammars.

II.1.1.1 Definitions

When annotating corpora automatically with syntactic information, the two basic decisions to be taken are 1) the choice of the theoretical framework, i. e. the language model, 2) the choice of the language processing tool which is compatible with the language model to be used. While 1) is a question of theoretical interest, 2) is largely influenced by the availability of tools. From an engineering viewpoint, we can argue that the more restricted our formalism is, the more efficient implementation we can use, which is undoubtedly an important aspect in natural language processing. Moreover, one could say that when linguists argue in favor of a more powerful formalism, the examples they cite are usually artificial: sentences that occur mostly in linguistics papers or textbooks, but rarely in spontaneous utterances. However, whereas computational linguists focus on the generative capacity of their grammar, (theoretical) linguists want their language models to have an explanatory power, which requires more elegant modeling of linguistic phenomena. After a description of formal language models and current parsing techniques, we will present some of the available language processing software. Subsequently, we will explain the motivation behind our choice of linguistic modeling framework. After that, we will present NooJ, the software that we have selected for implementing the shallow parser.

Syntactic parsing is the task of recognizing an input sentence and assigning a syntactic structure to it (Jurafsky and Martin, 2000). Either deep or partial parsing is needed for a wide range of NLP applications (information extraction, machine translation, question answering). Parse trees typically serve as an important intermediate stage of representation for a deeper analysis. Parsing is done by an algorithm according to a set of rules which define a formal grammar, and typically includes a disambiguation module to choose

between possible parse trees at the end of the analysis. Formal grammars can be represented by rewriting rules which manipulate terminal and nonterminal symbols. Grammars can be classified according to the complexity of their structure. The Chomsky hierarchy (Chomsky, 1956) specifies a hierarchy between formal grammars. Classes of grammars in this hierarchy are defined by the formal constraints that each rewriting rule has to satisfy. Besides theoretical interest, such a classification is useful because it allows to find a more suitable algorithmic implementation of the grammar. In what follows, we will sum up the important characteristics of formal grammars, based on (Jurafsky and Martin, 2000) and (Alberti, 2006).

The most restricted formalism in the Chomsky hierarchy corresponds to regular grammars. Regular grammars generate regular languages. Their rewriting rules take the form

$$A \rightarrow xB$$

$$A \rightarrow x$$

There is a single nonterminal on the left-hand side of the rule, while on the right-hand side we have a single terminal, possibly followed (in right-linear grammars) or preceded (in left-linear grammars) by a single nonterminal. The languages generated by regular grammars cover the same set as languages that can be decided by a finite state automaton (FSA) or described by regular expressions.

Context-free grammars generate context free languages. They are defined by rules of the form

$$A \rightarrow \gamma$$

where A is a nonterminal and γ is a string of terminals and nonterminals. The computational tool which corresponds to CF languages is a pushdown automaton (also known as stack automaton). A pushdown automaton is in fact a finite state automaton enhanced with a last in, first out (LIFO) memory: at each moment only the last data stored in the memory is accessible for reading or deleting. The early phrase structure grammars were based on context free grammars, e.g. the early chomskyan grammars (Chomsky, 1957) before the introduction of transformations; as well as Generalized Phrase Structure Grammar (Gazdar et al., 1985), Lexical Functional Grammar (Kaplan and Bresnan, 1982)

or Head-Driven Phrase Structure Grammar (Pollard and Sag, 1994). Current techniques for full parsing typically use a CF formalism.

The next level in the hierarchy is context sensitive grammars. They generate context sensitive languages. Context-sensitive rewriting rules take the following form:

$$\alpha A \beta \rightarrow \alpha \gamma \beta$$

where A is a nonterminal and α, β and γ are strings of terminals and nonterminals. The strings α and β may be empty, but γ cannot. Context-sensitive languages are languages that can be recognized by a linear bounded automaton.

Finally, Type 0 grammars are unrestricted: they include all formal grammars that can be generated by a Turing machine. The only restriction on the form of the rules is that the left hand side cannot be empty. These languages are also known as recursively enumerable languages.

II.1.1.2 Formal Grammars and Natural Languages

It is of both theoretical and practical interest to know what language group natural languages belong to and why; in other words, which linguistic phenomena can be modeled with which kind of grammar. The more restricted the formalism behind the grammar is, the more efficient (faster) the processing tool can be. Finite state methods are thus particularly appealing for the processing speed they offer. They are typically used in the domain of phonology and morphology, but implementations exist for other purposes, including shallow parsing. On the other hand, the expressive power of the grammar has also be taken into account, and - as we will see - this aspect favors the use of more powerful languages.

As to the mathematical aspect of the question, Partee et al. (1990) formally proved that English is not a regular language. The language data this proof is based on are sentences with a potentially infinite number of recursive center-embedded phrases. However, the assumption that center-embedding can go on indefinitely is widely challenged because of limitations on the length of natural language sentences as well as on the processing capacity of human brain (Karlsson, 2007). From a more pragmatical viewpoint it is acceptable to put a practical limitation on the number of recursion, since natural

language sentences with a center-embedding structure with several recursions will be extremely rare in corpora. Another argument against FSAs in language modeling, though not of a mathematical nature, concerns formal description of long-distance dependencies, subcategorization and agreement phenomena. Many of these phenomena in English can be modeled by finite state methods, as shown among others by Pullum and Gazdar (1982) for number agreement in English questions. Nevertheless, even if no mathematical factors prevent us from constructing FSAs for such phenomena, we observe that their expressive power is often too low to provide linguistically understandable formalisms. Cascaded grammars (Abney, 1996) offer a solution to this problem, meanwhile staying within the finite state framework. An inconvenient of his finite state cascades is, however, that they operate in a way that earlier steps of the annotation do not remain accessible during the upper levels of processing.

We can conclude that there is a set of natural language phenomena for which FSAs cannot provide an expressive model. Naturally, this does not mean that we have to write them off definitively, as they can be efficient in modeling a big set of phenomena. As Jurafsky and Martin (2000) state, "there is a considerable price to be paid in terms of processing speed when one switches from regular languages to context-free ones". Thus, computational linguists continue to use FSAs in several domains: morphology (Koskenniemi, 1984; Beesley and Karttunen, 2000; Silberztein, 1993, 1997), POS tagging (Silberztein, 1993, 2007), as well as in syntactic analysis, especially in partial parsing: they are particularly useful for chunking continuous constituents (Abney, 1996; Roche and Schabes, 1997). It is also possible to automatically build a regular grammar which is an approximation of a given context-free grammar. Pereira and Wright (1997) and Evans (1997) propose two different algorithms to approximate context free grammars by finite state calculus (both of them intended to be used in speech processing.)

We have seen that natural languages are not regular; the next question is whether they are context free. A context-free language is the set of strings which can be derived from a particular context-free grammar. Many linguists gave a negative answer to this question, even before the correct mathematical proof has been published by Shieber (1985). Chomsky (1957) claims that context free grammars easily become desperately complicated and lose their linguistic expressiveness unless we introduce transformations into the

formalism. Bresnan (1978) suggests that even context-sensitive phrase structure rules are inadequate to describe long-distance dependencies. Several arguments based on linguistic observations have been published against context-freeness of natural languages, mostly with the intention to prove that transformations are needed in generative grammar. This is in accordance with linguistsńeed for a more powerful and thus comfortable language description formalism, but conflicts mathematicianséffort to define constraints on natural languages (Alberti, 2006) as well as engineers who try to construct computationally efficient tools for language analysis. Pullum and Gazdar (1982), without claiming that natural languages are context free, systematically enumerate and disprove the previously published arguments against it. The decisive argument comes from Shieber (1985): he used evidence from Swiss German to show that at least some natural languages were not context free. This leads back to the question in Pullum and Gazdar (1982): "If a human language that is not a CFL is proved to exist, (...) a different question will be raised: given the non-context free character of human languages in general, why has this property been so hard to demonstrate that it has taken over twenty-five years to bring it to light since the issue was first explicitly posed?" (Kornai, habilitation thesis) explains that the "dubious grammatical status", as well as limited iteration possibilities and extremely limited frequencies make such structures marginal in natural language grammar. The status of corpus frequencies, and in particular whether we can infer the non-existence of linguistic structures from their absence in corpora, is evidently an open question (Saul and Pereira, 1997; Stefanowitsch, 2006). However, from an engineering point of view it is defensible to narrow down the scope of investigation to linguistic structures which actually occur in corpora (and not only in linguistics textbooks). Kornai (2008) concludes that it is unacceptable to waste the resources to marginal linguistic phenomena with frequencies far below the error rate of current state of the art parsers.

However, rare and marginal examples of non CF linguistic phenomena put aside, linguists may resort to more powerful formalisms in order to provide linguistically correct models for long-distance agreement. Since indexing or feature unification is still not allowed in CFGs, these phenomena cannot be described in a linguistically transparent way. They do not allow meaningful generalizations and often yield descriptions which seem to be ad hoc. For these considerations, in some cases it will be difficult to reuse

CF grammars for semantic interpretation. Moreover, this can lead to an explosion in the number of rules. Let us consider the example of agreement and recapitulate how it can be dealt with by finite state and context free formalisms. This problem might be less obvious when parsing English texts because agreement is by and large limited to number and person (2*3 possibilities). On the other hand, languages with a rich morphology and thus broader possibilities for agreement present a real challenge for FSAs. For instance, there are 20 nominal cases in Hungarian (according to the annotation of the Hungarian National Corpus (Váradi, 2002)). Two NPs can be coordinated with a conjunction between them, which would correspond to the following rewriting rules.[1]

- Finite state grammar for NP coordination:
  NP → DET N'
  N' → ADJ N
  N' → N
  COORDP → CONJ NP
  NP → NP COORDP

- Context free grammar for NP coordination:
  NP → DET N
  NP → DET ADJ N
  NP → NP CONJ NP

We can conclude from these examples that CF grammars do not need to resort to auxiliary nonterminal symbols such as N' and COORDP, thus providing a more transparent description. However, the real difficulty is in the modeling of agreement between the coordinated NPs and, as we will see, FS and CF grammars present the same weakness regarding this phenomenon. Coordination can take place if and only if both of the head nouns are inflected with the same case suffix, which would lead to the following rule systems:

[1]For simplicity's sake, we only consider NPs with the internal structure limited to a noun, preceded by a determiner and an optional adjective. Terminals are not taken into account.

- Finite state grammar for NP coordination revised:

  for each x in CASES:

  $NP_x \rightarrow$ DET N_x

  $N_{x'} \rightarrow$ ADJ N_x

  $N_{x'} \rightarrow N_x$

  $NP_x \rightarrow NP_x$ CONJ NP_x

- Context free grammar for NP coordination revised:

  for each x in CASES:

  $NP_x \rightarrow$DET N_x

  $NP_x \rightarrow$DET ADJ N_x

  $NP_x \rightarrow NP_x$ CONJ NP_x

Since indexing and feature unification are not allowed neither in FS nor in CF formalisms, our only choice is to enumerate all the twenty cases as independent possibilities. This is not a linguistically meaningful procedure since it does not capture the observation that we are dealing with an agreement. The following figure illustrates a graph which corresponds to a finite state grammar to recognize coordination of basic NPs. Its structure corresponds to the rewriting rule NP $\rightarrow$ NPcase_x + NPcase_x for every case. Although the grammar is finite state, the CF description would not be different in the number of paths.

The intuitively correct formalization of NP coordination in Hungarian by feature unification would be of the structure:

NP $\rightarrow$$\text{NPcase}_i$ CONJ NPcase_j ,

where $i = j$.

GPSG and its derivates (HPSG or LFG) offer a solution by means of feature structures which define a set of meta-rules to make the creation of grammar rules easier and linguistically motivated.

Another important issue for context free models of natural languages, and especially non-configurational languages as Hungarian, is the phenomenon of scrambling. The respective order of sentence constituents in Hungarian, though not completely arbitrary, is not influenced by the grammatical functions of constituents. In shallow parsing as well as in dependency parsing, where delimitation of constituents and exploration of dependency relations between them are at stake, we need to apply a formalism capable to deal with any possible constituent order. Moreover, we prefer to do so in a linguistically transparent way and without an explosion in the number of grammar rules describing the different surface realizations of the same structure. This is clearly not the case of context free rules. Let us consider the following example of a simple Hungarian sentence, comprising a verbal predicate with two arguments in different orders:

1. A fiúk megetették a galambokat.
   the boys+NOM fed the pigeons+ACC
2. Megetették a galambokat a fiúk.
   fed the pigeons+ACC the boys+NOM
3. A galambokat a fiúk megetették.
   the pigeons+ACC the boys+NOM fed
4. Megetették a fiúk a galambokat.
   fed the boys+NOM the pigeons+ACC

A CF grammar to recognize verbal argument structure in this sentence would comprise the following rules:

$\text{VP} \rightarrow \text{NP}_{nom}\ \ \text{V}\ \ \text{NP}_{acc}$

$\text{VP} \rightarrow \text{V}\ \ \text{NP}_{acc}\ \ \text{NP}_{nom}$

$\text{VP} \rightarrow \text{NP}_{acc}\ \ \text{NP}_{nom}\ \text{V}$

$\text{VP} \rightarrow \text{V}\ \ \text{NP}_{nom}\ \ \text{NP}_{acc}$

$\text{V} \rightarrow \text{megetették}$

$NP_{nom} \rightarrow$ a fiúk

$NP_{acc} \rightarrow$ a galambokat

The description is further complicated by the fact that a several number of adjuncts can modify the verbal predicate. The grammar has to take into account that adjuncts can be positioned between the verb and its argument or between two arguments:

1. Megetették tegnap a galambokat a fiúk a kertben.
   fed yesterday the pigeons+ACC the boys+NOM the garden+INE

2. A galambokat a kertben a fiúk tegnap megetették.
   the pigeons+ACC the garden+INE the boys+NOM yesterday fed

This increases the number of rules in our grammar considerably, and makes it more difficult to maintain and refine in light of additional language data. A more powerful grammar in turn can deal with such structures by means of e.g. storing recognized elements in variables and using feature checking to model lexical relations such as verbal subcategorization. These enhancements thus favor the construction of linguistically correct descriptions, albeit at the expense of processing speed.

As to the mathematical side of the problem, as Pullum and Gazdar (1982) suggested, "the obvious thing to do if natural languages were ever shown to be not CFLs (...) would be to start exploring such minimal enhancements of expressive power to determine exactly what natural languages call for in this regard". This effort lead to the definition of mildly context sensitive (MCS) languages (Joshi and Vijay-Shanker, 1985). MCSLs are languages generated by MCS grammars, while MCS formalism is as restricted as possible in its formal power when compared to the unrestricted grammars which are equivalent to Turing machines. Joshi and Vijay-Shanker (1985) defined Tree-Adjoining Grammar (TAG), a mildly context sensitive tree generating system, extended to deal with long distance dependencies by the operation of composition called *adjoining*. Joshi and Vijay-Shanker (1985) proposed a set of criteria for a grammar to be considered mildly context sensitive, including: 1) CF languages are properly contained in MCS languages; 2) languages in MCSL can be parsed in polynomial time; 3) MCSGs capture only certain kinds of dependencies, e.g., nested dependencies and certain limited kinds of crossing dependencies.

Various mildly context sensitive grammar formalisms have been proposed besides TAGs, including but not limited to Head Grammars (Pollard, 1984) or Combinatory Categorial Grammars (Steedman, 1996), all of them being weakly equivalent, as demonstrated by Joshi et al. (1991). Parsers that implement TAGs include the XTag system (Doran et al., 1994) or SemTag (Gardent and Parmentier, 2007).

To conclude, we can say that ever since the Chomsky hierarchy has been defined, considerable effort has been invested into finding the more restricted grammar formalism which is still able to generate every possible sentence of any natural language. Mathematical linguists are interested in answering the question which language group natural languages belong to; language engineers want to work with formalisms which allow fast and robust implementation of rule systems; linguists are looking for formalisms with a sufficient expressive power for building language models which satisfy both descriptive adequacy and explanatory adequacy.

II.1.2 Parsing: State of the Art.

Context-free grammars (or phrase structure grammars) are the most commonly used formalism for full parsing. Most of the generative linguistic theories were based on context-free grammars (such as Head-Driven Phrase Structure Grammar, Lexical-Functional Grammar, Government and Binding (Chomsky, 1981), and Construction Grammar(Kay and Fillmore, 1999). A context-free grammar consists of a set of rules expressed over a set of non-terminal and terminal symbols. An example of top-down parsing algorithm is the Earley parser (Earley, 1983), which can parse all context-free languages. An important difficulty that parsers have to face is structural ambiguity: grammars often assign more than one parse tree to an input. The most widely discussed case of systematic syntactic ambiguity is probably the problem of PP attachment. Many natural language sentences can have more than one possible parses, and NLP systems need to be able to choose the correct one(s). Several ambiguities pass unnoticed by humans because of their semantic or pragmatic inconsistency. However, this underlines the fact that semantic analysis is often unavoidable in order to perform a correct disambiguation. A possible solution in symbolic parsing is to manually refine the grammars until all the parse trees yielded by the system are correct (or if the analysis is still ambiguous, it can be considered as inherent

structural ambiguity). This operation often necessitates to consider semantic information. Manual correction of grammars, however, is a very costly operation, and unambiguous analysis seems to be a very ambitious target. An alternative way to deal with structural ambiguity is by enhancing our grammar by probabilities of the parse trees they produce. These probabilities can be learnt from a treebank. The resulting grammar, called Probabilistic Context-Free Grammar (PCFG) or Stochastic Context-Free Grammar (SCFG), was proposed by Booth (1969). Disambiguation of a sentence S can be thus performed by PCFGs: the task is to find the most likely derivation tree, given S. PCFGs have two important weaknesses: a) (false) independence assumptions: each grammar rule is given one probability value independently of the position it occupies in the tree, and b) lack of lexicalization, i.e. probabilities are attributed to structures of nonterminals, not taking wordsĺexical properties into account.

II.1.2.1 Rule-Based Shallow Parsing

Partial or shallow parsing does not aim to build complete parse trees for input sentences. Instead, the goal is to identify sentence constituents, and eventually to recognize some instances of dependency relations in the sentence. The basic task is chunking, which consists in 1) delimiting non-overlapping sentence constituents, typically NPs and potentially other phrases (with a content word as their head) 2) assigning the correct label to these phrases. Chunking does not imply recognizing hierarchical relations between constituents, its output corresponds to a bracketing/labeling of the input text. Chunking can be done by rules written by linguists, implemented as finite state automata. Such rules are usually applied from left to right, finding the longest match in the sentence. Another path in partial finding is to use cascaded FSTs "to produce representations that closely approximate" complete parse trees (Abney, 1996; Pereira and Wright, 1997). These approximations typically produce flatter trees compared to full parsing by CFGs. "This flatness arises from the fact that such approaches generally defer decisions that may require semantic or contextual factors, such as prepositional phrase attachments, coordination ambiguities, and nominal compound analyses. Nevertheless, the intent is to produce parse trees that link all the major constituents in an input." (Manning and Schuetze, 1999). A major benefit of the finite state approach is the ability to use the

output of earlier transducers as inputs to subsequent transducers to form cascades. In partial parsing, this technique can be used to more closely approximate the output of true context-free parsers. In this approach, an initial set of transducers is used to recognize a subset of syntactic base-phrases. These base-phrases are then passed as input to further transducers that detect larger and larger constituents such as prepositional phrases, verb phrases, clauses, and sentences.

Koskenniemi (1990) presents a method for finite state shallow parsing where input sentences are represented as networks already containing all the possible roles and representations of their units. Syntactic rules correspond to constraints which exclude ungrammatical readings of the sentence, thus, they basically perform a disambiguation function. Voutilainen (1997) describes a similar parsing method, but his grammar specification is obtained from a semi-automatically annotated corpus of example sentences. He uses the so-called dependency-oriented functional tagset: besides POS tags, each input word is furnished with a tag indicating a surface syntactic function. This representation aims to be at the same time sufficiently expressive for stating grammatical generalizations, but also sufficiently underspecified to make for a structurally resolvable grammatical representation.

Pereira and Wright (1997) present an algorithm to approximate CF grammars by FS grammars. The motivation behind their work is that although phrase-structure grammars provide an effective representation for important syntactic and semantic aspects of natural languages, they are computationally too demanding for use as language models. The algorithm has been used to construct finite-state language models for limited-domain speech recognition tasks. However, they note that even when the result FSA accepts the same language, the original CF grammar is still necessary because interpretation algorithms are generally expressed in terms of phrase structures.

By introducing the notion of lexicon-grammar, Gross (1975, 1997) aims to justify the use of finite state methods from a generative linguistics point of view. He highlights that ever since the efforts of Chomsky (1956), amongst others, to demonstrate that a FS formalism was not sufficient to describe human languages, attention has shifted to the study of a set of carefully selected examples of complex long-distance dependencies. This lead to neglecting of the fact that the "description of the linear order of words or

grammatical categories with rather simple dependencies holding between them" was far from being resolved: he states that "detailed attempts of systematic applications have revealed an endless number of subclasses of exceptions, each of them require a special treatment" (Gross, 1997). He directs his research towards the study of local syntactic phenomena, assuming that "the global nature of language results from the interaction of a multiplicity of local finite state schemes". Accordingly, he sets the goal of constructing extensive grammars in the generative linguistic terminology: grammars that are capable of generating (or parsing) every sentence in a corpus of the given language. In this sense, his work does not fall within the scope of shallow parsing; however, his envisages starting linguistic description by the strictly local syntactic phenomena and proceeding step by step towards a complete description of sentence structure.

Lexicon-grammars as defined by Gross are highly lexicalized: they comply with the linguistic approach according to which phrase structure rules that operate on grammatical categories often prove to be inadequate when confronted to real corpus data. The reason for this is that lexical items belonging to the same category frequently show different grammatical behavior, i.e. different distribution. Moreover, syntactic rules are far from being general among members of categories: they more frequently operate either on specific lexical items or on subclasses of parts of speech. Hence, as Mason (2004) notes, an efficient grammar should rather include rules with different degrees of generalization, but most preferably we should have different rules for each lexical item. He is in accordance with Gross (1975) who observes that members of the same grammatical categories show a surprising variety with respect to their distribution. He concludes that categories should be established upon real distributional groupings of words, as found in corpora.

II.1.2.2 Chunking as Classification

As we have seen above, one method to do shallow parsing is by the application of syntactic rules written by linguists or defined on the basis of annotated example sentences. An alternative is to see chunking as a supervised classification task. Annotated training data can be extracted from a treebank or, less typically, they can be produced manually for the task. Supervised machine learning algorithms can then be applied to learn the classification from the data. Ramshaw and Marcus (1995) introduced this approach by

defining the so-called IOB tagset which allows to treat chunking similarly to POS tagging, i.e. as a classification problem (where B stands for the beginning element of the chunk, I stands for internal and O for external elements). The advantage of this scheme is that the size of the tagset is reduced: (2n +1) where n is the number of different chunk categories. Annotation of existing treebanks can be converted to such a scheme in order to provide training data. A representation of the training data can be built by extracting additional features from the treebank, e.g. POS tags, word forms, chunk tags from surrounding tokens within the context window of n tokens. A variety of matching learning methods have been used in constituent chunking (see the CoNLL 2000 shared task on chunking (Tjong Kim Sang and Buchholz, 2000)). Methods using a Hidden Markov Model (Molina and Pla, 2002) or a Maximum Entropy Model ((Koeling, 2000; McCallum et al., 2000) necessitate a relatively reduced feature set and careful feature selection in order to achieve high accuracy. Support Vector Machines ((Kudo and Matsumoto, 2001)) in turn have been proven to be efficient with high dimensional feature spaces; hence, taking lexical information into account or working with larger context windows pays back in this mode. Applying Conditional Random Fields (Lafferty et al., 2001) to shallow parsing was proposed by Sha and Pereira (2003), and has became a largely used method ever since.

II.1.3 Choice of the Computational Tool.

II.1.3.1 Expressive Power, Computational Efficiency

This section presents the available corpus annotation tools and explains the motivation behind choosing NooJ (Silberztein, 2004, 2005) as an environment for implementing a shallow parser for Hungarian. As a summary of the previous sections, we can conclude that two basic features determine the choice of a language processing framework and a corresponding tool: robustness and expressive power. First, the tool has to be robust enough to be able to annotate large amounts of texts in real time. From the same viewpoint it is advantageous if it can process different input text formats. Additionally, since we intend to use our grammars in NLP applications (as will be presented in detail in section 2.3), we also need the system to be robust in the sense that potentially unrecognized sequences cannot destroy the whole analysis. Probabilistic shallow parsers can deal with this problem since they incorporate algorithms to predict the most likely analysis even

for sequences which do not have an exact analogue in the training data. Nevertheless, this often leads to false analyses which are difficult to correct, since increasing the size of training data is not always possible. Rule-based systems on the other hand often fail when the input that cannot be parsed according to the grammar. It is crucial in such cases that the system be able to defer decisions about the grammatical status of such unrecognized sequences to a higher level of analysis, where more information is accessible. The second feature is expressive power. As discussed in section 2.1.1, computational aspects and linguistic expressiveness of formal grammars are in conflict: the more powerful the linguistic formalism is, the less robust implementation it allows. However, linguistic expressiveness largely contributes to the usability of the grammar, especially for rule-based systems. The construction or amelioration of a rule-based parser, as well as its adaptation to new tasks or domains, constantly require humans to interpret the rules. Therefore, it is desirable for the rule system to be concise, coherent and well-structured (i.e. structured according to linguistic description levels and units). For instance, troublesome issues like the multiplication of rules for the sake of computational efficiency - which was exemplified by the formalisation of agreement in a FS or CF framework in section 2.1.1 - are to be avoided. In what follows, the main characteristics of some important grammar development and corpus annotation tools will be presented in light of the criteria above.

II.1.3.2 Corpus Processing and Linguistic Development Software

A variety of corpus processing toolkits are available for research (or pedagogical) purposes. Examples include the FST toolkits from Xerox (XFST, (Karttunen et al., 1997)), the University of Helsinki (HFST, (Yli-Jyrä et al., 2006)) or the University of Stuttgart (SFST, (Schmid, 2005)), NLTK: the Natural Language Toolkit (Loper and Bird, 2002), the CL@RK System (Simov et al., 2003), GATE: the General Architecture for Text Engineering (Cunningham, 2002), and NooJ (Silberztein, 2004). Amongst available implementations of machine learning algorithms, we have to mention Treetagger (Schmid, 1994) and the CRF chunker (Phan, 2006).

The FST toolkits (XFST, HFST and SFTS) are a collection of software tools for the manipulation and processing of finite-state automata and transducers. They include tools for compiling lexicons, grammars or the intersection of these two into finite state trans-

ducers. They support a wide range of transducer operations (e.g. to compare, compose, minimize, intersect, determinize transducers etc., or to display the strings recognized by a transducer). A finite state transducer (FST) maps strings from the regular surface language onto strings from another regular language (analysis language). Transducers can be used to generate analyses for a surface form and to generate surface forms for an analysis. Transducer specification is made via system-specific languages made up by regular expressions. Important applications of FSTs are lemmatization, tokenization, lexicon representation, spell checking or low-level parsing. However, these pieces of software are mainly intended for the implementation of morphological analyzers and other tools which are based on weighted and unweighted finite-state transducer technology. According to Koskenniemi and Yli-Jyrä (2009), the main application domain of FSTs remains word-level processing and similar phrase-level processes. Full scale syntactic description of languages is usually done with more powerful frameworks, even if finite-state devices are used in some parts of them. They criticize freely available FS tools (like SFST and HFST), because in these toolkits "the functions of the basic calculus are more or less integrated with some interface for compiling rules or interpreting regular expressions and most packages have only limited documentation which makes it difficult for a programmer to extend or utilize the existing code." From the point of view of linguistic soundness, FST-based tools are not preferred as they accept only regular expressions as input, which cannot provide a correct model for several linguistic phenomena.

NLTK, the Natural Language Toolkit, (Loper and Bird, 2002) is a suite of Python program modules available under an open source license to support a variety of NLP tasks. NLTK runs on all platforms supported by Python, including Windows, OS X, Linux and Unix. Important advantageous features of NLTK are modularity, ease of use, consistency and extensibility. NLTK covers symbolic and statistical natural language processing. Each NLTK module is devoted to a specific NLP task: string manipulation, tokenization, POS tagging, collocation discovery, chunking, parsing, classification, probability estimation etc. Its visualization modules define graphical interfaces for viewing and manipulating data structures, and graphical tools for experimenting with NLP tasks. In particular, it is provided with a corpus annotation interface. NLTK includes a chart parser which accepts CF grammars as input. It also allows to use probabilistic CF grammars. However, NLTK

is still mostly used for pedagogical purposes rather than for implementing large scale CF grammars. This is because the toolkit is not intended to provide a comprehensive set of NLP tools; indeed, it is primarily designed for students in NLP. Thus, it is not optimized for runtime performance.

CL@RK System (Simov et al., 2003) is a corpus annotation tool which was conceived to minimize human effort during manual annotation of BulTreebank. The development of the CLaRK system started under the Tübingen-Sofia International Graduate Programme in Computational Linguistics and Represented Knowledge. ClaRK is an XML-based tool, compliant to a lot of already developed standards for corpus description, such as CES (Corpus Encoding Standard 2001) and TEI (Text Encoding Initiative 1997). The core of the CLaRK system is an XML Editor which is the main interface to the system. With the help of the editor the user can create, edit or browse XML documents. It has also been enhanced with facilities that support linguistic work, namely a finite-state engine that supports the writing of cascade finite-state grammars and facilities that search for finite-state patterns. It expresses regular expressions as input, entered via the graphical interface, and it allows automatic or semi-automatic annotation of the corpus. Moreover, it supports the XPath query language which facilitates navigation over the whole set of mark-up of a document. The basic use that was envisaged for the system is to facilitate manual corpus annotation in the BulTreebank project. Therefore, whereas the interface of ClaRK System is user-friendly, the engine has not been optimized for performance.

GATE (Cunningham, 2002) stands for General Architecture for Text Engineering. It is a freely available, open-source framework and development environment, based on Java and XML standards. The main motivation behind the creation of GATE is to reduce the amount of software coding needs by providing a general architecture. As an architecture, it defines the modular organization of a language engineering system. It comes with a set of prefabricated, customizable software building blocks. The prefabricated modules are able to perform basic language processing tasks such as POS tagging and semantic tagging. A wide range of applications have been developed within the framework of GATE, e.g. for named entity recognition or semantic annotation/information extraction. However, two types of barriers put limitations on the reusability of language processing modules: the incompatibility of type of information used by different modules, and the

incompatibility of the representation of the information in these modules.

II.1.3.3 Statistical Chunkers

Widely used probabilistic chunkers are the Treetagger (Schmid, 1994) and the CRF chunker (Phan, 2006) . Treetagger was conceived as a tool for annotating text with part-of-speech and lemma information. However, it can be used as a chunker for any language for which an annotated training corpus is available. Chunking, as described in the previous section, can be seen as a classification task where each token in the training corpus is labelled with a tag which denotes its position in (or outside) a chunk. The basic version of Treetagger (for English) used hidden markov models. Later on, it has been extended in order to be able to deal with languages with a more complex morphology and thus a bigger set of context features. A decision tree algorithm was added to the system which reduces the number of context parameters (the feature space) in order to provide more reliable probability estimates for low frequency words. Besides, a more sophisticated smoothing algorithm has been included, which yielded important improvements for languages where only a reduced amount of training data are available. The CFR chunker is based on conditional random fields. CRFs has been shown to offer advantages over HMMs or classification methods in sequential labeling tasks (Lafferty et al., 2001). In fact, CRFs can be seen as a generalized version of HMMs which relax certain assumptions about the input and output sequence distributions. They have the advantage over classification methods that they can model dependencies between annotations. Besides shallow parsing, they have been applied with succes to a variety of tasks including named entity recognition, POS tagging or learning syntactic functions from a treebank (Moreau and Tellier, 2009). A freely available implementation of CRF chunker for English by Xuan-Hieu Phan is available at http://crfchunker.sourceforge.net/.

II.1.3.4 NooJ

NooJ (Silberztein, 2003) is a linguistic development environment that allows users to construct large formalized dictionaries and grammars and use these resources to build robust NLP applications. It can also be used as a corpus annotation tool: whereas grammars and other linguistic resources can be applied to the text fully automatically,

NooJ also allows to select correct hits in a concordance table and thus perform a semi-automatic annotation.

NooJ has evolved from and shares many functionalities with an other piece of linguistic software, Intex (Silberztein, 1993). Intex's technology was based on the principles described by Silberztein (1987), and its first version was released in 1992. NooJ is largely different in its software architecture and linguistic methodology, but remains to a large extent compatible with Intex. Considerable differences between Intex and NooJ are always in the direction of simplification, especially with respect to the unified processing of morphological and syntactic grammars, by the same linguistic engine. In both systems, grammars are represented as graphs and can be edited via a graphical interface. However, as opposed to the strictly finite state approach of Intex, NooJ is not limited to FS formalisms. Grammars constructed by linguists in NooJ can be

- FS automata or transducers represented by graphs;
- graphs containing further embedded graphs;
- recursive graphs;
- finally, graphs enhanced by variables to store segments of the input. Grammars constructed using one variable correspond to pushdown automata and generate CF languages; grammars with more variables correspond to linearly bound automata and generate CS languages.

In particular, graphs enhanced with variables and constraints allow linguists to model long-distance dependencies and agreement phenomena in a straightforward and simple way. We will see in II.2 how agreement phenomena or lexical constraints can be handled with very simple and transparent NooJ grammars with only one path.

At the same time, NooJ is computationally efficient: it has the power of a linear bounded automata and processes CS grammars in linear time. It is characterized by a bottom-up architecture where each level of analysis is processed by a different, specialized (and therefore efficient) computational device (Silberztein, 2007). One can use simple devices such as finite state machines to represent a large number of linguistic phenomena, whereas more complex grammars such as CF grammars or even context sensitive ones

can be used for more complex phenomena. It is therefore possible to reach beyond mildly context sensitive grammars.

The linguistic/methodological background of NooJ corresponds to the point of view of Gross (1975) on language description: it was conceived as a tool to implement local grammars to formalize local linguistic phenomena, and proceed step by step towards a complete description of sentence structure. Lexical properties play a central role in this corpus-driven methodology. The starting point is not an enumeration of syntactic rules which precise the syntactic behavior of grammatical categories. Instead, categories are established upon a detailed examination of the corpus. As noted by Gross (1975) and Gross and Danlos (1988), a precise linguistic description needs to rely highly on the properties specific to lexical units. Hence, the role of generalization is reduced: a large-scale linguistic analysis requires using a much finer grained categorization than traditional POS categories. Moreover, the study of French verbs and the syntactic contexts in which they can occur lead to the conclusion that "no two verbs have the same set of syntactic properties; as a consequence, verbs have to be described individually and not in terms of intensional classes." (Gross, 1991). Besides syntactic idiosyncrasy, the other argument for individual lexical representation of linguistic units comes from the field of semantic interpretation. Several linguistic units, even though they show a predictable syntactic behavior, need to be enumerated in the lexicon because of their non compositional semantic interpretation: "The proportion in the lexicon of idiomatic sentences, of metaphoric and technical sentences that have non compositional meanings, is very high. (...) The consequence is that they must be described individually, that is without reference to other classes of lexical combinations or of interpretation rules." (Gross, 1991).

Accordingly, the study of lexical information about the behavior of simple or complex lexical units has gained considerable importance both in the field of linguistic description and natural language processing (Gross and Danlos, 1988; Mason, 2004; Silberztein, 2007; Laporte, 2000). It is thus highly important for a grammar to be able to include rules with different degrees of generalization, as well as lexical rules specific to individual items. We made use of this lexicalization potential while constructing our grammars; however, at this first stage of shallow parsing, our work concentrated on the general structure of sentence constituents and dealt with lexicalized phenomena only to a lesser extent.

Since NooJ (like its predecessor Intex) was conceived to comply with this principle: its structure promotes the construction and use of hierarchical dictionaries and lexical resources (Silberztein, 1987, 1993). It uses a bottom-up approach to the formalization of natural languages: linguistic description starts at the level of morphology and normalization, and then proceeds to higher and higher linguistic levels, moving up towards the sentence level. "NooJ provides parsers that operate in cascade at each individual level of the formalization: tokenizers, morphological analyzers, simple and compound terms indexers, disambiguation tools, syntactic parsers, named entities annotators and semantic analyzers." (Silberztein, 2004) Each type of grammar (morphological, syntactic etc.) is ran by a different engine, optimized for the given linguistic module. An important feature of NooJ is its 'non-destructive' annotation system. An annotation is a pair (position, information) that states that a certain sequence of the text has certain properties. When NooJ processes a text, it produces a set of annotations, stored in the Text Annotation Structure (TAS); annotations are always kept synchronized with the original text file, which is never modified. NooJ's parsers communicate via the Text Annotation Structure that stores both correct results and erroneous hypotheses (to be deleted later). The non-destructive engine makes it possible to perform a large number of operations in cascade or in parallel. For instance, the integration of morphology and syntax allows NooJ to perform morphological operations inside syntactic grammars. Therefore, while the structure is modular, these modules are not strictly separated: on a more abstract level, information produced by lower level grammars will remain accessible for higher level grammars. This makes it possible to delegate decisions to subsequent levels of analysis. The best illustration of this feature is the tagging/disambiguation approach put forward by Silberztein (2007): he advocates that certain ambiguities cannot be resolved at the morphological level using morphological analysis and context features, but are best treated at higher (syntactic, semantic) levels. The decision about certain ambiguities is thus delegated upwards until they are solved or retained as real structural ambiguities.

NooJ is implemented within the .NET framework[2], following a Component-Based Software approach. It benefits from many built-in facilities of the .NET framework, and hence it allows to process multilingual texts and corpora, in over 100 file formats. By

[2]A Linux- and MacIntosh-compatible implementation also exists under MONO.

reason of its robustness, modular bottom-up structure and XML-compliance it has been chosen as a tool to develop a complete tool chain for analyzing Hungarian corpora (Váradi and Gábor, 2004). The tool chain includes tokenization, sentence splitting, lemmatization, morphological analysis (Vajda et al., 2004), named entity recognition and shallow parsing (Gábor, 2007). The tool enables users to annotate their texts and run queries on its elements according to morphological features or syntactic functions.

To sum up, the following advantages of NooJ motivated our choice:

1. not limited to a particular formalism or type of grammar,
2. processing speed, modularized optimization,
3. allows a linguistically transparent description,
4. annotations accessible and modifiable at any level of processing,
5. provides a complete tool chain for corpus processing,
6. user-friendly environment; functions facilitating grammar development; easy access for non-linguist users as well.

II.2 The Shallow Parser for Hungarian

II.2.1 Sentence Constituents in Hungarian

Hungarian language has some relevant characteristics which make it a challenge to process it with strictly finite state tools. First, it is a highly inflective language: nominals show a much extended inflectional paradigm with about nineteen different cases (the exact number depends on the criteria used, which differ among linguists and theories). A typical case suffix is the rightmost suffix on a noun or a word belonging to a nominal category. Being part of the inflectional paradigm, typical case suffixes are totally productive. Besides case suffixes, Hungarian also has a rich postpositional system. From a syntactic point of view, case suffixes and postpositions have a similar function: they encode the grammatical function of the nominal constituent. Although they are not in a complementary distribution, as postpositions follow case-marked NPs, we still consider

that they have the same syntactic function: when added to top-level NPs, they mark the type of the dependency relation between the NP and the predicate. Verbs can subcategorize either for cases or for postpositions; postpositions, in turn, subcategorize for the case of the NP they are added to. Both case suffixes and postpositions can be ambiguous: they can encode more than one potential grammatical function depending on the context.

Second, Hungarian language is non-configurational in that it encodes constituents' grammatical functions not in their respective order but at the morphological level. This means that dependency relations between top-level sentence constituents, i.e. the predicate and its complements and adjuncts, are not reflected in the surface syntactic structure. É. Kiss (2002) argues that the Hungarian VP has a flat structure: all complements are generated in parallel behind the verb, without a hierarchical structure. Some of these complements can subsequently be moved to a preverbal position, e.g. focus, topic, contrastive topic etc. Nevertheless, these positions correspond to discourse functions rather than grammatical functions. Consequently, Hungarian language is non-configurational with respect to grammatical functions.[3] Surányi (2006) on the other hand claims that there is a configurational encoding of verbal complementation in the deep structure, but in surface syntax these constituents can be mixed to produce alternative orderings: this phenomenon is known as *scrambling*. Whichever analysis we accept, the conclusion is that surface constituent order does not reveal the internal structure of the verb phrase in Hungarian.

However, as opposed to the sentence level, NPs and other nominal constituents present a bound internal word order with the case-marked head being the rightmost element. We therefore hypothesized that local grammars can perform fairly in recognizing NPs, APs and other phrases with a strict word order.

[3] É. Kiss (2002) claims that Hungarian is a discourse-configurational language, which we do not contest: however, instead of discourse functions, we are interested in the correlation between surface syntax and grammatical functions (or the lack of it).

II.2.2 Parsing in NooJ. Structure of the Grammars

II.2.2.1 Overall Structure

The goal of our work was to build up a large-scale shallow parser for Hungarian, implemented in the corpus processing tool NooJ. The syntactic parser is a part of the Hungarian NooJ module (Gábor, 2007). It is part of the Hungarian NooJ tool chain which includes a morphological analyzer (Vajda et al., 2004) based on the vocabulary of the Concise Dictionary of Hungarian (Pusztai, 2003) and on the inflectional description of Elekfi (1997). The NP grammars by Váradi (2003) were integrated into the Hungarian parser at an early stage.

The shallow parser is implemented in NooJ as a bottom-up cascade of patterns which exhibit different degrees of specification. Patterns or rules are represented by graphs, and they perform a partial analysis at the sentence level. This method ensures robustness: the analysis of the sentence does not fail even if some part of the input has not been recognized by any prior level of analysis. Another advantage it offers is that lack of analysis at earlier steps of the process (e.g. tokens unknown by the morphological analyzer) could be inserted in the pattern, thus errors made at an earlier stage does not necessarily propagate upwards. The structure of the parser can be conceived as an ordered set of grammars applied to a text in cascade. Grammars typically correspond to general phrase-building rules accepting POS categories as input; however, they can exhibit different degrees of lexicalization and even extend to very specific lexical patterns. Indeed, one of the most attractive features of NooJ grammars in parsing is their flexibility: they allow for modeling lexically constrained collocations, semi-frozen expressions and syntactic rules with the same description method, i.e. with graphs. The parser comprises the following major levels, corresponding to the basic tasks:

1. sentence segmentation;
2. chunking (noun phrases, adjectival phrases, postpositional phrases);
3. constituent coordination;
4. clause boundary detection grammars;
5. predicate and VP detection (verb prefixes, auxiliaries and infinitives).

Tasks 1) and 2) are accomplished by a series of grammars organized in a cascaded process. Grammars are represented by structured sets of graphs whose outputs are linguistic annotations: constituent labels, completed by attributes that describe properties specific to the given phrase. Such properties are e.g. the number, case or the semantics of the NP's head noun, which can be percolated up to the phrase via built-in NooJ functions. Most grammar patterns refer to the annotation which was inserted to the text by the preceding grammars, thus all the information of lower level grammatical analysis have to be accessible for syntactic graphs. Contrary to Abney's method (Abney, 1996), grammars' outputs do not replace the input at subsequent levels: they are stored in parallel to the source text. Hence, the output of previous analysis as well as unannotated sequences can be referred to in the grammars. This also reduces the risk of destroying the full analysis of the sentence by leaving a string unanalyzed at a lower linguistic level, because such unanalyzed strings may still serve as a valid input for subsequently applied local grammars. Another useful feature of this method is that rules can easily modify the output of the lower level grammars if necessary.

II.2.2.2 Disambiguation

An important feature of the parser is that it does not require the input text to be disambiguated. This complies with the philosophy of NooJ, according to which an important amount of ambiguities, even at the word level, cannot be solved without having recourse to syntactic or semantic information, as well as information about the phrase structure. However, this kind of information is not available when the POS tagging should be performed, before syntactic analysis. Thus, NooJ supports parallel analyses during syntactic processing. However, we aim at producing an unambiguous final output, to be exported as a correct XML structure. Hence, syntactic grammars have to be designed and applied in a way that they produce an unambiguous annotation. Our shallow parser accepts two types of input: raw texts, to be analyzed with the NooJ-internal Hungarian morphological module, or POS-tagged texts, analyzed with the tool chain developed for tagging the Hungarian National Corpus (Oravecz and Dienes, 2002) (HNC POS tagger from now on). The Hungarian morphological analyzer in NooJ does not perform any disambiguation. NooJ in turn provides several utilities for manual disambiguation. Two types of POS

ambiguities can be differentiated: systematic and contingent ambiguities. Systematic ambiguities occur when two or more suffixes coincide in the same inflectional paradigm: e.g. in Hungarian, past tense suffixes and past participle suffixes coincide for many verbal paradigms, resulting in a large set of ambiguous verb forms. Disambiguation by complex syntactic grammars can be especially fruitful for disambiguating this kind of ambiguous word forms. On the other hand, contingent ambiguities either have to be dealt with one by one, or can be left to the grammar. The disambiguation module in NooJ allows to write disambiguation grammars of both types, and they can be applied at any level of the processing. The shallow parser presented here includes a very limited set of explicit disambiguation grammars for a few frequent word forms (e.g. ég – *sky* vs *burn* or mert – *because* vs *ladle out*). The rest of the ambiguities are not handled explicitly but a big set among them are implicitly disambiguated by the syntactic grammars.[4]. To provide an alternative to this implicit partial disambiguation, we also created a conversion script to transform the output of the HNC POS tagger into NooJ-compatible XML. The advantage of using this tool is that a better coverage can be achieved, especially for special text types such as spoken language corpora. The morphological analyzer in the HNC POS tagger is Humor (Prószéky and Tihanyi, 1992), which includes a guesser, whereas the coverage of the NooJ-internal Hungarian morphological analyzer is limited to the 60.000 entries of the Hungarian Concise Dictionary. The conversion tool allows a choice between working with pre-tagged, disambiguated corpora or following the method of Silberztein (2007) by performing a gradual disambiguation.

The preprocessing of the input includes sentence segmentation and the application of the disambiguation grammars. Tokenization is a built-in functionality within NooJ, based on the language-specific alphabet: each character that does not belong to the alphabet is considered as a separator. However, this initial tokenization can be overwritten by additional rules, e.g. for named entity or multiword expressions recognition.

[4]This means that the grammars do not explicitly associate POS tags to words inside chunks, but presume that the tag which was matched by the grammar is the correct one. This implicit disambiguation could easily be explicited; however, this does not imply that every word inside the chunks would be disambiguated

II.2.3 Constituent Chunking

The starting point of the syntactic parsing is chunking: the cascaded application of grammars responsible for the recognition and labeling of phrases that may express potential verbal arguments in the sentence: NPs, APs, Postpositional Phrases. They are all characterized by a relatively bound word order, and their head occupies the rightmost position within the phrase. The first role of the grammars consists in finding and annotating the edges of the phrases. Although chunking is generally identified as the task of recognizing base phrases, i.e. without embedding or recursion, and thus can be accomplished by finite state grammar rules, we did not differentiate between base phrases and complex phrases. Embedded structures and agreement phenomena occur even at the lowest levels of linguistic description of Hungarian sentence constituents, which makes it difficult and useless to strictly separate phenomena which can be described by FS grammars from more complex ones. Moreover, our goal was to annotate top-level sentence constituents, and as we will see, this requires the use of more complex description methods than FSAs.
As NPs show the biggest complexity among top-level sentence constituents, their processing requires the most complex grammars. The basic structure of a typical NP is composed of:

- a determiner and or/numeral quantifier on the left edge of the phrase;
- a nominal head on the right edge;
- a finite number of (mostly adjectival) modifiers which precede the verb;
- an eventual possessor which also appears between the determiner and the head noun.

Although this structure is very schematic, it can serve as a starting point for the description. In what follows, we will give an overview (based on (É. Kiss, 2002) and (Váradi, 2003) on the internal structure of NPs in light of the factors that may complicate their automatic labeling, and present the solutions we opted for when constructing the chunker.

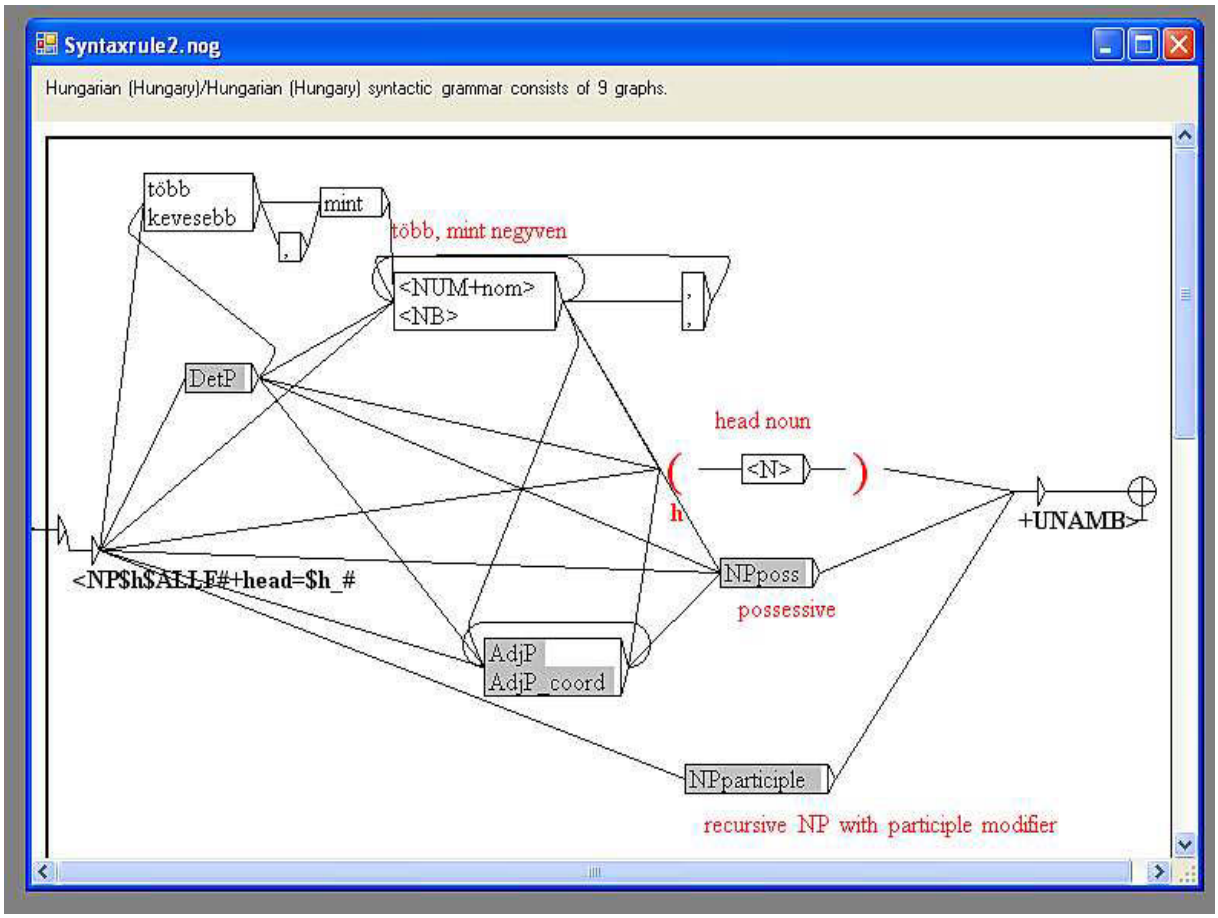

Figure II.1 : Top-level structure of the NP grammar

II.2.3.1 Determiners

First, as Váradi (2003) notes, determiners cannot be relied on as anchor points to indicate the boundaries of the noun phrase since they can be omitted. This especially occurs in the case of a generic interpretation or when the NP occupies a verb pre-modifier position[5]:

Diákok is járnak bulikba.

Students also go to parties.

János moziba ment.

John cinema+ILL went.

John went to the cinema.

[5]The examples below are taken from (Váradi, 2003)

Determiners, when they are present, can show a more complex structure than a simple definite or indefinite article. Numerals can appear both in the position of the determiner or following it. Definite numerals (e.g. 3, három – *three*) and some of the indefinite quantifiers (annyi – *that many*, néhány – *a few*) can be further modified by comparative structures (több, mint – *more than*). The comparison can be specified:

néhánnyal kevesebb, mint 750 jelölőszelvény
a few+INS less, than 750 endorsement-coupon
'a few less than 750 endorsement coupons'

Although Hungarian has a very rich morphology and especially complex nominal case system, NP-internal agreement is not a typical feature of Hungarian. Adjectives modifying the head noun do not agree with it; they are invariable in this position. Demonstrative pronouns, however, agree with the head noun in case and in number:

azt a tányért
that+ACC the plate+ACC
'that plate' (accusative)

The correct modeling of this phenomenon requires thus introducing a constraint in our NP grammar. Since the sequence constituted by a pronoun, a determiner and a noun is frequent and structurally ambiguous, not considering case agreement would lead to a high number of erroneous analyses, i.e. sequences that should be labelled as [PRO] [DET N] could be annotated as a single NP of the form [PRO DET N]. Thus, recourse to feature checking is necessary even at this low level of the chunking task.

II.2.3.2 Adverbs inside NPs

Various types of adverbs can occur inside NPs in different specific positions. Their lexical properties largely influence the structural positions that they can fill: some adverbs modify adjectives; a different set of adverbs precede quantifiers. Many of these adverbs can appear as a top-level sentence constituent, directly modifying the predicate. In addition,

participle modifiers of nouns can be extended by another set of adverbs, which coincides with the adverbs modifying predicates. In order to avoid false bracketing results, at this point we had to lexicalize our grammar and select adverbs which can occur in these specific positions. The lexical features introduced in our grammar cannot be directly linked to semantic properties. Adverbs modifying quantifiers include e.g. mintegy *'about'*, körülbelül *'approximately'*, pontosan *'exactly'*, nagyjából *'roughly'*. The set of adverbs modifying adjectives is larger, covering more semantic fields, e.g.: különösen *'especially'*, nagyon *'very'*, teljesen *'completely'*. A comprehensive list of adverbs occurring in these two positions were added to the grammar of adjectival phrases, as well as to the NP grammar.

II.2.3.3 Possessive Structure

Possessive structures introduce recursion in the NP grammar, since NP-internal possessors can be expressed by NPs. In Hungarian, it is the possessee that is inflected for possessive relation, whereas possessors can either be in nominative or in dative (there is no genitive case in Hungarian). The possessive inflectional suffix on the head noun agrees in number and person with the NP or pronoun which denotes the possessor. Whereas the inflection on the head noun serves as a good indicator of the presence of a possessive structure, the major complication arises from the fact that dative possessors can leave the NP and move apparently independently in the sentence:

Érvénytelenítették a jelölteknek a szavazatát.
were-voided the candidates+DAT the vote+POSS

A jelölteknek érvénytelenítették a szavazatát.
the candidates+DAT were-voided the vote+POSS
The votes of the candidates were voided.

A jelölteknek megszámolták, majd érvénytelenítették a szavazatát.
the candidates+DAT were-counted, then were-voided the vote+POSS
The votes of the candidates were counted, then voided.

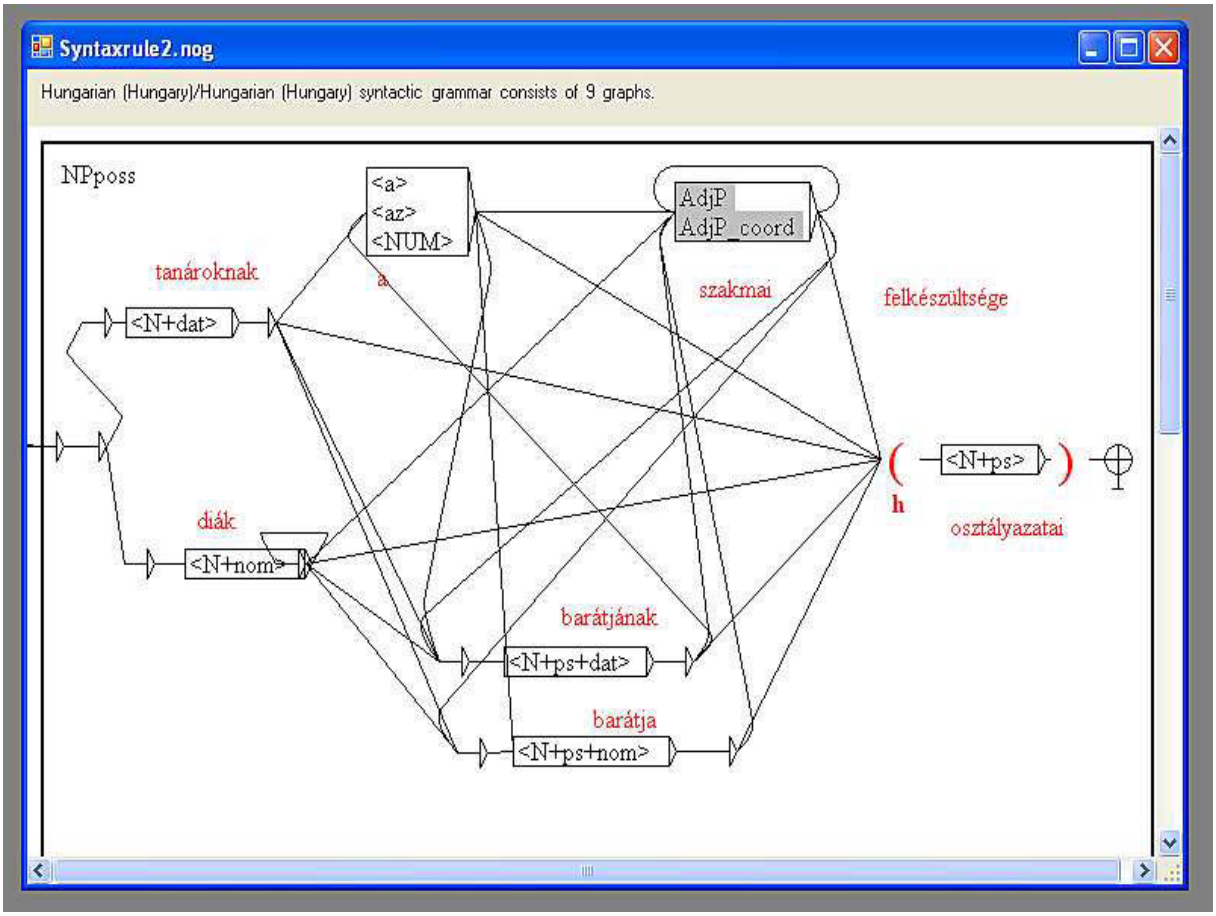

Figure II.2 : Grammar for Possessive NP

The strategy we adopted was to consider continuous NPs with a dative possessor as one constituent (contrary to the annotation of the Szeged Treebank (Csendes et al., 2004). This constituent can be continuous: in this case, it is distributionally equivalent to a possessive NP with the possessor in nominative case and, correspondingly, we annotate it as such. On the other hand, the constituent can be discontinuous, in which case we do not currently have a solution for annotating it in our framework. Since in this chunking phase we are interested in bracketing continuous constituents, we associate such discontinuous possessives with two NP tags. In order to disambiguate between dative possessors and other top-level sentence constituents in dative, we need verbal argument structure to be recognized. Thus, this task is propagated to a higher level of syntactic analysis.

II.2.3.4 Ellipsis

It is not only the determiner (the right-hand edge of the NP) that can be omitted, in certain cases the head noun itself may be absent. Ellipsis can take place if the head is identifiable from the context. In cases like this the case marking suffix appears on the last constituent of the remaining NP:

A képviselő háromszáz érvényes szavazatot kapott, és egy érvénytelent adott le.
the deputy received three-hundred valid vote+ACC and one void+ACC gave prefix
The deputy received 3.000 valid votes, while he casted a void one.

Adjectives and quantifiers can thus function as NPs and as such, they are inflected for case the same way noun heads are in complete NPs. We can talk about ellipsis when the head noun missing from the NP is present in the tight textual context (usually in the same sentence). However, the noun can also be omitted when it can be identified either from the extra-linguistic context by general knowledge, or from the semantic selectional constraints of the predicate. Evidently, the boundary between these categories is vague: how do we know that the AdjP replaces the NP complement of a verb or it fills a complement position in its own right? An argument in favor of the latter assumption is provided by the observation that a variety of verbs accept or even require adjectival phrases as complements. Typical adjectival complements have a predicative function (lesz *'become'*, változik vmilyenné *'change into'*, ábrázol vmilyennek *'depict as'*). These adjectives are inflected for case and can have modifiers, but they are not preceded by a determiner, since they are not referring to an entity. However, the syntactic role of top-level AdjPs is not limited to the predicative function, they participate in a set of more or less idiomatic complementation patterns:

- rosszat gondol vmiről *'to think bad (things) about sg'*
- kedveseket mond vkinek *'to say nice (things) to sy'*
- teszi a szépet vkinek literally: to do the nice, *'to woo sy'*

Structures in the first two examples are productive, open patterns: the adjectives in these expressions can be replaced by other, semantically compatible ones, and they can

be put into comparative or superlative. The third example on the other hand is idiomatic: the adjectival phrase is invariable since the meaning of the structure, unlike the other ones, is completely non-compositional. As these structures need to be included in the lexicon as verbal complements, we can say that in some cases AdjPs are explicitly subcategorized for by verbs. Thus, the category of the top-level, non-predicative AdjP will be present in our grammar. In a number of other cases, the grammatical category of verbal complements is underspecified: they can be expressed either by NPs or by (referential) AdjPs. If we take ellipsis into account, we can state that any verbal complement — with the exception of lexically bound idiomatic complements — which can be expressed by an NP can equally be expressed as an AdjP. Therefore, we will consider structures like the above as instances of top-level adjectival phrases and not as NPs with a missing head.

II.2.3.5 Participles

The NP grammar is divided into two grammars: one of them describes 'simple' NPs (including those with a possessive structure), while the other one — which, in turn, contains self-recursion — is designed for recognizing NPs that contain a participial structure. Two types of verbal derivates, namely past and present participles, appear frequently inside noun phrases as modifiers. In effect, the predicative use of participles is by and large avoided in written texts, which means that participles almost always occur inside NPs. From the point of view of chunking, participles are peculiar in that they can take any number of their own arguments, or even adjuncts, in front of the NP head. Hence, NPs containing participle pre-modifiers can be embedded recursively in each other. Consider the following examples:

[NP [NP a bizottság] által javasolt módosítások]
the committee+NOM by suggested modifications+NOM
the modifications suggested by the committee

[NP [NP [NP a parlementben] felállított bizottság] által javasolt módosítások]
the parliament+INE set-up+PARTIC committee+NOM by suggested modifications+NOM
the modifications suggested by the committee set up in the parliament

[NP [NP [NP [NPa parlamentben] felállított bizottság] által javasolt módosítások] folyományaként elfogadott törvények]

the parliament+INE set-up+PARTIC committee+NOM by suggested modifications+NOM following accepted laws+NOM

the laws accepted following the modifications suggested by the committee set up in the parliament

As shown by the examples, the head of the NP can take a (past or present) participle as a modifier. Being verbal derivates, participles can in turn be accompanied by complements or adjuncts of the base verb. Since complements and adjuncts can be expressed as NPs, the resulting structure will contain an NP inside an NP, which is indeed an example of center embedding. This phenomenon is not rare in Hungarian. Indubitably, the frequent appearance of one (two, three) NP modifier inside an NP does not dissolve the doubts raised earlier, namely, whether structures with a potentially infinite number of embedding only exist in linguistic competence but not in performance, due to the limited processing capacity of human brain. However, while we want our grammar to serve for annotating a corpus, we also aim to provide a linguistically plausible model of Hungarian language phenomena. Moreover, for a rule-based parser to be easy to handle, adapt and improve, its rules have to be easily interpretable. This requirement answers to the same principles as the explanatory power: the grammar has to capture the observation that participles inside an NP can have NPs as modifiers (and not just accidentally the same set of symbols that happen to appear elsewhere in our grammar). Note that there is a difference between embedded NPs and top-level NPs: embedded NPs cannot be in nominative[6]. The subject function of the participle is marked with the postposition által. Thus, the actual recursion is one level deeper, at the level of embedded NPs. However, to avoid unnecessary multiplication of embedded grammars, we added self-recursion at the top level and handled this phenomenon with a constraint which checks whether the case of the embedded NP matched by the grammar is other than nominative.

Besides recursion, the complexity of the grammar dealing with participles is due to

[6]Unless they are followed by a postposition

the fact that participles can be extended by almost any sort of arguments or modifiers that the base verb can have. This corresponds basically to the complexity of the top-level sentence structure, but within the boundaries of the NP: i.e., between the determiner and the head noun. A theoretically correct solution would be to check whether the modifiers present before the participle are subcategorized for by the base verb of the participle. This solution, however, goes far beyond the scope of chunking. Thus, we settled for a heuristic, yet efficient solution: the NP grammar accepts any potential complement, as long as it occurs between the determiner and the participle. This approach is an approximative one, since it only applies to NPs starting with a determiner. Nevertheless, NPs modified by a participle but not containing a determiner are rare, which boils down to two reasons: first, they show significant structural ambiguity which is difficult to process even for humans. Second, NPs modified by participles are usually introduced as definite entities, thus requiring definite articles. This is because the participle is meant to complete the reference of the NP by a more or less precise, but usually concrete description of an earlier event, of which the reference of the NP is a participant.

Besides delimiting phrases, the second function of the chunking grammars is to extract relevant features of the head of the phrase, and percolate these features up to the level of phrases. The result is an attribute-value structure which can be exported as an XML annotation. Relevant features include number, person (agreement features), case, possessive, and potentially semantic features. They are all used in verbal argument specifications; hence, they are needed for matching verbal argument frames at a later phase. The grammar uses the feature inheritance function of NooJ to copy the head's feature values to that of the phrase via built-in functions.

II.2.3.6 NP Coordination

The third operation of the NP grammar serves the recognition of coordinated NPs. The difficulty of this task is induced by the fact that conjunctions which occur between coordinated NPs may also connect clauses. The precision of the grammar can be increased by using the output of the NP annotation, more precisely the value of the case feature: only NPs with the same case suffix can be coordinated. At this point, we make use of

NooJ's constraints to check whether the two NPs on both sides of the conjunction have the same case suffix. At a previous step, we recognized the two simple NPs and identified their head noun; the case feature of the head has subsequently been percolated to the higher level annotation as values of the Case feature of the NPs. The coordination grammar hence takes the two NPs as input (as in a cascade of grammars). The two NPs are stored in two variables; when storing a recognized sequence in a variable, NooJ provides means to refer to its lexical properties, in our case the Case feature. The constraint then checks whether the two NPs share the same value for Case:

<$NP1$case=$NP2$Case>

This formalism allows to deal with coordination and, more generally, feature agreement phenomena using a compact representation.

II.2.3.7 Postpositional Phrases

Since postpositions do not show a homogeneous syntactic behavior, their categorization needed to be refined in order to design a grammar which recognizes postpositional phrases correctly. Three types of postpositions can be differentiated with respect to their morphological and syntactical subcategorization. The first and biggest set of postpositions require NPs in nominative case and follow directly the NP they subcategorize for:

[POSTP [NP az asztal] alatt]
the table+NOM under
'under the table'

Other postpositions subcategorize for different cases. This property was encoded in the dictionary of function words used for the syntactic analysis. Feature agreement is verified in the grammar via constraints.

[POSTP [NP a határon] túl]
the border+SUP over

'over the border'

Finally, there are postpositions, originally derived from nouns, which behave similarly to possessive structures in that the NP they take can be either in nominative or in dative. In the latter case, an additional determiner (definite article) can show up between the NP and the postposition:

az igazgató számára
the director+NOM for
az igazgatónak a számára
the director+DAT the for
for the director

II.2.4 Predicates

Predicates were defined as finite verbs which subcategorize for complements or can be modified by adjuncts, all of these being expressed by top-level sentence constituents. The important features to be retrieved for a predicate are its lemma, tense, number and person. However, predicates can appear as infinitives accompanied by modal or temporal auxiliaries. In this case, the auxiliary is inflected for tense, number and person, but the complementation pattern belongs to the infinitive. Nevertheless, not all the infinitives are modified by an auxiliary: they often belong to the argument structure of a content verb, e.g. verbs of motion can productively take infinitives. Content verbs with an infinitival complement or adjunct can express their own arguments, and the arguments of both verbs can appear in the same sentence. The same statement holds for adjunct infinitives as well. This can result in structural ambiguity when the attachment of a complement or, more typically, an adjunct cannot be determined based on surface structure:

Tegnap megpróbáltam teniszezni az öcsémmel.
Yesterday I tried to play tennis with my younger brother.
Tegnap [megpróbáltam [teniszezni az öcsémmel]], de nem ért rá.

Yesterday I [tried [to play tennis with my younger brother]], but he was busy.

Tegnap [megpróbáltam [teniszezni] [az öcsémmel]], de nem tudtunk pályát foglalni.

Yesterday [I tried [to play tennis] [with my younger brother]], but we were unable to book a court.

The correct attachment of adjuncts in sentences like the above would probably require extra-linguistic information.

We adopted the hypothesis according to which auxiliaries differ from content verbs with an infinitive complement (or adjunct) in that auxiliaries do not have other modifiers, they only add a piece of modal, temporal or aspectual information to the meaning of the infinitive. Thus, they cannot have other complements in the sentence: top-level constituents belong to the infinitive, which in this case can be considered as the predicate of the sentence. The definition of auxiliaries was carried out according to the criteria defined by Kálmán C. et al. (1989). The grammar recognizes infinitives accompanied by auxiliaries as predicates, and yields an annotation which contains the lemma of the predicate and the lemma of its auxiliary. Besides, the relevant morphological properties of the auxiliary (tense, mood, person, number, definiteness) are copied to the predicate's feature structure so that its agreement with the subject (number, person) and the object (definiteness) could be checked at a potential deep analysis phase. This method ensures that verbs annotated as predicates will correspond to our definition: they will be annotated for tense, number, person and other morphological features, so that the argument structure of the sentence can be correctly matched at the level of deep parsing. On the other hand, when the finite verb is not an auxiliary according to our criteria, it will be marked as a predicate, and the infinitive is simply annotated as an infinitive, i.e. as a potential complement or adjunct of the finite verb.

The predicate's grammar is also responsible for the correct lemmatization of predicates. This task is complicated by the fact that verb prefixes (preverbs) can be detached from the verb under certain conditions, namely in case of negation or focusing. Negation can easily be described, but focusing is often unnoticeable from the structure of a written sentence (though it has very specific properties on the level of prosody). In sentences with focus, the focused constituent immediately precedes the predicate. If the predicate has a verb prefix, it is detached from it and will follow the verb. Many preverbs, however,

are homonymous with adverbs or pronouns and thus produce ambiguity when they are detached. The only way to deal with this problem is to check, by means of a lexical constraint, whether the lemma of the predicate can occur prefixed by the given word form, i.e. whether they form a potential lexical unit together.

```
többé nem <PRED Number="sg" Person="3" Definitness="indef"
head="visszatér">tér  vissza</PRED>

<NP Number="sg" Case="acc" head="Kossuth">Kossuthot</NP> <PRED
Number="pl" Person="1" Definitness="def" TenseMood="past"
head="odatesz">tettük  volna oda</PRED>,

<AUX>sikerült</AUX> <PRED Number="sg" Person="3"
Definitness="indef" TenseMood="past" modified_by_aux="sikerül"
head="megállapodik">megállapodniuk</PRED> <NP Number="pl"
Case="dat" head="fél"> a feleknek </NP>

<NP Number="sg" Case="nom" head="megye">a megye</NP>
<AUX>próbáljon</AUX> <PRED Number="sg" Person="3"
Definitness="indef" TenseMood="subj" modified_by_aux="próbál"
head="segít">segíteni</PRED> <NP Number="sg" Case="sup"
head="maga">magán</NP>
```

Figure II.3 : Predicate annotation.

II.2.5 Clause Boundary Detection

As a sentence may consist of more than one clause, each of them having its own predicate with its arguments, it is a necessary step in shallow parsing to find and annotate clause boundaries. We consider clauses as the domain inside which verbal argument frames have to be matched. We have to note, however, that clauses themselves can also play the role of a complement. Clause boundary detection is a problematic task for several reasons. Since in Hungarian not all clauses contain a finite verb and the order of phrases within the clause is free, it did not seem feasible to describe the structure of the Hungarian clause as such. Therefore, we decided to try to describe the local context of the boundary itself. As a working hypothesis, we accepted that there is a finite set of clause boundary markers (i.e. punctuation marks, conjunctions, certain adverbs and

pronouns) whose occurrences (or co-occurrences) reveal the exact location of the boundary. Gábor et al. (2003) accomplished a corpus-based analysis of the boundary markers that resulted in a classification of the conjunctions, adverbs and pronouns according to the position they occupy in the clause. These findings have been incorporated in the clause boundary detection module. The module consists of cascaded local grammars referring not only to the boundary marker elements but to the syntactic annotation produced by the earlier steps of the process. They are all organized around the types of probable boundary markers. Finally, the grammar is supplemented by a "guesser rule". Since we supposed that there is always a clause boundary between two finite verbs in a sentence, but some of them may not be covered by the rules, it is necessary to examine all the points in the domain which are susceptible to be boundaries. Thus, this rule operates on sequences between finite verbs which do not contain any clause boundary annotation yet, and it annotates possible boundaries (i.e. punctuation marks and conjunctions at the top sentence level). Clause boundary grammars are based on a typology of potential boundary markers. The following markers were taken into consideration:

1. Relative clauses, introduced by a relative pronoun (optionally preceded by a comma);

2. Clauses introduced by one conjunction. If the conjunction is preceded by a comma and it is not inside an already marked phrase, the clause boundary annotation is inserted before the conjunction. However, some conjunctions do not occupy the first position in the clause but may be preceded by one (top-level) constituent – in these cases, the annotation is inserted between the comma and the phrase.

3. Citations. They are followed by a dash, which is directly followed by a finite verb (e.g. *declare, announce, report*)

4. More than one adjacent conjunctions, or one conjunction and one question word: they are labeled as boundary markers if the conjunction has not been recognized by earlier grammars.

5. Question words, if they co-occur with a comma — adjacently, or with at most one top level constituent placed between them — are also exact boundary markers.

6. Sentence boundaries are naturally considered as clause boundaries.

7. Past and present participles are not considered as predicates in our grammar, thus, they do not form a clause-like domain with their arguments. On the other hand, adverbial participles and their arguments may produce a structure similar to clauses with a finite predicate, in that they can be expanded by complements, even by complement clauses, as shown by the example below. These clauses could not be linked to the finite predicate of the main clause, as they are in predicate-argument relation with the adverbial participle. For example:

Nemzetközileg egységes könyvviteli normák általános bevezetésére tett javaslatot a Price-WaterhouseCoopers, [clauseboundary] azt remélve, [clauseboundary] hogy ily módon visszaszerezhető a befektetők bizalma.

Price WaterhouseCoopers suggested to introduce international accountancy standards, [clauseboundary] hoping [clauseboundary] that this will help them regain the investors' confidence.

Finally, the guesser rule is applied to the text. This rule is applied to sentences where the previous grammars did not identify a boundary between two predicates. Any punctuation mark or conjunction (external to phrases) is then annotated as a potential clause boundary.

II.2.6 Evaluation

The chunker was evaluated on a subcorpus of the manually annotated Szeged Treebank (Csendes et al., 2004). As the annotation guidelines of the Treebank do not completely coincide with the output of our chunker, we concentrated on the evaluation of the chunker uniquely for the task of NP chunking. The other important functions of our chunker differ significantly in their principles from the coding of the Treebank, especially with respect to the notion of verb phrase and to the delimitation of propositions.

The test corpus is composed of an extract of 500.000 tokens (including punctuation) from the Treebank. The same subcorpus was used to evaluate HunChunk, the stochastic NP chunker of Recski (Recski and Varga, 2009; Recski et al., 2010). This chunker uses the Start/End labeling convention (instead of the more frequently used IOB convention):

the tag set includes: O (for elements outside the NP chunks), B-NP (first token inside the NP), E-NP (last token inside the NP) and I-NP (any token in-between). The test corpus, converted to this format, was provided by courtesy of Gábor Recski. We analyzed the corpus with our chunker, using the Humor morphological analyzer and the POS tagger of Oravecz and Dienes (2002) as input to the chunker. Subsequently, we converted the XML output of our chunker to the Start/End format and compared to the gold standard.

Precision	61.20%
Recall	73.40%
F-measure	66.74%

Table II.1 : Evaluation on Szeged Treebank. Full matches

F-measure was calculated as the harmonic mean of precision and recall:

$$F = 2\frac{(Prec * Recall)}{(Prec + Recall)}$$

Only full matches of maximal NPs were taken into account, i.e. where the left and the right boundary of the NPs are both correctly identified; partial matches were not considered.

If we want to make use of partial matches in a meaningful way, we can consider a partial match as useful if the head of the NP is correctly identified. To give an approximative indication about useful partial matches, we estimated that the head is always to the last word of the NP (which corresponds to our hypothesis used when constructing our chunker). As the Treebank does not specify concretely the head of the NP in its annotation, we had to resort to this approximative solution. Hence, we calculated the precision and recall values for the tags E-NP (last element of a multiword NP) and 1-NP (last and only element of a one-word NP).

Precision	69.20%
Recall	82.00%
F-measure	75.00%

Table II.2 : Evaluation on Szeged Treebank. Correctly recognized NP heads

When interpreting these results, we have to bear in mind that the Szeged Treebank does not use the same notion of NP as the one we defined when constructing our grammars.

A number of systematic — and intentional — differences merge from the errors when evaluating our chunker on the Szeged Treebank, concerning:

- the processing of dates (systematically considered as NPs in the gold standard, but not in our system),
- the analysis of possessive structures (a possessive structure with the possessor immediately preceding the possessee is considered as one top-level NP in our chunker, while it is annotated as two separate NPs in the Treebank),
- bare adjectives or participles are annotated as NPs in the standard, while our system requires NPs to have a nominal or pronominal head (however, depending on the morphological analyzer and the POS tagger, such 'predicative' adjectives often end up having a nominal category[7]).

The conceptual divergences above are due to intentional and, we believe, linguistically motivated choices which we do not consider as errors. On the other hand, the error analysis brought into light real sources of errors in the output of our chunker, the most important one being the inadequate treatment of numerals. A significant amount of unrecognized tokens with the label 1-NP, i.e. NPs consisting of one token, are digits (either bare or with a case suffix) or numerals written with letters. They are not treated as NPs in our system although, especially from a syntactic /distributional point of view, they can in some cases behave like NPs. They can also be positioned as head inside an extended NP, though their NP-internal distribution is different from that of nouns. Their correct disambiguation with respect to syntactic function and distribution requires further refinement of our grammars.

The other important direction in further development is to increase the flexibility of the chunker with respect to lexicalization. Currently, our chunker mostly operates on POS categories as input. As we have seen, lexicalization is taken into account to some extent when describing function words, postpositions, adverbs and pronouns inside NPs. The specific distributional properties of these elements are captured by syntactic features in the dictionary. However, in the light of the error analysis, we concluded that elements

[7]None of the available morphological analyzers can be considered as totally coherent with respect to the POS tag of adjectives that can behave as nouns.

inside these groups may require specific grammars, since the precise distribution of these elements may vary from word to word.

II.2.6.1 Comparison with Hunchunk

Hunchunk, the stochastic NP chunker of Recski and Varga (2009); Recski et al. (2010), was developed at SZTAKI Artificial Intelligence Research Group several years after the first version of our system was presented (Gábor, 2004). The statistical model of HunChunk is a Hidden Markov Model with emission probabilities provided by a Maximum Entropy model. Recski uses the following features: the lemma and each individual morphological feature of the word and the surrounding words in a 5-token long window, the sequences of POS tags before and after the word in a 3-token window, as well as every possible subsequence of tags of length l where l<3. The chunker was trained on 1.000.000 tokens coming from randomly extracted sentences from the Szeged Treebank. The remaining 500.000 tokens were used for evaluation. Recski (2009) reports a precision of 89.4%, a recall of 89.97% and correspondingly, an F-measure of 89.68% on the same task and the same test corpus. However, the direct comparison between the two systems is complicated by two factors. First, Hunchunk being trained on the Treebank, it has the same notion of NP as the evaluation standard, while ours is not conceived according to the same guidelines. Second, Hunchunk has been trained and evaluated on random sentences extracted from Treebank; thus, the training corpus and the test corpus contains sentences coming from the same texts. Even if there is no direct overlap between the training corpus and the test corpus, it is possible that the chunker may overfit the Szeged Treebank and would not produce as precise results on a different test corpus as it does on the Treebank. It would be desirable, therefore, to compare the two systems either on a different gold standard, or by integrating them into the same NLP application.

II.2.6.2 Portability and Use in Applications

Other important aspects of evaluation include portability and adaptability to different tasks. The chunker presented here was used in three major lines of research:

1. Creation of a shallow-parsed corpus for Hungarian,

2. Automatic annotation of prosodic elements, in the project Realization of Near-natural Human-Machine Speech Interaction in Information Systems (Tamm et al., 2008)

3. Different studies on psychological content analysis (emotion detection (Pólya and Gábor, 2010), automatized narrative psychological studies of social identity (Vincze et al., 2010; Liu and László, 2007; Ehmann and Garami, 2010).

Project 1) above has been realized by the author of the present thesis and will be discussed in the following section. Some of the other applications were realized in direct cooperation with the responsible project members, while others were or are being realized independently by researchers outside the RIL HAS.

NP chunking is not the only function of our shallow parser, while the shallow parser itself is only a module of the continuously growing amount of data and grammars available for Hungarian in NooJ. We would like to emphasize that the grammars constructed for parsing Hungarian have been used with success for over five years in diverse applications and thus have proven flexible enough to be adapted to different text types and different annotation tasks. Moreover, the Hungarian NooJ community is constantly growing and researchers outside area of linguistics and the natural language processing are benefitting from both the advantages of NooJ as a piece of software and from the Hungarian data and grammars made available to them. These results are due to the fact that NooJ provides an interface which makes it accessible and easy to learn for non-NLP specialists, but also to the resources provided with NooJ for Hungarian (morphology and syntax). As recent works by Ehmann and Garami (2010) have shown, resources created outside NooJ, as well as the output of external language processing systems, have been successfully integrated into the Hungarian Nooj module and used in a pipeline together with NooJ-internal linguistic modules.

II.3 Building a Syntactically Annotated Corpus for Hungarian

II.3.1 Motivation

This chapter describes the construction of a syntactically annotated corpus, created automatically using the shallow parser described in the previous chapter. The motivation behind constructing a syntactically analyzed corpus with the shallow parser developed in NooJ to was twofold. First, the process of annotating a corpus can be conceived as the validation of the usability of our shallow parser. Second, the resulting corpus can be used for machine learning experiments, especially in the domain of verbal argument structure studies at the syntax-semantics interface, as it will be described in details in chapters VI and VII.

Syntactically annotated corpora are of crucial importance for computational linguistics research and for the development of NLP tools. First, they serve as a gold standard for evaluating applications. They can also be used for testing linguistic theories, and they constitute a valuable source of data when building rule-based NLP systems. Moreover, they can be used as training and test corpora for machine learning systems. The past few years have seen a significant growth in the popularity of machine learning methods in several branches of computational linguistics: syntactic rules are acquired automatically from corpora, lexical resources are based on or validated against corpus data, and are more and more likely to contain frequency or probability information. Moreover, data-driven methods function as a filter that erases the inconsistencies practically unavoidable in human-made linguistic resources. However, languages differ with respect to the availability of annotated resources: while many European languages possess POS-tagged corpora, only the most frequently used ones have robust, validated parsers which can be used for creating large, syntactically analyzed corpora.

As of today, the following corpora are available for Hungarian:

- a morphologically analyzed and POS-tagged corpus: the Hungarian National Corpus, (Váradi, 2002),

- a manually annotated treebank: the Szeged Treebank(Csendes et al., 2004),
- and a raw corpus of texts collected from the world wide web, the Webcorpus (Halácsy et al., 2004).

The Szeged Treebank consists of 1.2 M words; therefore, it is not suitable to extract reliable information about less frequent words or structures: the need arises to automatically annotate bigger corpora with syntactic information. As of today, there is no available validated large-scale parser with proven robustness for Hungarian. Thus, the corpus creation experiment will constitute a robustness test and a validation process for both the annotation tool itself and its Hungarian module. At present the following syntactic analyzers exist for Hungarian:

- a set of rewriting rules automatically learned from the Szeged Treebank (Hócza, 2006),
- a syntactic analyzer based on handwritten grammatical rules (Babarczy et al., 2005),
- the huchunk chunker, based on a maximum entropy model and trained on Szeged Treebank (Recski et al., 2010)
- the shallow parser presented in the previous section.

The idea is to adapt a rule-based system, namely the Hungarian module of NooJ, to create a syntactically analyzed corpus which could be used for machine learning experiments. The aim of the project is to create a 10M words corpus of Hungarian, in order to be able to run machine learning experiments on its data. The conceptual goal is to produce a reusable corpus which represents general vocabulary, but at the same time can be decomposed to thematically different subcorpora which represent different domains. The 10.000.000 words corpus will be annotated entirely with NooJ.

II.3.2 Constitution of the Corpus. Text Sources

The aim of the project is to create a 10M words corpus of Hungarian. We opted for a reusable corpus with general vocabulary and with a possible structural decomposition into subcorpora representing different thematic domains.The corpus texts comprise 10.85M

words and come from the Hungarian National Corpus. The final structure of the corpus is made up as follows:

TEXT TYPE	WORD COUNT
press	4.5 million
scientific	2.2 million
official	2.08 million
literature	2.07 million
ALL	10.85 million

Table II.3 : Composition of the Corpus

II.3.3 POS Tagging

The Hungarian module in NooJ includes a morphological analyzer (Vajda et al., 2004), based on the 60.000 lemmata found in the Concise Dictionary of the Hungarian Language. The analyzer recognizes all the inflected forms of these words, according to the inflectional tables of Elekfi (1997). However, as studies showed later (Nagy et al., 2007), the inclusion of an external analyzer in the process improves coverage and speed, mostly because of the large number of productive derivation possibilities present in Hungarian morphology, which were not included in the inflectional dictionary. Hence, the morphological analysis of the corpus vocabulary was imported from the annotation of the Hungarian National Corpus tagged by the HNC POS tagger (Oravecz and Dienes, 2002). However, function words required a different treatment in that they were re-analyzed by our NooJ-internal function word dictionary which contains all the necessary information to comply with the syntactic rules included in NooJ.

II.3.4 Parsing. Adaptation of the Grammars

When implementing a parser for Hungarian, we have to face the fact that Hungarian language presents a challenge for two reasons. First, because Hungarian is a non-configurational language in that it encodes constituents' grammatical functions not in their respective order but at the morphological level. Verbal arguments appear in an order which is independent of their complement functions. Another problematic issue for

shallow parsing is that not only the respective order of verbs and complements is free, but there may also be optional adjuncts between them (such as sentence modifiers, adverbs of time, location or manner, or even nested clauses), which are left unanalyzed by previous grammars. These issues were discussed in the previous chapter from a theoretical point of view. On the other hand, when analyzing a corpus, grammars developed for dealing with non-adjoining constituents may affect performance. However, as opposed to the sentence level, NPs and other nominal constituents present a strict internal word order with the case-marked head being the rightmost element. This implies that local grammars designed for recognizing NPs, adjectival phrases and other constituents with a strict word order can perform fairly, while we need to have recourse to NooJ's special functionalities for exploring inter-constituent dependencies at the sentence level.

At this point we need to recapitulate that the annotated corpus is meant to be used for machine learning purposes, especially in the domain of learning verbal argument structure and acquisition of distributional/semantic verb classes. Special attention was paid to the correct identification of verbal predicates. The goal was to produce and export an output which contains the correctly recognized verbal predicates at the level of clauses, and the labeled phrases which represent potential complements or adjuncts of the predicate, annotated with the features relevant for the identification of the argument structure. Ten cascaded grammars (simple or composed of several sub-grammars) were used in a strict order to perform the annotation.

NooJ grammars used in corpus annotation (in order of application):

1. Sentence segmentation

   This rule produces <S> tags, exported in the output.

2. Frequent ambiguities

   A set of grammars for the disambiguation of a few instances of word-level ambiguities.

3. NP Grammar

   Recognizes and annotates NPs, at the same time disambiguating past participles inside NPs (homonymous with verbal past tenses).

4. Postpositional phrases

Application of three subgrammars, corresponding to three different types of postpositions, as described previously.

5. NP Coordination

   The grammar unifies coordinated NPs under a common NP tag. The head of the coordinated NP is not specified, but the features of the two NP heads are percolated to the upper level NP tag.

6. Predicate

   Annotates finite verbs as predicates, recognizes complex verb forms and detached prefixes. Relevant verb features like morphological features, definiteness and the lemma of the verb are copied into the <PRED> tag.

7. Clause boundary detection

   The heuristic clause boundary grammars are unified under one graph. The guesser rule, however, was not applied in NooJ because of performance issues (examining every position between two finite verbs turned to be a too expensive operation); it was added as a postprocessing module.

8. Auxiliaries and infinitives

   Two subgraphs are distinguished, impersonal constructions are dealt with in a separate grammar and receive an additionafeature; the rest of the grammar corresponds to the content of the previous chapter.

9. Content verbs with infinitival complements

   When the infinitive appears with a finite verb which is not listed as an auxiliary, the latter will be considered as predicate, and the infinitive is annotated as a potential complement.

The result of the shallow parsing is exported in XML. A postprocessing phase was added to complete the xml annotation by the header and by a unique identifier for each sentence.

Chapter III

Lexical Syntax and Semantics of Verbs. Theory.

"Fully to define the association between a particular subcategorization structure and a given predicate, however, one must go beyond listing of syntactic frames. Full account of subcategorization requires specifying the number and type of arguments that a particular predicate requires, predicate sense in question, semantic representation of the particular predicate-argument structure, mapping between the syntactic and semantic levels of representation, semantic selectional restrictions or preferences on arguments, control of understood arguments in predicative complements, diathesis alternations, and possibly also further details of predicate-argument structure."

(A. Korhonen, 2002)

III.1 Subcategorization and Lexical Semantics

III.1.1 Introduction

Chapters III deals with theoretical issues about lexical syntax and semantics of verbs. Our research aims at building lexical resources for the computational processing of Hungarian in order to recognize dependency relations between sentence constituents. The focus of the work described in this chapter and the following ones is verbal complement structure.

In section II., we have presented a rule-based constituent chunker and shallow parser

for Hungarian. This parser was conceived in a way that it produces the input of further processing, e.g. by a dependency parser. For instance, see the sentence below:

Iraki erők megrohamozták az iráni menekültek táborát.
Iraqi forces attacked the Iranian refugee camp.

The constituent chunker produces the following labelled bracketing for this sentence:

<NP case="nom" number="pl">Iraki erők</NP> <PRED>megrohamozták</PRED>
<NP case="acc" number="sg">az iráni menekültek táborát</NP>.

As opposed to shallow parsing, a dependency parser or deep parser has to 1) recognize and annotate dependency relations between the verb and its syntactic dependents (both complements and adjuncts), 2) identify the verbal subcategorization frame. We would like to assign to the sentence above an annotation structure which provides at least the following information:

PRED lemma="megrohamoz" SUBJECT=NP+head="erők"
PRED lemma="megrohamoz" OBJECT=NP+head="táborát"
subcategorization.frame=megrohamoz SUBJECT OBJECT

The major difference between constituent chunking and dependency parsing in Hungarian lies in the different amount of lexical information needed for each of these tasks. Although we have seen in part II that some subtasks in constituent chunking require lexical information to be taken into account, i.e. the internal structure of sentence constituents is influenced by lexical properties of words, this is a more marginal phenomenon. This means that we can see POS categories as elementary constituents of the basic structures. E.g., the elementary constituent of a NP is the head noun, which can be extended by determiners, quantifiers, adjectival modifiers, possessive structures etc. The most complex phenomenon is the modification of an NP by a participle, which in turn can be modified by a large set of complements inherited from the base verb. From a method-

ological point of view, the construction of an NP grammar can start from the description of basic structures – those which can be described without taking lexical information into account. Later, in parallel with the testing of the grammar, it can be extended so that more sophisticated subcategories will be established for more specific phenomena. We can continue with more and more specific structures until a desired coverage of the grammar is achieved.

On the other hand, lexical conditioning is necessary from the beginning when we describe and parse verbal argument structures. The number of arguments a verb can take, their syntactic realization, the syntactic alternations they participate in, as well as their semantic roles are specific to the given lexical unit. Verb-complement dependency relations are to be coded in the lexicon previously to the construction of dependency grammars. The role of a lexicon is to encode unpredictable linguistic phenomena, including lexical subcategorization properties of words. Since such properties cannot be derived from general principles of the syntax of a language, they have to be stored and referred to during parsing. Lexicalized parsers have been widely used (Schabes and Waters, 1993; Abeillée and Candito, 2000), and proved to be more accurate and efficient in solving problems such as attachment ambiguities. Such parsers use lexical databases of syntactic (and semantic) dependency relations.

Lexical databases have to satisfy a certain number of requirements in order to be usable for parsing or other NLP purposes. First, they have to be *explicit*: they cannot rely on the user's linguistic intuition (as it is often the case with paper dictionaries). Second, they have to be *coherent*: similar constructions have to be coded similarly. In order to meet these requirements, a set of objective and wide-coverage linguistic tests is needed to ensure that the intuition of human coders does not interfere with the coherence of the database. Most importantly, when constructing a verbal subcategorization lexicon for a new language, one needs to provide a consistent notion of verbal complement and one or more corresponding tests which can be used to differentiate complements from adjuncts. The following chapter deals with theoretical and practical issues related to the notion of complement and verbal subcategorization in a non-configurational language. Precisely, we will argue that due to morphosyntactic specificities of Hungarian as a non-configurational language:

- existing complement tests, either Hungarian-specific or language independent, cannot be used satisfactorily on a wide range of language data,
- the reason for this is that there are no systematic differences between the syntactic behavior of complements and adjuncts in the surface syntax of Hungarian,
- neither the obligatory character of complements, nor the productivity of adjuncts cannot be confirmed empirically,
- for these reasons, speakers of Hungarian do not have a strong intuition about particular cases of the complement-adjunct dichotomy.

In a Hungarian sentence, the context of a verbal predicate is made up of a set of case-marked nominal constituents (eventually, postpositional phrases). At the semantic level, some of these constituents enter in a predicate-argument relation with the verb, others modify the complete predicate-argument structure externally. A predicate-argument relation is characterized by a mutual semantic interdependence, whereas external modification implies semantic autonomy. However, the syntactic realization of arguments and adjuncts is similar: they occupy the same syntactic positions (at least in the directly accessible surface structure) and are marked morphosyntactically by case suffixes or postpositions. These morphological markers are ambiguous in that the same case suffix or postposition can encode several different grammatical functions, including both complement and adjunct functions. The semantic distinction, therefore, does not go together with any morphosyntactic clue. Since the semantic level cannot be accessed directly, the question raises how is it possible distinguish arguments from adjuncts?

In what follows, we are going to see that there are incoherence problems with the argument definitions and tests across linguistic theories. These problems are partially masked by the fact that most complements are syntactically obligatory in well-studied and well-resourced european languages. Therefore, the speakers of these languages have a strong intuition about whether a given constituent is a complement or not. On the other hand, several recent studies pointed out that the obligatory character of complements is insufficient and suggested that the distinction between arguments and adjuncts could be

captured quantitatively, based on corpus data (Fabre and Bourigault, 2008; Merlo and Esteve Ferrer, 2006). We subscribe to the point of view rejecting a strict dichotomy between complements and adjuncts and we propose to proceed by a parallel syntactic and semantic role labeling. It has been known for a long time that syntactic complement structure correlates with lexical semantic properties of verbs. We extend this hypothesis and apply it to a non-configurational language: we argue that knowing the lexical semantic properties of verbs is *necessary* in order to predict which case-marked adjunct NPs can appear in the context of a verb yielding a grammatical sentence. The same case suffix can encode different relations depending on the lexical semantic class of the verb they appear with. Therefore, knowing the lexical semantic class of a verb is a prerequisite to telling which argument structure(s) it takes. Our methodology defines lexical semantic classes on the basis of distributional properties. Besides addressing the important theoretical concern about the consistency of argument-adjunct definitions, relying on such a classification has the advantage of increasing the amount of predictable information and therefore allowing a more compact lexical representation.

After presenting the state of the art in argument realization studies (3.1) and highlighting the problematic issues in common syntax-based and semantics-based argument definitions (3.2), a suggested model of lexical representation will be presented (3.3). This model takes into account the specificities of Hungarian syntax. The applicability of our approach will be discussed in chapters IV-VI., which deal with methodological and implementation issues. The content of the following chapters is to be conceived as a workflow aiming to put into practice the model presented hereafter. We will describe a hand-constructed verbal subcategorization lexicon and two methods for enhancing it with relevant lexical semantic information: a method for manually constructing verb classes, and an automatized lexical acquisition method applied to corpora.

III.1.2 Lexical Semantics and the Status of Arguments

(The ideas presented in sections 3.1 – 3.3. result from a tight collaboration with Enikő Héja and contain overlap with (Gábor and Héja, 2006))

The following section presents how different language theories deal with the argument-adjunct distinction. Semantic and syntactic definitions of arguments will be described, along with the hypotheses and presuppositions formulated or implied within the different theories. The coherence of these definitions and the usability of the corresponding argument tests will be discussed. Subsequently, an alternative approach will be described, using compositionality as a starting point and relying on Levin-type verb classes.

III.1.2.1 Mapping Theories and the Semantic Basis Hypothesis

A common presumption within current (lexicalist) theories is that lexical units constitute the input of syntactic rules. This implies that lexicons have to encode at least the pieces of idiosyncratic information that cannot be inferred from general syntactic rules. For instance, a verbal lexical entry has to contain a list of the semantic arguments the verb takes and the description of the syntactic realization of these arguments – if it is not predictable. Therefore, the argument structure of a predicate can be conceived as a lexical property of the verb corresponding to the predicate. This approach determines how lexicalist theories view grammatical functions: since complements are considered as the syntactic realization of semantic arguments, the linguistic description of complements is seen as a priority. On the other hand, adjuncts are characterized as optional and productive elements both at the semantic and at the syntactic level. Their description is referred to the syntactic module of the grammar.

Nevertheless, the syntactic properties of elements in verbal subcategorization frames are predictable to a certain extent from the semantic properties of the verb in question. The Distributional Hypothesis of Harris (1954) states that words that occur in the same or similar contexts tend to have similar meanings. The causal link has been pronounced as the Semantic Basis Hypothesis (SBH) (Koenig and Davis, 2003). The weak form of the SBH states that there is a *correlation* between the semantic content of a word and its distributional contexts: the degree of semantic similarity between two linguistic expres-

sions is a function of the similarity of the linguistic contexts in which the expressions can appear (Zarcone and Lenci, 2008). The strong form of the hypothesis implies a certain amount of *predictability*: knowing the semantic content of a word helps us to predict its distributional behavior to a certain amount, and there is a causal link between lexical semantics and syntax. Applying this hypothesis to verbs leads to the statement that semantically similar verbs tend to have similar subcategorization frames. The assumption is that verbal semantics determine argument structure and the surface realization of arguments. As formulated by Koenig and Davis (2006):

SEMANTIC BASIS HYPOTHESIS: *If you know a language and you are presented with a new verb, you can predict a fair amount of its subcategorization (possibly all of it, if the verb is not syntactically irregular).*

This line of research reached linguistic theories and gave rise to a big amount of work aiming to provide a semantically motivated account of verbal subcategorization. Lexicalist theories such as LFG (Kaplan and Bresnan, 1982) and HPSG (Pollard and Sag, 1994) as well as chomskyan language models tried to explore the link between morphosyntax and semantics by tracing back surface syntactic dependents to lexical semantic representation of verbs. Mapping theories (or linking theories) try to derive surface syntactic structure (argument realization) from a lexical semantic representation of verbal arguments. The basic presumption is that semantic arguments are realized as syntactic complements, and there is a set of lexical-syntactic rules which perform the mapping between the two structures. The semantic representation which serves as a basis for formulating generalizations over the behavior of arguments is often formalized by thematic roles. Since thematic roles are semantically motivated concepts and as such, they are prone to subjectivity, a certain number of constraints have been created in order to try to avoid ad hoc representations, mistreatment or subjective interpretation of the roles. Linguists suppose that there exists a finite set of theta roles, all universal across languages. The theta criterion Chomsky (1981) states that every argument in a sentence must be assigned one and only one theta role, and each theta role must be assigned to one and only one argument. The Uniformity of Theta Assignment Hypothesis(Baker, 1988) ensures that identical thematic relation-

ships between items are represented by identical structural relationships between these items, i.e. each thematic role is realized in a consistent way among verbs. The Transparency Principle (Lightfoot, 1979), adapted to verbal argument realization (Koenig and Davis, 2003) states that constraints on argument realization must be captured in terms of semantically natural classes of properties. Most linking theories accept these principles and build on thematic roles in their lexical semantic representation to predict the surface realization of arguments. However, as to the present, the prediction of many linguistic phenomena seem to fall beyond the scope of thematic roles. In parallel, hierarchical lexicon models came to existence and flourish, mostly within the framework of lexicalist language theories (Jackendoff, 1990; Levin, 1993; Levin and Hovav, 2005; Pustejovsky, 1995; Copestake, 1992; Bresnan and Zaenen, 1990; Koenig and Davis, 2006; Goldberg, 1995). In what follows, we are going to point out the role and mechanism of mapping within the formalism of two lexicalist models, and describe the ontological/grammatical presumptions shared across different mapping theories.

III.1.2.2 Mapping in Lexicalist Theories

In GPSG (Gazdar et al., 1985), as well as in early HPSG (Pollard and Sag, 1994), (verbal) subcategorization was represented as list of syntactic dependents ('SUBCAT' feature), the value of which contained only syntactic information. Therefore, the formalism did not provide any tool to derive surface syntactic valence from the underlying (semantic) representation of argument structure.

Since HPSG II, syntactic and semantic features have been integrated under the SYNSEM feature; this allowed to capture the constraints the head puts on syntactic and semantic properties of its arguments. Since the value of the SYNSEM feature contains both syntactic and semantic information, the value of SUBCAT can now be specified more appropriately as a list of syntactic and semantic properties. Later on, the SUBCAT feature has been subdivided into SUBJ (subject) and COMPS (other complement) features (Borsley, 1989) mostly in order to allow a proper treatment of complement-taking prepositions which do not have subjects. As Manning and Sag (1999) note, the SUBCAT feature continued to exist as a separate, but not independent feature: it was limited to summarize the valence list of the verb. When authors started to formalize and explain different

phenomena involving a dissociation between argument structure and valency frames (e.g. phonologically empty complements in pro-drop languages), the SUBCAT feature started a life on its own. Newer literature refer to SUBCAT as ARG-ST, conceived as an abstract, underlying representation of the argument list, as opposed to SUBJ and COMPS features which are linked to surface syntax. "The ability to dissociate argument structure from valence (...) takes HPSG a certain distance from the monolevel, monostratal roots of GPSG and early HPSG." (Manning and Sag, 1999). The third level of representation is the CONTENT feature, which specifies the semantic relations between the verb and its arguments. Linking theory in HPSG is based on a typology of semantic relations and a parallel hierarchy of predicate types. Following Koenig and Davis (2003), we can state that inheritance hierarchies are present at both levels. First, at the level of semantic relations (CONTENT feature), the very specific semantic role of arguments (e.g. "runner" for the subject argument of run) belong to and inherit attributes from more general types of relations. These types are often, but not necessarily conceived in terms of thematic roles. Second, there is an inheritance hierarchy of predicate types, defined by how they link attribute values within their CONTENT to members of the ARG-ST list. "Such a semantic hierarchy, which encodes the (linguistically relevant) relations between categories of situations, helps restrict the grammatical constraints on the realization of semantic arguments to the proper semantically-defined class of verbs." (Koenig and Davis, 2003)
A meta-theoretical constraint ensures that the hierarchy of semantic relations and the hierarchy of predicator types correspond to each other: if the semantic type i is a subtype of another semantic type j, then the verb class with semantic type i will be a subtype of the class with semantic type j (with respect to their linking properties). This parallel type hierarchy constitutes the formal toolkit for describing argument realization in terms of semantically motivated linking constraints and rules in HPSG.

Subcategorization and mapping in Lexical Functional Grammar(Kaplan and Bresnan, 1982) highly relies on the notion of grammatical function. Functional information is present at three different levels of representation, linked by a mapping function:

- a-structure encodes lexical information including argument structure,
- c-structure encodes constituent structure (it corresponds to surface syntax and un-

like transformational grammars, does not allow transformations, moving etc.),

- f-structure represents functional structure itself, which is composed of universal syntactic primitives (functions).

The f-structure therefore provides an abstract level of representation of grammatical information, which does not constrain surface representation; language-specific surface syntax is represented at the level of c-structure.

Argument structure is lexically encoded in the a-structure. By argument structure we mean semantic arguments, i. e. unfilled slots in the meaning of the verb. Argument roles are specified in terms of thematic roles. The first step of mapping between argument structure and surface realization takes place in the lexicon by the annotation of lexical structure or grammatical function assignment, i. e. grammatical functions are associated to argument roles. Similarly to the Uniformity of Theta Assignment Hypothesis referred to above, a constraint has been formulated in LFG by Bresnan (1982)) to ensure the strictly one-to-one relation between semantic arguments and grammatical functions within the argument structure of a lexical form. An important and distinctive principle of LFG states that subcategorization can only refer to grammatical functions of complements instead of their syntactic realization (e.g., POS categories). The monotonicity or direct syntactic encoding principle ensures that only lexical rules can change the mapping between argument structure and grammatical functions - this entails that syntactic alternations and other processes affecting argument realization (e.g. passivation) can only be dealt with at the lexical level. The surface realization of constituents with a complement function is produced by context free rules of the c-structure. Since languages differ with respect to how they realize grammatical functions, this has to be specified for each complement function in a language.

The Lexical Mapping Theory (Bresnan and Zaenen, 1990), developed within LFG, provides a way to formulate generalizations and regularities about the mapping between grammatical functions and arguments, instead of stipulating redundant lexical rules. This theory relies on the universal hierarchy of thematic roles (Grimshaw, 1990) and on the idea of decomposing syntactic functions with two binary features: +/- thematically restricted (r) and +/- objective (o). These features define four separate complement classes: subject (-r -o), object (-r +o), obliqueΘ (+r -o) and objectΘ (+r +o), where Θ stands

for specific thematic roles. A markedness hierarchy of syntactic functions can be established using the features above, the subject being the least marked and thematically restricted oblique complements being the most marked functions. Arguments are mapped onto these functions following their thematic hierarchy, under the condition that they are compatible with respect to the +/-o and +/-r features. These features are assigned to thematic roles according to a set of lexical mapping principles. Some of the thematic roles intrinsically and universally bear some features which affect their realization (i. e. the syntactic function assignment), e.g. the agent is always -objective. Other features are added by language-specific morpholexical rules. Finally, there are default mapping principles, e.g. the highest ranked thematic role receives a -r feature. The application of lexical mapping principles respects monotonicity in that they cannot delete or change already assigned features. The bi-uniqueness of argument-function mapping is preserved within the Lexical Mapping Theory.

III.1.2.3 Relevant presumptions shared across mapping theories

Let us now consider the basic assumptions about the status of predicate-complement structures in natural language, underlying the work on mapping theories and argument realization. These claims can be formulated explicitly at three distinct levels, all of them corresponding to a different level of representation. The first one is the ontological level: the presumptions and hypotheses formulated at this level refer to the relation between linguistic notions and entities external to language:

- Verbs typically describe events.
- Verbal arguments refer to relevant participants of the events.
- Adjuncts refer to external coordinates of the events.

At the next level we find semantic presumptions:

- Each verb has an argument structure.

- These arguments can combine with further arguments and hence change the type of the event[1] encoded by the verb.
- The thematic role of the arguments determine their morphosyntactic realization.

Finally, morphosyntactic presumptions:

- Arguments are realized as complements of the verb.
- Adjuncts are attached externally to the syntactic phrase formed by the verb and its complements, by the application of productive syntactic rules.

The next section deals with the implications of the semantic and syntactic presumptions above, with respect to argument definitions and tests.

III.2 Argument Definitions and Tests

III.2.1 Semantic Definitions

At the ontological level, arguments are described as participants of the event denoted by the verb. At the semantic level, arguments are considered as unfilled slots in the verbal meaning, as opposed to adjuncts that express external coordinates of the action denoted by the verb. Finding a straightforward way to formalize what exactly being a semantic argument entails is a complex task. Koenig and Davis (2003) introduce two criteria to distinguish arguments from external coordinates and test them to provide psychological proof for the existence of the argument/adjunct dichotomy. The first criterion relies on the compulsory character of semantic arguments. They formulate the obligatoriness criterion at the ontological/pragmatic level when they presume that the recognition of a verb activates a certain amount of information regarding the participants involved in the action/event described by the verb. This information is presumed to be encoded in the

[1] On a detailed discussion of the notion of *event type*, see 3.2.

mental lexicon of (native) speakers.

Semantic Obligatoriness Criterion: *If r is an argument participant role of predicate P, then any situation that P felicitously describes includes the referent of the filler of r.*

E.g., the verb sleep describes a sleeping event, and its subject refers to the sleeper participant of that event. Since no sleeping event can occur without a sleeper, this role corresponds to an argument participant. However, the authors also admit that the Semantic Obligatoriness Criterion does not ensure an exhaustive identification of semantic arguments: on the ground of this criterion one should categorize location and time adjuncts as arguments, since they compulsorily characterize most of the events. This difficulty is tackled by the introduction of the Semantic Specificity Criterion which states that every argument participant role is specific to a predicate or a restricted class of predicates:

Semantic Specificity Criterion: *If r is an argument participant role of predicate P denoted by verb V, then r is specific to V and a restricted class of verbs/events.*

The idea behind this criterion is that while most events occur at a certain time and a certain location, very few of them include e.g. a sleeper participant. Moreover, they argue that argument participant roles take on additional properties for specific types of events – beside the properties shared by most participants of the same kind. Their argument is based on the distinction introduced by (Dowty, 1991). He refines the notion of thematic role by distinguishing individual thematic roles (specific to certain predicates or classes of predicates) from proto-roles (generalizations over properties of individual roles). For instance, besides the set of properties that characterize agents of all verbs, the agent of a singing event must adduct its vocal folds in any event that sing felicitously describes, while this is not true for the agents of many other verbs. This implies that argument participant roles are lexically required to bear additional properties aside from those which are characteristic of the role. To summarize, a semantic definition of arguments relies on the obligatory character of the argument (where obligatory is understood at the semantic/pragmatic level), and on the specificity of the semantic role which links the argument

to the predicate. As we will see later, there is a certain number of difficulties raised by this definition and its application to language data – most importantly, a discrepancy between semantic definitions and their syntactic counterparts. At this moment we limit ourselves to noting that neither semantic/pragmatic obligation, nor semantic role specificity cannot be tested directly.

III.2.2 Syntactic Definitions

III.2.2.1 GB tests: Structural Position

Attempts were made to give a purely syntactic definition of arguments. In what follows, the argument-adjunct dichotomy and corresponding syntactic complement tests will be presented in a configurational (GB) and in a lexicalist model (LFG).

According to both theories, complements are the syntactic realization of arguments and take different structural positions than adjuncts. In Government and Binding theory (GB) (Chomsky, 1981), complements are compulsory constituents which appear in the close local context of the verb. Being a complement is conceived as a relation: constituents with a specific complement role with respect to a given predicate do not necessarily have the same function in the context of other predicates. Predicates' ability to take complements is their idiosyncratic lexical property. Consequently, lexical entries of verbs have to contain as much information as necessary for the syntactic rules to generate surface form of complements. They will contain a syntactic description (minimally, the syntactic category) of their complements. Moreover, since syntactic complements are the surface representation of semantic arguments, it is worth coding the thematic roles of semantic arguments in the lexicon: a part of the syntactic structure can be derived from thematic roles. In configurational language theories, the structural difference between complements and adjuncts can be summarized by the following rewriting rules:

1. Argument NP: $V' \rightarrow V + NP$
2. Argument NP: $VP \rightarrow V' + NP$
3. Adjunct NP: $V' \rightarrow V' + NP$

4. Adjunct NP: $VP \rightarrow VP + NP$

Complements and adjuncts thus occupy different positions in the parse tree of a sentence. Complements are located in the syntactic tree in a sister node of the V head and together they form a V' projection. Adjuncts are sisters of the V' projection and form a new V' with it. Syntactic complement tests build on this structural distinction. However, there are only language-specific tests for verifying the different structural position of given constituents. (Radford, 1988) shows the following tests for English:

I) Passivation: An NP constituent of a complement prepositional phrase can be passivized, while an NP constituent of an adjunct PP cannot:

[This job] needs to be worked at by an expert.
*[This office] is worked at by a lot of people.

II) Pronominalization: The do so phrase substitutes a V' constituent. The V' can optionally substitute a V' which includes adjuncts (1), but not necessarily (2), while complements are always included in the V' the do so phrase stands for (3), they cannot be explicit (4):

John will [buy the book on Tuesday] and Paul will do so as well.
John will [buy the book] on Tuesday and Paul will do so on Thursday.
John will [put the book on the table] and Paul will do so as well.
*John will [put the book] on the table and Paul will do so on the chair.

III) Surface order: Complements are closer to the verb than adjuncts because they connect to the verb in the syntactic tree earlier than adjuncts, and crossing branches are forbidden.

IV) Ellipse: Any phrasal category can be ellipted. Constituents of the category of V' can be ellipted if they consist of the verbal head with its complements and adjuncts (1), the head with its complements but without adjuncts (2), but the head with one of its

complements and without the other one does not form a constituent, hence it cannot be ellipted (3):

1) - Who might be going to the cinema on Tuesday?
- John might be ... .
2) - Who might be going to the cinema when?
- John might be ... on Tuesday.
3) - Who will put the book where?
- *John will ... on the table.

Although structural positions, including the specific positions attributed to complements and adjuncts are presumed to be universal according to GB, we have to note that the arguments for a distinction between complement and adjunct positions in this framework come from evidence in configurational languages. In languages like Hungarian, where grammatical functions are not configurationally coded, the surface structure does not reveal the distinction. In fact, none of the tests above can be used for Hungarian, not even to an extent limited to prototypical cases. The passivation and the pronominalization tests do not apply to Hungarian due to the lack of such constructions, i.e. there is no passive voice in Hungarian and we cannot find a pronominal element which substitutes verb phrases. Criterion no. 3 regarding surface order is simply not met in Hungarian sentences: it is easy to create counter-examples by switching the surface order of complements and adjuncts:

A szomszéd lenyírta a füvet délben.
The neighbor cut the grass at noon.
A szomszéd lenyírta délben a füvet.
the neighbor cut at noon the grass.

Although the constituent 'at noon' is most likely to be an adjunct, it is encrusted between the verb and its object complement in sentence b). There are two distinct features of Hungarian syntax which may yield sentences similar to that in example b). First, as

described in (é.Kiss 2002), in the neutral sentence verbal complements and adjuncts follow the predicate within the VP. However, practically any of the complements or adjuncts can be topicalized or focused, hence moved outside the VP to a position preceding the verb. Furthermore, verb modifiers, i.e. verb prefixes, adverbs or bare NP complements also precede the verb they modify. Since these operations work on complements as well as adjuncts, some combinations of them may result in a constituent order which contradicts the criterion formulated in 3). Second, complements and adjuncts in a post-verbal position are free to mix with each other, resulting in an arbitrary constituent order: we cannot rely on constituent order to tell apart complements from adjuncts.

The problem with the test no. 4) is that different conditions apply to ellipse in Hungarian. We will demonstrate this by two counter-examples. As a first example, we will use the Hungarian counterpart of the English sentence, the verb tesz (to put), presuming that similarly to its English equivalent, it has three complements (subject, object, locative complement).

IV/1 Ellipse of verb phrase with complements and adjuncts:
- Ki teszi a kulcsot a lábtörlő alá indulás előtt? - János ... .
- Who puts the keys under the doormat before leaving? - John ... .

IV/2 Ellipse of verb phrase with complements, without adjuncts:
- Ki teszi a kulcsot a lábtörlő alá mikor?
- János ... az indulás előtt.
- Who puts the keys under the doormat when?
- John ... before leaving.

IV/3 Ellipse of the head with one of its complements and without the other one:
- Ki teszi a kulcsot hová?
- János ... a lábtörlő alá.
- Who puts the keys where?
- John ... under the doormat.

Sentence 3) shows that a grammatical sentence is obtained when ellipting the verb with one of its two complements. This means that either the locative phrase is not a complement of the Hungarian verb, or the test is not valid for Hungarian. As a second example, we use the verb bánik (to treat), which is a verb with two very prototypical compulsory complements (bánik N.INS ADV = to treat sg/sy somehow).

IV/1.1 Ellipse of verb phrase with complements and adjuncts:

- Ki bánik méltánytalanul Jánossal az üzemben?
- A főnök ... .

- Who treats John unfairly at the factory?
- The boss ... .

IV/2.1 Ellipse of verb phrase with complements, without adjuncts:

- ? Ki bánik méltánytalanul Jánossal hol?
- ? A főnök ... az üzemben.
- Who treats John unfairly where?
- The boss ... at the factory.

IV/3.1 Ellipse of the head with one of its complements and without the other one:

a) interpreted as an echo question:

- Ki bánik Jánossal hogyan?
- A főnök ... méltánytalanul.
- Who treats John how?
- The boss ... unfairly.

b) interpreted as distributive quantification:

- Ki hogyan bánik Jánossal?
- A főnök ... méltánytalanul, a kollégák ... kedvesen, a személyzetis ... előzékenyen.
- Who treats John how?
- The boss ... unfairly, the colleges ... nicely, the HR manager ... politely.

Sentences 2) and 3), including the questions, have a dubious grammaticality status. However, they can be transformed to have a distributive/contrastive interpretation: for each potential subject [how s/he treats John]? for IV/3 a), or [where does s/he treat John unfairly] for IV/3 b). This interpretation allows to construct answer sentences ellipting the verb and one of its complements ('bánik Jánossal' – 'treats John'), butt keeping the other one ('méltánytalanul' – 'unfairly').

One can conclude from the examples above that neither surface sentence structure nor the syntactic tests building on this structure can be used to evidence the supposed syntactic difference between complements and adjuncts in Hungarian. The surface structure does not translate directly to the parse tree of the sentence (since this model allows covert movement), thus, we cannot reject that complements and adjuncts occupy different structural positions, but this cannot be demonstrated in Hungarian.

The tests above are admittedly language-specific, but no equivalent structural tests have been proposed specifically for Hungarian data. However, É. Kiss (2002) considers that complements occupy sister nodes to the verb, while adjuncts are attached to it with recursive rewriting rules. This structural difference implies that complements can change the syntactic distribution of the verb (since a new category is created by the addition of a complement), while adjuncts cannot change the distribution of the constituent they combine with (the recursion implies that the syntactic contexts of the VP are the same as the syntactic contexts of the VP + adjunct structures).

III.2.2.2 LFG Test 1: Obligatory vs Optional

In the LFG model, different structural levels of the sentence carry the same functional information. However, information about grammatical function is present at every level of representation. Komlósy (1992), in his LFG analysis of Hungarian, abides by the notion of grammatical function and considers complement and adjunct at a strictly syntactic level. Though he admits that by this time no widely accepted and applicable criteria exist, he tries to enumerate some criteria on the grounds of which at least in some cases one could decide about the exact nature of a given constituent. Being syntactic, the criteria cannot refer to the meaning of the verb at all, only formal aspects of the verbal behavior (its dis-

tribution) are taken into account. The underlying semantic analysis relies on the notion of unfilled slots ? — argument positions — in lexical entries of predicates. These slots can be filled up with constituents which refer to participants of the event expressed by the predicate. These constituents are syntactically realized as complements. Argument realization is guided by thematic roles. Therefore, every argument has to be assigned a thematic role. Komlósy emphasizes that argument realization and the syntactic functions need to be looked at as purely syntactic notions, and the only adequate way to distinguish between them is via distributional tests. He argues that a full coverage, language-specific criterion system would necessitate the knowledge of the complete grammar. At the current state of linguistic research, we have to limit ourselves to enumerate a few criteria which do not provide full coverage but — he claims — work reliably where they apply to tell apart complements from adjuncts.

> 1) *If a constituent is obligatory in any level of the sentence structure, it is a complement (where obligation implies that omitting the constituent yields an elliptic structure or a different meaning).*

This criterion is widely used across languages to test whether a given constituent is a complement (e.g. (Somers, 1984)). In Hungarian, however, many complements can be omitted even in sentences where the conditions for ellipse are not met. For instance, the verbs eszik (eat), épít (build), magyaráz (explain), megbocsájt (forgive) can be used intransitively, without an antecedent which would co-refer with an ellipted object. Evidently, one can say that there are more than one lexical element corresponding to these verbs, one of them being transitive, the other one being intransitive: this is what Komlósy (1992) claims as well[2]. To support his argument, he explains that the intransitive variants of these verbs semantically imply their omitted object to be understood as an existential quantification over a restricted semantic class. For instance, the sentence "Jóska eszik" ('Joe eats') implies that Jóska eats something edible, while the transitive verb does not have this restriction (e.g. "Jóska homokot eszik", 'Joe eats sand'). The different semantic

[2]Though his concept of lexicon allows to generate the intransitive entries from the transitive ones by systematic lexical rules.

selection constraints indicate, according to Komlósy, that we are dealing with two distinct lexical entries.Kálmán (2006) contests this analysis: he argues that the variance in semantic selection is too widespread to be considered as a lexical property and attributes the specific semantic constraints to the intransitive structure itself rather than to the specific verbs that can enter in this construction. We concur with Kálmán's analysis and do not consider the intransitive variants as separate lexical entries for two reasons 1) the phenomenon is general and predictable for a big amount of Hungarian verbs 2) the difference between the meaning of the transitive structures and that of the intransitive forms is not proven to be more important than the difference we can find between the meaning of any verb used with two semantically different direct objects.

Another problem with the obligatoriness criterion is that it disguises a more important difference: the difference between compositional and non-compositional constituents. While "lakik valahol" ('live somewhere') and "törődik valamivel" ('care about sg') both come with an obligatory NP complement, the first one is a compositional structure – the obligatory NP might be substituted with any adverb denoting a location. On the other hand, "törődik valamivel" is non-compositional and hence more prototypical as an argument – in semantic terms, it fulfills Koenig's Semantic Specificity Criterion. This cannot be said about the NP in "lakik valahol". In other words, upon the application of this principle, structures as "lakik valahol" and "törődik valamivel" end up in the same category, while the locative NPs in "lakik valahol" and "alszik valahol" ('sleep somewhere') would not have the same status. The strictly formal definition is therefore in conflict with the semantic presumption about complements being semantic arguments of the predicate. This discrepancy is hardly defendable on semantic grounds. While the semantic argument will be represented as the argument of the verbal meaning, the adjunct will take the whole verb + arguments structure as its own argument, which makes it very hard to capture the similarities between the two structures. It is clearly a matter of choice which principle is taken as more cardinal: compositionality or the obligatory character of the complement.

Nevertheless, if every obligatory complement is considered as the syntactic realization of an argument, they have to be assigned a meaningful theta role, in accordance with their participant status. As Dowty (1991) notes:

"In order for such systems to work, the meanings of all natural-language predicates must turn out to be of a very particular sort: for every verb in the language, what the verb semantically entails about each of its arguments must permit us to assign the argument, clearly and definitively, to some official thematic role or other. It cannot be permitted to hover over two roles, or to 'fall in the cracks' between roles and what the meaning entails about every argument must always be distinct enough that two argument clearly do not fall under the same role definition. This is a very strong empirical claim about natural-language predicates." (p. 549)

Assigning a thematic role to complements is challenging in some cases and may result in thematic role types which are not semantically meaningful. This is especially problematic for obligatory complements other than NPs. For instance, the verb bánik ('treat'), used in the examples IV/1-3 above, has an obligatory object and an obligatory adverb as complements:

János toleránsan bánik a tapasztalanlanabb kollégákkal.
John tolerantly treats the less experienced colleges.
John is patient with the less experienced colleges.

What kind of semantically motivated thematic role could be assigned to the adverbial complement of bánik in the example above? In such cases, as Koenig and Davis (2003) note, linguists often resort to constructs "that are syntactic diacritics": they do not correspond to semantically natural classes, they are only posited in order to be able to formulate constraints in mapping theory.

III.2.2.3 LFG Test 2: Event Type Shift

The following complement test builds on the assumption that the semantic difference between arguments and adjuncts is reflected in syntactic distribution. Komlósy presumes that adjuncts refer to external coordinates of events. Events are expressed by full propositions; adjuncts are operators that take propositions as their semantic arguments, and semantically different adjuncts are compatible with different types of events. Arguments,

on the other hand, can modify the type of the event referred to by the predicate. Therefore, if a constituent modifies the type of the event, it has to correspond to an argument. We know that the event type has been modified if it accepts different types of adjuncts. It is important to note that although the notion of event type can be interpreted at the semantic level, for Komlósy it is truly a syntactic notion, anchored to distributional criteria.

> 2) *If a constituent's appearance in the structure allows to expand it further by an optional adjunct (which could not be present in the original structure), than this constituent is an (optional) complement.*

Illustrating the criterion with Komlósy's example:
Mari hízott egy héten át / *egy hét alatt.
Mary was getting fatter for a week / *in a week.
Mari hízott öt kilót *egy héten át / egy hét alatt.
Mary put on five kilos *for a week / in a week.

Komlósy claims that because "egy hét alatt" ('in a week') cannot be joined to the sentence without the constituent "öt kilót" ('five kilos') being present, "öt kilót" has to have a place in the verbal argument structure, which means that it is a complement. The point behind this criterion is that there are NPs in Hungarian which, through being adjoined to the verb, change the argument structure of the given verb. This is the case of the constituents, such as five kilos in the example above, which add the meaning component 'end point' to the meaning of a structure denoting a process. In such cases, as one would expect, complementary sets of time adverbs can modify the resulting structure. This reflects the vendlerian (Vendler, 1967) distinction between accomplishments and activities.

Levin and Hovav (2005) illustrate event type shift with the resultative construction:

> The blacksmith hammered the metal.
> The blacksmith hammered the metal flat.

The sentences above refer to different types of event, since they can be modified by different types of time adjuncts:

*The blacksmith hammered the metal in to hours.

The blacksmith hammered the metal flat in two hours.

They come to the same conclusion, though their argument is semantic in nature : 'flat' cannot be an adjunct, since it adds an extra 'meaning component' to the predicate. This implies that event type shift changes the meaning of the predicate. Adjuncts cannot change the meaning of the predicate as they refer to external coordinates of the event denoted by the verb. Arguments, on the other hand, are part of the verbal meaning and thus adding an optional argument to the predicate can change its meaning. This extra meaning component can yield an event type shift. It is important to note that Komlósy's second complement test is the exact syntactic counterpart of the semantic argument of Levin and Rappaport-Hovav.

The first questionable aspect of taking this kind of expressions as part of the verbal complement structure is that adding this specific meaning component to the verb meaning is probably not a lexical property as this operation is completely predictable. Second, as (Dowty, 1991) points out, it is not even trivial whether such kind of expressions should be assigned a thematic role (although this would be obligatory if they belonged to the verbal complement structure). Komlósy himself does not specify a thematic role suitable for such purposes in his list of thematic roles. For the same considerations, Butt (2006) argues that such constituents cannot be considered as complements, and notes that "this type of data remains an issue which has not as yet received a good/standard solution within modern syntactic theories".

The third argument against event type shift as an indicator of argument-hood comes from a counter-example published earlier as Gábor and Héja (2006). We investigated whether it is possible that adjunct constituents can trigger event type shift in Hungarian. The semantic dichotomy (according to which adjuncts refer to external coordinates, while arguments denote participants of the event), as well as the widely accepted hypothesis about structural difference (adjuncts are attached to the verb phrase by a recursive rule,

complements change the category) are at stakes: if we can show that some adjuncts can trigger event type shift, we can state that these hypotheses are not well-founded. In order to understand the example, first consider the sentences below:

a) A hordó szétrobbant a gázok-tól.
the barrel exploded the gases+ABL
The gases made the barrel explode.

b) János tántorgott a bor-tól.
John wabbled the wine+ABL
The wine made John wabble.

We can see that NPs in ablative case denote a certain type of constituent the semantic role of which can be scribed as direct physical cause. This constituent can only appear with predicates that do not have an agent subject. However, we found that adding a constituent which emphasizes a non-agentive interpretation can systematically modify the type of the event denoted by the predicate so that it accepts a direct physical cause adjunct:

c) *Anna fizetett a gyógyszer-től.
Anna payed the pill+ABL
*The pills made Anna pay.

d) Anna lassan fizetett a gyógyszer-től.
Anna slowly payed the pill+ABL
The pills made Anna pay slowly.

In sentence d), the presence of the adjunct 'slowly' allows to extend the structure by another adjunct (gyógyszertől, the pill+ABL). If the complement test 2 is reliable, we have to conclude that lassan (slowly) is an argument of the verb fizet (pay), which contradicts our intuition as well as any possible definition of complements/arguments

cited previously. Hence, we can conclude that we found an adjunct that can change the syntactic distribution, as well as the event type of a predicate. It is also important to note that we are not dealing with a marginal phenomenon: many agentive verbs can be substituted to the examples above with the same result.

The distributional test no.2 is therefore not reliable, most probably because it is semantically ill-founded.

III.2.2.4 LFG Test 3: Optional Complements without Event Type Shift

The third distributional test is based upon the observation that verbs can be very similar in meaning, and at the same time have a different complement structure:

> 3) *If a word X has an expansion Y, and there is a word Z which can systematically replace X+Y, and can replace X when X is not expanded by Y, but cannot replace X when Y is present, then Y is an optional complement of X.*

In other words, the substitution of another verb which does not have the same syntactic complement yields an ungrammatical sentence. Komlósy's example for test no.3 is:

Mindannyian órákon át csodálkoztunk az eredményen.
All of us for hours were wondering the result+SUP
All of us were wondering about the result for hours.
Mindannyian órákon át csodálkoztunk.
All of us for hours were wondering.
All of us were wondering for hours.

He demonstrates that "eredményen" ('the result+SUP') is the complement of the verb by the fact that there are other verbs denoting the same type of event which do not allow the same dependent to appear in the sentence:

Mindannyian vihánccoltunk/unatkoztunk *az eredményen.
All of us were tittering/being bored *the-result+SUP

He affirms that if the NP in SUP case was an adjunct, such a substitution would not make the sentence ungrammatical. The motivation of this test is clear: if the same grammatical function is not allowed in the context of a different predicate, it is certainly a lexical property specific to a given verb (class). The problem with this criterion is the vague definition of the notion of event type. It is left unsaid how one could decide whether two predicates (except for near synonyms) refer to the same type of situation or not. As we have seen, the definition of 'event type' is grounded on the distribution of adjuncts: two predicates belong to the same event type if they can be modified by the same set of adjuncts. Using such a definition would, however, yield a tautology: if two predicates can be modified by the same set of adjuncts, the differences between their syntactic distribution must come from their different subcategorization patterns, i.e. complements. The implicit assumption behind this principle is that adjuncts are fully productive.[3] This assumption is tenable, though pragmatically less appealing: by limiting the notion of adjunct to (aspectually bound or fully productive) time adjuncts, we would refer many partially productive and predictable phenomena to the lexicon, increasing the number of counter-intuitive duplicates in the lexicon.

III.2.3 Syntax-Semantics Interface: Thematic Roles

We have seen in III.2.2 that there are inconsistencies between semantics-based and syntax-based argument definitions. Some of them are due to erroneous presumptions about a strict argument-adjunct dichotomy, others come from difficulties in deriving argument realization from constraints formulated over thematic roles. In what follows, we are going to raise an additional question about thematic role assignment in LFG and the underlying presumptions about the nature of thematic roles.

LFG divides complements into three classes: it differentiates between labelled complements, restricted thematic roles (such as location e.g.) and unrestricted thematic roles which are associated with certain syntactic positions such as subject or object. In the case of labelled complements, the exact morphosyntactic realization (case suffix or post-

[3]According to Komlósy (personal communication), event type is interpreted as aspectual agreement: two verbs express the same type of event if their aspectual properties are similar, and therefore are compatible with the same set of time adjuncts.

position in Hungarian) is formally required by the verb. In the case of arguments with restricted thematic roles, the verb requires an argument bearing a specific theta role, which can be instantiated by any possible category and morphological marker which is able to express the given role, e.g. any adverb or case-marked NP expressing a location. Finally, unrestricted thematic roles belong to subject and object position: they can be assigned different theta roles, e.g. there are agent subjects as well as patient or theme subjects, etc.

This trichotomy presents a theory-internal methodological contradiction. The LFG analysis aims to describe subcategorization on syntactic grounds, where complement and adjunct constitute different grammatical functions (although the difference is semantically motivated, the methodology strictly requires a syntactic underpinning of the semantic features and roles used in the analysis). However, the definition of this trichotomy refers to pieces of semantic knowledge when it attributes a meaning to case suffixes (even when they appear on complement NPs). The underlying presumption is that there is a set of case suffixes which have a meaning on their own: they can express the same role in the context of different predicates. The morphosyntactic marker of the role has a certain autonomy: its meaning, and therefore the meaning of the complement NP, does not come from the argument relation with the predicate. This violates the Semantic Specificity Criterion (location as a semantic role is not restricted to a set of predicates: in fact, it is one of the most prototypical adjunct roles), and contradicts the view according to which only adjunct express external coordinates of events. When putting theory into practice, we have to face that we lack an exact method to define the meaning of a case suffix. To illustrate that with an example, let us take the ablative case in Hungarian. It has more than one productive uses: with movement verbs, it can express the SOURCE thematic role (távolodik/odébbmegy valamitől, 'go away from sg'), while it can also productively express CAUSE with non-agentive predicates as we have seen in the examples in III.2.3.3 (János tántorgott a bortól 'the wine made John wabble'). According to Komlósy's argument test no.3, all of them are complements. How can we decide whether they are labelled or thematically restricted complements? Although the meaning of the verbs távolodik and odébbmegy is practically identical (except that odébbmegy cannot be used in a figurative sense), only odébbmegy allows its SOURCE complement to be expressed by a suffix or

postposition other than the ablative. If we posit a labelled complement in one case and a thematically restricted complement in the other, a valuable generalization is lost. Why should we indicate in the lexical entry of távolodik both the fact that it has an optional argument in ablative case and the fact that the semantic role of this argument is SOURCE, when we know from the lexical entry of odébbmegy that the ablative case is capable to express that role on its own? The case of the verb tántorog is slightly different as its CAUSE complement can also be expressed with the postposition miatt ('because of') – though this postposition is a typical adjunct construction allowed with most verbs. This corresponds to our intuition that thematically restricted complements are really close to adjuncts.

Another problem with this classification is that it defines subject and object as thematically unrestricted complements at a general level, while the definition of labelled complements and thematically restricted complements is verb-specific. This means that the subject of any verb will be considered as thematically free, despite any restriction the individual verbs put on their subject, just because subjects in general can have different theta roles. In reality, the subject of a given verb is always thematically restricted, just like its thematically restricted oblique complements.

III.2.4 Summary

In this section we attempted to give an overview on how verbal lexical entries are represented in lexicalist and configurational theories of language. With the goal of constructing a Hungarian verbal subcategorization lexicon in mind, we enumerated the different criteria for distinguishing lexical information from syntactic information. All of the theories cited above share the view that syntactic complements correspond to semantic arguments and that complements, as opposed to adjuncts, are not productive nor predictable. Their semantic role, as well as their syntactic realization, needs to be derived from the information stored in the verbal lexical entry. We have seen that there are substantial problems with semantic and syntactic definitions of arguments. The first problem is that lexical semantic representations and semantic argument positions are not directly accessible – whether a given element belongs to a particular verbal argument structure is most frequently tested through its syntactic behavior. However, syntactic complement tests are

language-specific and mostly rely on configurational positions. These configurational differences are not evidenced by surface structure in Hungarian. Other syntactic complement tests, like obligatoriness or event type shift, yield coherence problems at the level of lexical semantic representation. Syntactically obligatory complements must be considered as arguments and included in the lexical representation (they have to get theta roles), but this may result in "syntactic diacritics" in the lexical semantic representation when the obligatory complement cannot be assigned a meaningful theta role. Moreover, constituents with the same syntactic realization and the same semantic content can behave like an obligatory complement with certain verbs or like an optional adjunct with other verbs (e.g. él valahol 'live somewhere' vs alszik valahol 'sleep somewhere'). They will have different representations which is counter-intuitive. We have shown by virtue of an example that adjuncts can also trigger event type shift in Hungarian. This observation undermines the supposed structural difference between adjuncts and arguments: since in our example an adjunct changes the syntactic distribution of a predicate, it cannot be attached to the verb phrase by the application of a recursive rule. Finally, we have demonstrated that the trichotomy of thematic role assignment in LFG is formulated in an incoherent way, probably due to the fact that the analysis is constrained to categorize some constituents as complements despite their adjunct-like behavior.

III.3 Semantic Verb Classes and Semantic Roles

III.3.1 Claims

The arguments presented in the previous chapter demonstrated that a proper classification of Hungarian NPs as arguments or adjuncts cannot be achieved using (surface) syntactic clues. Therefore, a parallel syntactic and semantic processing is needed to classify case-marked NPs. Since we cannot define arguments with usable argument tests, we are bound to prioritize the description and enumeration of the different types of adjuncts.

Research on lexical semantics and semantically motivated mapping has been concentrating on predicting the syntactic realization of arguments, taking for granted (either explicitly or implicitly) that the distinction between arguments and adjuncts is known, and that adjuncts' syntactic realization is governed by productive syntactic rules, not lex-

ical properties. On the other hand, as mentioned in III.2.2.3, Komlósy (1992) claims that the event type (i.e. semantic category) of the predicate determines the scope of adjuncts that it can take. However, besides the correlation between verbal aspect or actionsart and time adverbs (e.g. Vendler, 1967 or Kiefer, 1992 for Hungarian), the distribution of adjuncts among verbs or verb classes did not receive significant attention, especially within the lexical semantics framework. The definition we propose for argument-adjunct dichotomy and the resulting methodology exploits the idea that lexical semantics not only influence complement structure but is the key to the argument-adjunct distinction and to the realization of adjuncts. The focus of the above cited lexical mapping/linking theories is to predict the syntactic realization of 'typical' arguments: they are primarily dealing with subjects, direct or indirect objects and examine oblique complements to a lesser extent, principally when they alternate with the functions above. Unlike these pieces of work, we are focusing on phrases which correspond to oblique complements or adjuncts, our ultimate goal being to find a method for deciding which of these have to be lexically coded. On the basis of the arguments presented in 3.2, we reject the hypothesis of a strict dichotomy between complements and adjuncts at the syntactic level, and we do not postulate a semantic dichotomy either. Instead, we claim that certain elements in the verbal complementation frame have a higher degree of semantic and syntactic autonomy than others, due to the fact that they are licensed by more general verbal meaning components and therefore are more productive than others. The constituents we are dealing with are typically case-marked nominal (pronominal, adjectival) phrases. A given case suffix as a morphological marker of syntactic function can mark diverse functions (complement, adjunct of time, adjunct of location, manner etc.), depending on its context. We claim that the syntactic function encoded by case suffix as well as the meaning it conveys depend on the lexical semantic class of the predicate it occurs with. Therefore, knowing the lexical semantic class of a verb is a prerequisite to telling which argument structure(s) it takes and which adjuncts it allows. The difference is that the adjunction process is productive within the given verb class: the same case suffix used within the same class always has the same semantic and syntactic function. In other words, the semantic role marked by a given case suffix is constant within verb classes. Hence, we can attribute a meaning to the morphological case when it occurs on adjunct NPs, and this meaning is

added compositionally to build up the meaning of the sentence. We are going to argue for compositionality as the main semantic characteristic of adjuncts. Compositionality in this context means that the grammatical marker of the adjunct — in the case of Hungarian, morphological case — is not a purely functional element but has its own meaning which adds up with the meaning of the predicate and that of the adjunct NP according to certain rules. We claim that argument NPs have a non-compositional meaning in that 1) their semantic role cannot be extensively defined without reference to the idiosyncratic meaning components of the individual predicate and 2) therefore, we cannot talk about real productivity, which means that they have to be treated in the lexicon.

III.3.2 Lexical Semantic Classes

We us the term 'lexical semantic class' in the approximate sense of (Levin, 1993), defined by distributional criteria on the one hand, and meaning components shared across class members on the other hand. Levin's work demonstrates that verbs sharing the same meaning component(s) participate systematically in the same diathesis alternations. Let's consider the following example:

The verbs break, cut, hit, touch have two arguments, realized as a subject and a direct object:

(a) Margaret cut the bread.
(b) Janet broke the vase.
(c) Terry touched the cat.
(d) Carla hit the door.

But they differ with respect to their diathesis alternations. Middle alternation only applies to cut and break:

(e) The bread cuts easily.
(f) Crystal vases break easily.
(g) *Cats touch easily.

(h) *Door frames hit easily.

On the other hand, the conative alternation is allowed for cut and hit, but not for break and touch:

(i) Margaret cut at the bread.

(j) *Terry touched at the cat.

Finally, only break does not participate in the third kind of alternation:

(k) Margaret cut Bill's arm.

(l) Margaret cut Bill on the arm.

but:

(m) Janet broke Bill's finger.

(n) *Janet broke Bill on the finger.

Levin explains these phenomena by suggesting a decomposition of verbal meaning (Levin, 1993; Levin and Hovav, 2005). She divides it into a *core meaning*, the idiosyncratic component specific to the individual verbs, and an *event schema*, which constitute a generalization over the meaning components shared among natural classes of verbs. These meaning components can be captured by the so called metapredicates. We accept this position, implying that metapredicates allow different kinds of syntactic operations (e.g. diathesis alternations in English). We are going to apply this schema to explain other types of distributional phenomena, namely, the appearance of optional adjuncts. Levin claims that cut, hit and touch share the CONTACT meaning component, which is responsible for the alternation exemplified by sentences (k)-(n). The conative alternation applies to verbs having the meaning component MOVE, i.e. cut and hit. Finally, middle alternation is allowed by the meaning component CAUSE CHANGE, shared by cut and break, but not hit and touch.

To sum up, Levin's methodology is based on the presumption that semantic mean-

ing components influence verbal syntax: they define the scope of diathesis alternations which apply to a given verb. Verbs can be grouped together according to their meaning components, where the resulting groups will be homogeneous with respect to one or more diathesis alternation(s). Only linguistically relevant metapredicates can be used in the classification, and this linguistic motivation can be ensured by relying on distributional criteria.

At this point, we can refer back to the notion of event type used by Komlósy (1992), discussed in 3.2. We propose the following semantic definition for event type:
Two predicates are said to belong to the same event type if they share the same syntactically relevant meaning components.
However, when we define syntactically relevant meaning components on the basis of distributional properties, we cannot restrict these properties to diathesis alternations (as does Levin), as we do not have syntactic complement tests to decide which constituents can participate in such alternations. Therefore, we return to our original claim: namely, that it is necessary to prioritize the description and enumeration of adjuncts: 1) if we want to give a syntactic definition of event types or lexical semantic classes, it can be done on the basis of adjuncts' distribution, 2) if we want to give a semantic definition, we resort to metapredicates, which, in turn, emerge from the study of the compatibility between verb meaning and adjunct semantic roles. We can also conclude that syntactic and semantic processing should not be done modularly but at the same level, in parallel: we rely on semantic notions when we describe lexical verb classes, and different types of adjunct are categorized according to the semantic role they fill. On the other hand, if we want to guarantee the linguistic relevance of our semantic features and the objectivity of description, semantic notions have to be anchored to distributional criteria.

III.3.3 Semantic Compositionality

We have seen that the three main characteristics usually attributed to adjuncts as opposed to complements are a) productivity, b) compositionality (as opposed to Semantic Specificity, see section III.2.1) and c) recursion (the fact that an arbitrary number of adjuncts can be added to the same constituent without changing its category/distribution). We rejected criterion c) by showing a counter-example in Hungarian. As to productivity,

we have seen that it only applies to a few, prototypical examples of adjunction. However, if we redefine productivity to be understood over verb classes, it will become accessible and interpretable to a wider range of phenomena. We claim that semantic compositionality and productivity are mutually interdependent. If we accept that different types of semantic roles are licensed by predicates' semantic metapredicates and hence can appear with verbs sharing the given metapredicate, we can say that these semantic roles are adjunct roles and the syntactic structure they display are productive within the verb group.

A prototypical example of adjunction is constituted by location adjuncts: 1) they are productively used with the majority of predicates, 2) they are usually optional, 3) their meaning or semantic role is stable (in the sense of Koenig's Semantic Specificity Criterion (2003): the semantic relation they convey is not specific to an individual predicate or predicate class). E.g.:

1) Mari a szobájában aludt.

Mary was sleeping in her room.

2) A fiam ebben a városban dolgozott.

My son used to work in this town.

Time adjuncts, as mentioned above, are less productive since verbal aspect constrains the type of adjuncts allowed; we refer to Komlósy's examples to illustrate this:

3) Mari hízott egy héten át / *egy hét alatt.

Mary was getting fatter for a week / *in a week.

4) Mari hízott öt kilót *egy héten át / egy hét alatt.

Mary put on five kilos *for a week / in a week.

Although these examples contain complex predicate structures, it has been stated that verbal aspect — at least in some cases — is lexically coded in Hungarian (Kiefer, 1992). Therefore, lexical properties of verbs can determine the scope of time adjuncts the verb takes, even with respect to the prototypical category of time adjuncts. Let us now take the example of an extensive group of adjunct-like, optional constituents that are only allowed with a certain type of predicates. The 'causalis' case suffix -ért ('for') for instance

is largely considered as an adjunct; however, it is only compatible with agentive verbs, the subject of which is undertaking an action on purpose:

5) János tanul/fut/dolgozik a jutalomért.
John is working / is running for the price.
6) *János megbotlik/ felébred a jutalomért.
John stumbles / awakes for the price.

These examples illustrate that adjuncts are — at different levels — sensitive to lexical semantic properties of verbs and they are only compatible with certain — smaller or larger — verb groups. On the other hand, this statement holds for arguments as well, to the extent that it even constitutes a basic principle of mapping theories. Consider now the example of verbs expressing a change of state, with a punctual aspect: felriad ('to be roused') / felkacag ('to start giggling') / felébred ('to awake'). The cause of the change of state can appear in the sentence as an optional element marked by the case suffix -ra/-re ('sublative'):

7) Jancsi felébred/felriadt a zajra.
Johnny was roused/awaken by the noise.
8) Jancsi felkacagott erre a mondatra.
Johnny started giggling at/when hearing this sentence.

The same case suffix appears systematically with another intuitively coherent verb group and with a completely different semantic role:

9) Mari állítása erre az érvre alapoz/épít/támaszkodik/apellál.
The statement made by Mary is based on / builds on / relies on / addresses this argument.

What is common in the examples 3-5) and 7-9) is that they include an optional constituent with a specific semantic role which is only compatible with a restricted set of

predicates: not even the more prototypical adjuncts are totally productive over all the predicates. The difference in productivity is primarily a quantitative one: some verb groups, like that with an agent subject, are far more extended than the group of verbs expressing a punctual change of state. On the other hand, we have no reason to consider the first group semantically more coherent than the second one. Instead of considering adjuncts' productivity as a function of the range of verbs they go with, we suggest testing whether the verb group forms a natural class. If a constituent is allowed for a natural class of verbs, it can be considered as an adjunct.

Since adjunct functions are coded by case suffixes (and prepositions/postpositions, depending on the language), and they are not defined in the lexical entry of the predicate (unlike semantic arguments), these grammatical elements are bound to have a lexical meaning on their own. The meaning of a Hungarian case suffix can be defined as a semantic relation between the referent of the case-marked nominal element and that of the predicate. This relation can be transparent (e.g. the 'temporalis' suffix -kor) or less transparent (e.g. the examples 7-9 above), depending on productivity and on the ambiguity of the case suffix. The case-marked NP can appear in the context of a predicate if the semantic role encoded by the case is compatible with the meaning of a predicate. To put it more formally, a set of metapredicates (meaning components) can be used to describe the lexical semantic properties of verbs. The metapredicates included in the lexical entry of the verb license certain semantic roles. Verb groups defined by shared metapredicates go together with the same set of adjuncts: i.e. the same set of semantic roles expressed by the same morphosyntactic markers (akin to Komlósy's notion of event type, our predicate classes are are made up by verbs which can be extended by the same group of adjuncts, albeit with a different semantic interpretation). Our verb groups are very analogous to those of Levin (1993): they are established based on distributional criteria, in order to ensure that they constitute natural classes with linguistically relevant meaning components. The main difference is that we are (deliberately) not concentrating on diathesis alternations but on optional adjunct-candidates. Second, as we are not dealing with arguments, our notion of semantic role does not coincide with the usual set of thematic roles (although there may be overlaps, especially with respect to instrument, comitative, source or goal adjuncts; see also Rákosi (2006) who argues that some dative

experiencers in Hungarian should "systematically be treated as adjuncts that bear a thematic role").

As we have seen, the metapredicates we use can be very general, applying to most of the predicates (e.g. in the case of time or location adjuncts), and on the other edge of the scale we find more and more specific groups, such as the above mentioned 'cause change of state', or 'quantitative change'.

After having defined productivity in terms of natural classes, we can come back to the compositionality of adjuncts and attempt to define it within the new interpretation frame: *If the case suffix x can systematically appear in an NP syntactic dependent of any verb belonging to the natural class C, and the NPs have the same semantic role among any verb member of class C, then the semantic role corresponds to the lexical meaning of the case suffix x.* The meaning of a complex adjunction structure is built up compositionally, i.e. equals the sum of verb meaning + semantic role + NP meaning. Semantic arguments on the other hand have a verb-specific semantic role which cannot be generalized among classes, neither can it be expressed independently of the verb: or at least, such a tentative description of the semantic role would not be extensive. Therefore, the semantic role has to be included in the lexical representation of the predicate. Accepting this definition implies that we do not have to include the adjunct semantic role in the lexical meaning of each individual verb, even if the role is restricted to a set of predicates. This way we can restrict argument slots to real semantic arguments, which corresponds to the standard assumptions. on the other hand, we have to lexically encode the natural class the verb belongs to (by means of semantic meaning components).

According to this analysis, punctual change of state (examples 7-9) is a relevant meaning component and the NPs with the sublative case suffix (-ra/-re) are an example of adjunction on the same account as e.g. perfective time adjuncts with accomplishment verbs. We do not postulate qualitative difference between the two groups: they only differ with respect to the quantity of verbs belonging to the group. Since we claim that productive syntactic rules are those which apply to verb groups, once our groups are defined based on objective criteria, we can expect the number of arguments significantly reduced in our lexical representation.

III.3.4 Consequences

The claims expressed in 3.3.1 raise a certain number of theoretical and methodological questions. While the latter will be addressed in detail in chapters IV-VI which deal with the research and development done on the basis of this theoretical background, we will attempt to resume and address the most important concerns about the coherence of our claims.

III.3.4.1 Semantic Predicate Classes

Since lexical semantic classes constitute the foundation of our definition, it is extremely important to define them on a solid basis. We are aware of the fact that by introducing semantic criteria in the definition of syntactic functions we run the risk of losing objectivity. This implies that semantic metapredicates have to be underpinned by independent distributional tests to guarantee that we indeed deal with natural classes. Moreover, as lexical semantic classes come into the picture when differentiating adjunct roles from complement roles, defining them solely on the basis of one adjunct shared would be a circular argument for the relevance of semantic verb classes. Levin (1993); Levin and Hovav (2005) use diathesis alternations. We are going to adapt the distributional tests to the structure of Hungarian and use either alternations involving arguments or adjuncts, or simply different complementation patterns with potentially different semantic roles. The idea behind this decision is that if there are several adjunction possibilities shared among a group of predicates, they are likely to belong to the same semantic class (or express the same event type in Komlósy's terminology). Productive morphological derivation involving change in complementation patterns can also be considered. More examples and a detailed description of the methodology will be given in chapter IV, here we limit ourselves to illustrate the point by presenting the lexical semantic group which can be described as 'cause change in mental state'. Verbs belonging to this class include among others meglepődik ('to be surprised'), elszomorodik ('to get sad'), felvidul ('to cheer up'). The distributional pattern for this class would involve the alternating syntactic function of CAUSE.

10a) Mari meglepődött/elszomorodott/felvidult.

Mary got surprised/sad/ cheered up

Mary got surprised/sad/ cheered up.

10b) Mari meglepődött/elszomorodott/felvidult a hírtől.

Mary got surprised/sad/ cheered up the news+ABL

Mary got surprised/sad/ cheered up hearing the news.

11) A hír meglepte/elszomorította/felvidította Marit.

The piece of news surprised Mary/ made Mary sad / cheered Mary up.

12)János meglepte/elszomorította/felvidította Marit a hírrel.

John surprised Mary/ made Mary sad / cheered Mary up the news+INS.

John surprised Mary/ made Mary sad / cheered Mary up by (telling her) the news.

Sentences 10-12 show that the semantic role CAUSE can be filled by different oblique syntactic functions, all of them being optional. While hírrel in sentence 12 could potentially be analyzed as an instrument and as such, characteristic to a much more extensive verb group (agentive/intentional actions), this verb group would not allow the alternation between 11) and 12). On the other hand, the alternation between 10) a-b and 11) characterize any predicate with the meaning component 'cause change', but they do not accept 12) unless they belong to the more restricted group 'cause change in mental state'. Therefore, the totality of the alternating syntactic structures designate exactly the lexical semantic group which can be paraphrased as 'cause change in mental state' [4].

III.3.4.2 Semantic Roles

The other problematic issue concerns the objectivity of the notion of semantic role. Our thesis is that adjunct-type semantic roles can be extensively defined without making reference to the core meaning of the verb. This is exactly the property that enables us to attribute the expression of this role to the morphological marker, in other words not to include it in any way in the lexical representation of the verb. This is not a novel statement

[4]On a more detailed analysis of this group, see 5.1.2

in itself: the above cited argument realization theories, although at an implicit level, also presume that only semantic arguments are included in verbal lexical entries. /footnote-However, they remain silent as to the nature of adjuncts' semantic roles. What makes the difference is that we use this property of adjuncts as a differentiating feature. When we say that adjunct-type semantic roles can be extensively defined without reference to verb meaning, we accept Levin's distinction between core meaning and generalizable meaning components. Adjunct roles can be defined according to metapredicates or independently (e.g. direct physical cause, mode or time are considered to be extensive definitions); on the other hand, an extensive definition of any argument role needs to refer to core meaning. Since core meaning is idiosyncratic, such roles depend on verbs. This idea can be interpreted analogously to Koenig's Semantic Specificity (2003) in terms of Dowty's (1991) thematic proto-roles and individual thematic roles. As opposed to the specificity of arguments (here in terms of individual predicates or an enumerated/arbitrary list of predicates), we claim that an adjunct-type semantic role is generalizable (to one or more natural classes of predicates). The extensive description of an argument-type role in our view has to be similar to an individual thematic role in Dowty's terminology, and as such, a priori specific to its governing predicate (e.g. singer, trustee etc.). Although Dowty suggests generalizations over thematic roles in terms of proto-roles, as explicitly stated, these roles do not constitute extensive descriptions. We presume that adjuncts in turn can be extensively described at the level of proto-roles, since their role is identical in the context of every verb which disposes of the meaning component allowing for the given adjunct. We conceive a representation of adjunct roles in the way proposed by Dowty for proto-roles: by a set of entailments. Chapter V concentrates on the syntactic and semantic description of verb classes and adjunct roles.

III.3.4.3 Lexicalized Verb+NP structures

In many cases, the description of an adjunct role does not match the actual role of a particular NP in the given adjunct-like position, as the structure itself is lexicalized with a different meaning. The resulting verb+NP structure is therefore non-compositional, the semantic relation does not correspond exactly or is completely different from what we would expect. E.g.:

compositional:

Tegnap a szomszéd városban ebédeltünk..

Yesterday the neighboring town+INE we-had-lunch

Yesterday we had lunch in the neighboring town.

non-compositional?:

Tegnap étteremben ebédeltünk.

Yesterday restaurant+INE we-had-lunch

Yesterday we had lunch at a restaurant.

compositional:

A fiam a szomszédba jár.

My son the neighbors+ILL goes.

My son goes to the neighbors.

non-compositional:

A fiam iskolába jár.

My son school+ILL goes.

My son goes to school.

In the compositional examples above, the predicate has a subject and an oblique NP adjunct: eating, as any activity, can have an optional location adjunct, as well as going, or any directed movement can have an optional directional adjunct. The interpretation of the semantic role is regular in both cases. However, the non-compositional examples show how certain institutionalized activities (eating at a restaurant or going to school) confer different entailments: eating at a restaurant entails ordering and paying, besides being at the restaurant while eating; going to school entails being enrolled and attend classes. The syntactic particularity of these institutionalized activities is that in Hungarian, they occupy a specific position, namely, the 'verb modifier' position (É. Kiss, 2002). The NP, verb prefix or adverb in this position immediately precedes the verb, and it has to be non-

referential (only bare NPs can go to this position). Let us now consider the example below:

Tegnap buszon ettünk.

Yesterday bus+INE we-ate

Yesterday we ate on a bus.

In order for the sentence to be grammatical, it has to have a non-flat structure in É.Kiss's terminology, involving a non-neutral discourse configuration. Depending on prosody, it can be interpreted either as "buszon" ('on the bus') being part of a contrastive topic or as a verb modifier (*igevivő*). If the sentence has a contrastive topic intonation, the corresponding context would be for instance *"Yesterday we ate on a bus, as opposed to other days when we eat at home/in the car..."*. If the verb is pronounced with no stress, the predicate structure is interpreted as an institutional activity, and the complex would imply that the speaker has the habit of eating on a bus as a pass-time. Although this interpretation is by far less natural than the non-compositional meaning of going to school, the distinction seems to be only pragmatical.

The fact that virtually any kind of complement or adjunct can be forced to be interpreted as an institutional activity while the quality of this institutionalization depends on pragmatical factors is an argument for not considering these NPs automatically as complements. Complements receive their semantic role from the verb; their syntax and semantics depend on the lexical properties of the verb. The verb modifier structure is best analyzed as a syntactic structure with its particular constraints on semantic interpretation. This also holds for other configurational phenomena involving focus, contrastive topic and other discourse functions, as well as quantification or negation.

Structures with a verb and a verb modifier constitute a frequent, but obviously not a unique type of non-compositional structure with predicates and unpredictable semantic roles. Many multiword expressions belong to this category. However, this does not affect our analysis of the adjuncts as compositional structures for two reasons. First, the fact that one particular instance of adjunct behaves in an unpredictable way with respect to semantic interpretation does not affect the other, regular realizations of the same role.

Moreover, there is no motivation to analyze a specific role as an argument role only because it allows lexicalization, since multiword expressions show semantic or syntactic properties which differentiate them from complements. Second, any kind of syntactic dependent can be lexicalized with a predicate, even modal adverbs, which indicates that the specificity of the original semantic relation marked by the given morphological marker does not influence the possibility of lexicalization. If we considered that every structure that can be lexicalized automatically qualifies as a complement, the predictive power of our model would decrease drastically.

Discussing the types and properties of multiword expressions goes far beyond the scope of this thesis and we are not going to draw a conclusion about their status hereby. We limit ourselves to state that any structure with a non-compositional meaning has to be lexically coded, but in the case of multiword expressions, contrary to verbal arguments, lexical specification applies not only to the verb but to the complement or adjunct NP itself. Therefore, they constitute a different category and their study falls beyond our current objectives.

III.3.5 Conclusion

In chapter III. we outlined the concept of semantic argument and syntactic complement in linguistic theories. We analyzed the attempted argument/complement definitions in different theoretical backgrounds and concluded that they lead to incoherence problems. In particular, we have criticized two basic theoretical presumptions about verbal complements and adjuncts: 1) semantic arguments are syntactically realized as complements; their semantic argument status is formally encoded by their unique thematic role; 2) adjuncts cannot change the syntactic distribution of the constituent they are attached to, since they are added by recursive rules. We have shown that these claims, though widely accepted across argument realization theories, raise a number of problems. The presumption that semantic arguments always correspond to syntactic complements and vice versa poses problems for thematic role assignment to some elements which syntactically behave like complements. By virtue of a counter-example, we have proven that not only complements but also adjuncts can change the syntactic distribution of the constituent they are added to. We have shown that current syntactic complement tests

cannot be applied satisfactorily to data from a non-configurational language, due to the fact that there are no systematic syntactic differences between complements and adjuncts in the flat VP structure of Hungarian. We have also demonstrated that the supposed productivity of adjuncts is significantly less widely tenable than it is implied in literature.

Since we do not dispose of reliable complement tests, neither have we a strong intuition about complement status — this intuition being based on the obligatory character of complements in other languages — we suggested introducing semantic compositionality as a distinctive feature. We believe that the semantic relation between a verb and an adjunct corresponds to the lexical meaning attributed to the morphosyntactic marker of the relation, while in the case of arguments, the semantic relation is lexically encoded in the lexical entry of the verb. Adjuncts are productive in terms of lexical semantic predicate classes. We can then talk about compositional meaning when the same morphological marker expresses the same semantic role on the NPs appearing in the context of any verb belonging to the given verb class. The appearance of an optional adjunct is made possible by the fact that members of the semantic verb class share a certain number of meaning components that are compatible with the semantic role filled by the adjunct. In order for our argument to be valid, lexical semantic classes have to be defined on the basis of independent syntactic criteria.

The following chapters present the work completed on this theoretical basis: the construction of a lexical database of verbal complement structures (IV), a methodology for the systematic study of adjuncts and the manual creation of distributional/semantic verb classes (V), and an attempt to automatically induce lexical semantic verb classes by an unsupervised learning method (VI).

Chapter IV

The Verbal Valency Lexicon

IV.1 Introduction

Lexicon is conceived in linguistic theory as a repository of idiosyncratic information. Idiosyncrasy refers to pieces of information specific to a given lemma: in case of complementation patterns, idiosyncrasy covers form-meaning pairs which cannot be decomposed to smaller units in a way that the bigger unity can be generated from the small one by the application of productive grammar rules. With respect to verbal subcategorization, this view of the lexicon entails that some elements in the context of the verb are predictable and can be parsed/generated by productive rules (adjuncts), while others are idiosyncratic and have to be encoded in the lexical entry corresponding to the predicate (complements). [1] In the previous chapter, we argued in favor of a gradual differentiation between lexical and productive syntactic phenomena. This chapter presents our efforts on putting this approach into practice when constructing lexical resources for verbs. In what follows, we will present a Hungarian verbal lexicon built for natural language processing purposes. The primary purpose of the lexicon is to support parsing for various NLP applications; correspondingly, it has been adapted and integrated into several applications such as information extraction (Prószéky, 2003), machine translation (Prószéky and Tihanyi, 2002; Tihanyi, 2006), prosodic annotation for human-machine interaction

[1] Another approach of the lexicon, adopted mostly by NLP researchers, puts forward frequency as the main criterion for including subcategorization frame candidates in a lexicon (see 4.1.2.2); however, we accept idiosyncrasy as a principle for selecting lexical information to be stored.

(Tamm et al., 2008), and psychological content analysis (Vincze et al., 2010; Pólya and Gábor, 2010; Ehmann and Garami, 2010). Considering the resource-demanding nature of the task of building a large-coverage subcategorization lexicon, it was an explicit goal to construct a lexicon as general-purpose as possible, the main requirement being to provide an exhaustive description of unpredictable verb-specific subcategorization information. In this chapter we will present the composition of the database, the organization of the annotation process with a particular emphasis on the coding guidelines and its linguistic background (the syntactic definitions and tests to be used by the annotators), and the structure of the database. Instead of giving a detailed technical description of the properties encoded in the lexicon, we will present several individual linguistic phenomena to illustrate how such complex linguistic structures can be modeled inside a computational lexical database. Since the coding guidelines were dynamically updated according to our findings during the first phase of coding, we can say that our approach was a data-oriented one. As we have seen, Hungarian was a resource-scarce language in the beginning of our project: very few large-scale annotation/coding works had been undertaken previously, our project was among the first ones to involve confrontation to large amounts of data. In particular, we believe that the so far unexplored fields of a) adjunct classification and b) verb class-specific syntactic alternations require further detailed study in order to be able to implement a rule-based deep parser and potentially move forward towards a semantic processing of Hungarian. In chapters V and VI, we present two different experiments aiming at enriching the database by following these two goals.

IV.2 State of the Art in Verbal Lexicons

Theoretical linguists has shown little interest in the lexicon for a long time: research in syntax focused on the formulation of productive rules. The lexicon was seen as a simple list containing all the pieces of information which deviate from regular syntax. Neither the internal structure of the lexicon, nor the question of how lexical information could be listed were put in the centre of theoretical research. Although theoretical and computational linguistics were more in line with each other in the '50s and the '60s than at the present time, the role of the lexicon has always been an aspect of divergence. Compu-

tational linguists attributed more importance to the lexicon since, as we will see, lexical encoding is an integral part of basically every computational linguistic tasks or NLP application. One reason for this difference between linguistic theory and CL is that linguists work by introspection, the data they use are usually provided by the linguists themselves, according to their linguistic intuition as native speakers. Computational linguistics, on the contrary, is a highly data-oriented field which cannot afford to pay less attention to theoretically less appealing or less challenging issues. When developing human language technology applications or conducting research in this orientation, efficiency and recall are crucial goals: the application has to cover as much of the input natural language text as possible. Besides the specific characteristics of the input data, the application domain is also of a crucial importance. For instance, in a machine translation system, lexical specificities of the target language impose requirements on the encoding of source language: regular and irregular phenomena do not necessarily coincide across languages. However, we believe that well-established and linguistically grounded criteria can help to reduce contingency and dependence on application domain.

In the following section we give a short overview of the available machine-readable verb lexicons. The lexical databases presented hereby deal with different aspects of verbal syntax (complement structure: syntactic subcategorization, semantic selection of complements), semantics (meaning description, hierarchy of semantic relations between verbs) or syntactic/semantic classification of predicates.

IV.2.1 Manually Constructed Verb Lexicons

Computational lexicons built by linguists/lexicographers are considered to be more accurate than automatically acquired lexical databases (e.g. (Messiant, 2010), p.85). However, in order for this statement to be true, manual construction of lexical resources have to rely on unambiguous coding guidelines to ensure that coders working on the lexicon in parallel produce coherent output. We will summarize the coding guidelines and the main features of some important verbal lexicon projects below.

Lexicon-grammar, the representation method put forward by Maurice Gross (Gross, 1975), is a syntax- and data-oriented approach. As the name lexicon-grammar implies, Gross argues in favor of a uni-level encoding of grammatical and lexical information. The

French lexicon-grammar database classifies verbs into categories based on their syntactic base constructions. A base construction is composed of the set of complements that obligatorily appear with the verb in a simple sentence: in other words, it constitutes a minimal construction with the complements needed to make a grammatical sentence. Further constructions are derived from this base construction by means of transformations. These constructions are equally listed for each verb class. Since transformations do not necessarily apply to all the verbs in a class, it is coded by a binary feature (+/-) whether the individual verb accepts the transformation or not. The key to complement structure is the presumed obligatory character of complements (compléments essentiels). According to the definition, they are considered to be obligatory at the semantic level; however, this property is tested via syntactic obligatoriness. As the tables in their original format are not easily adaptable for NLP purposes (Gardent et al., 2006), two projects have recently emerged focusing on converting them into a more machine-exploitable format (SynLex, 5.244 verbs (Gardent et al., 2006) and LGLex, 5.694 verbs (Tolone, 2011)).

The authors of the DicoValence (formerly: PROTON) subcategorization lexicon (van den Eynde and Mertens, 2006) use distributional tests based on pronominal substitution of arguments(van den Eynde and Blanche-Benveniste, 1978). These tests eventually also rely on the supposed obligatory character of complements. The resource contains subcategorization information for 3.738 French verbs.

The Vallex subcategorization lexicon (Lopatková, 2003) for Czech verbs have been constructed in parallel to the manual annotation of the Prague Dependency Treebank (Hajič et al., 2001). Syntactic complement structure and semantic argument structure are differentiated both in the treebank and in the lexicon. The criteria for including a constituent in the complement structure description are 1) obligatoriness and 2) the semantic role of the constituent. Constituents with the following semantic roles are by definition considered as complements, whether they are obligatory or not: Actor (or Actor/Bearer, ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF). Constituents with a different semantic role are considered as adjuncts; however, obligatory adjuncts are also included in the valency of the verb. Obligatoriness is understood at a pragmatic level, tested by a 'dialogue test'. A third category is distinguished for the so called quasi complements: they are usually optional constituents which fill a constant semantic role specific

to a set of verbs. The syntactic realization of these quasi-complements is subcategorized for by the individual verbs according to Lopatková.[2] An example of quasi complements is the group of modifiers of amount/measure with verbs expressing quantitative change (prodlouit o hodinu 'to extend by one hour').

The FrameNet lexicon (Baker, 1988) aimed to formalize verbal lexical semantic representation in English based on the semantic representation theory of Fillmore (1982). This lexicon is focused on semantics: predicates (including verbs, predicative nouns and adjectives) are classified into fine-grained semantic groups. The portion of meaning shared between group members is defined as a frame, i.e. the totality of semantic roles specific to the group and assigned to the arguments of the words belonging to the group. E.g., the frame communication includes semantic roles such as message, addressee, channel, enunciator etc. The inclusion of an element in the frame does not depend on its syntactic function (whether it is a complement or an adjunct), but solely on its semantic role. It is important to note that semantic roles are generalized only at the group level, the frame elements diverge from group to group - whence the high number of different roles. This feature makes FrameNet semantics more difficult to learn for ML systems. Corpus data, e.g. example sentences can also be used to complete the lexicon or to guide the editors' intuition and to eliminate the inconsistencies as much as possible. As a by-product of the lexicon, a corpus annotated with semantic roles was also created (Baker et al. 1998). Similarly, the SALSA project at the University of Saarland attempts to create a lexical semantic database of German verbs according to FrameNet's representation. A German corpus was annotated with semantic roles first, and the lexical semantic database was built based upon the corpus data (Erk et al., 2003).

Jackendoff (Jackendoff 1983; Jackendoff 1990) suggests a semantic representation by predicate decomposition. Semantic decomposition of predicates is done via Lexical Conceptual Structures (LCS). Conceptual structures can be primitives (basic constituents of meaning, e.g. cause, go), semantic fields which are used to define selectional restrictions (e.g. time, location) or conceptual constituents that define a top ontology for predicates

[2]This question depends on how lexical entries are conceived. The group of quasi complements seem to overlap with our definition of partially productive adjuncts: while we consider that they are specific to verb classes, which can be represented by lexical semantic features; Lopatková suggests to include them in individual verbal lexical entries.

(event, state). He redefines thematic roles in a way that each role is characterized by the conceptual constituent of which the role can be a semantic argument. The lexical conceptual structures serve as a representation from which syntactic argument realization can be predicted. Levin (1993), whose semantic representation was discussed in detail in chapter II, as well asKipper et al. (2000) work along the lines of Jackendoff in that they aim to explore syntactically relevant, universal meaning components and define verb classes on the basis of lexical semantic representation. The VerbNet electronic lexical database follows and extends Levin's classification by defining verb groups according to semantic meaning components and diathesis alternations. This database is conceived for NLP purposes and pays special attention to learnability criteria. Hence, the syntactic/semantic information is always made explicit and for a better learnability, generalizations are made at a higher level (e.g. compared to FrameNet). Verbal arguments are typified using thematic roles.

IV.2.2 Automatically Constructed Verb Lexicons

Automatic acquisition of subcategorization frames presents a less time-consuming alternative to manual construction of lexicons. Automatically built lexical databases might be less precise than manual ones, but they often have a better coverage (especially for a specific domain). Different sources can be used for lexical acquisition. The most straightforward method is to use an electronic version of already existing, manually edited mono- or bilingual dictionary (Briscoe et al., 1990; Sanfilippo and Poznanski, 1992). Nevertheless, dictionaries for human use are often not formalized and are not explicit enough: they tend to rely on the readers' common sense and on other extra-linguistic knowledge. A more robust approach is to extract subcategorization information from large corpora (Manning, 1993; Brent, 1993; Briscoe and Carroll, 1997; Korhonen, 2002; Messiant, 2010). A first prerequisite to the acquisition task is to formulate a hypothesis about which quantifiable features, to be found in the corpus, represent the lexical information to be extracted. If we aim to extract complementation patterns, a viable hypothesis can be that complements differ from adjuncts with respect to specificity: in other words, the complement will be significantly more frequent in the context of its governing predicate than we would expect from other sentences in the corpus. Adjuncts on the other hand will have a uniform distribution across the verbs in the corpus. Thus, most of the subcategorization acquisi-

tion methods rely on co-occurrence frequencies of predicates and complement candidates and use different statistical tests (eg. binomial test, t-test) to prove that the verb and the complement candidate are not statistically independent. The annotation level of the corpus is also a determining factor. Having a syntactically annotated corpus/treebank at our disposal makes it possible to rely on precise and explicit complementation information (the nature of this information depending on the language model behind the annotation) (Kupsc, 2007; O'Donovan et al., 2005). However, existing treebanks are usually too small to provide sufficient information on many of the less frequent verbs. Otherwise, we can either define a heuristic approximation of what subcategorization patterns might look like and do without syntactic annotation (Brent, 1993), or use a parser to automatically produce the desired syntactic information (Briscoe and Carroll, 1997; Korhonen et al., 2000; Preiss et al., 2007; Zarcone and Lenci, 2008).

It would be misleading to say that automatically acquired subcategorization patterns rely exclusively on the hypothesis that every constituent appearing above a given frequency threshold with the verb is its complement. Although frequency (and the implicitly underlying obligatoriness criterion) plays a central role, lexical acquisition methods always contain more or less preconception about the structure of a potential subcategorization frame. This preconception is either implied in the annotation of the corpus, or comprised in the acquisition algorithm itself: e.g. in the form of how it generalizes over individual tokens of subcategorization frames by tracing it back to an abstract frame type, or in the form of a filter which verifies whether the candidate is likely to be a possible subcategorization frame in the given language. Moreover, automatically acquired data are often considered as a first version of a lexicon and constitute the input of a post-editing work phase where experts manually revise the entries before finally including them in the dictionary.

IV.2.3 Important Aspects of Constructing Lexical Databases

When building lexical resources manually, a typical problem is raised by the incoherence between lexicon editors. This makes it very important to give a precise definition of the formal structure as well as the content of the database, in order to minimize inconsistencies. In the case of a verbal argument structure lexicon this is particularly difficult

because, as we have seen previously, there are no widely accepted definitions or tests to distinguish complements from adjuncts.

Another critical issue is how to achieve consistency. Typically, several editors work on the database at the same time, all of them potentially having different intuitions about linguistic phenomena and a different understanding of the instructions given. For these considerations, it is desirable to define unambiguous criteria and concrete tests which help to make decisions on particular, atypical cases. Manually built lexical resources often rely on syntactic tests, though pragmatic or semantic tests can also be considered (e.g. the dialogue test for Vallex). Syntactic tests are preferred over semantic criteria because they are more formal and less intuitive. On the other hand, they are language-specific, and when using several criteria, the result of one test can be in contradiction with an other.

IV.3 The Verb Lexicon

The Department of Language Technology of the Hungarian Academy of Sciences undertook the construction of an extensive, monolingual lexical database for Hungarian in 2001. The primary function of this lexical database is to support parsing: it serves to identify verbal complements and to assign them a correct grammatical function label by differentiating them from adjuncts. The database is divided according to part of speech categories and it contains category-specific syntactic and semantic descriptions. In the first phase of the project (2001-2003), the guidelines were defined and a basic vocabulary of 20.000 lemmata, chosen on a frequency basis, was processed and included. [3].

The database of complement structures and other morphosyntactic and semantic properties is centered around verbal argument structure. However, in order to correctly assign complement structures to verbs, the features which specify the morphosyntactic and semantic properties of complements in verbal subcategorization had to be encoded in the lexical representation of the lemmata which can fill the complement position. Moreover, as we will see in the examples, the database contains syntactic descriptions which are not directly connected to the argument structure, but they are pieces of lexical information

[3]The author of the present thesis took part of this work as a coordinator in charge of conceiving the structure and the coding guidelines of the verb lexicon presented in IV, as well as coordinating the lexicon editors' work.

required for rule-based syntactic analysis of Hungarian. Throughout the construction of the database, there was a strong intention to keep it as independent as possible of any potential application domain, so that it could be adapted to any language technology application which requires syntactic analysis of Hungarian (e.g. information extraction, automatized content analysis, machine translation etc).

The vocabulary of the database was chosen on the basis of frequency in the Hungarian National Corpus (Váradi, 2002): the most frequent 20.000 lemmata, of which 2.800 verbs were included. All of these verbs figure in The Concise Dictionary of the Hungarian Language (Magyar Értelmező Kéziszótár) (Pusztai, 2003), which was used to discriminate senses for verbs in the basic vocabulary. The table below shows the distribution of POS categories in the basic vocabulary.

POS	lemmata	entries
verb	2.964	8.865
noun	11.325	11.325
adjective	2759	3024
adverb	639	639
conjunction	74	74
pronoun	162	162
postposition	85	85
ALL	20.008	26.174

Table IV.1 : Basic Vocabulary of the Lexical Database

In the second phase of the work, the lexical database, including the verb lexicon, was extended to cover the vocabulary of a 150.000 words corpus of short business news, being used for developing an information extraction system for Hungarian (Prószéky, 2003). About 3.000 records were added to the verb lexicon then. By the end of the year 2003, the complete verbal vocabulary and subcategorization frames occurring in the Szeged Treebank (Csendes et al., 2004) was covered by the lexical database.

Between 2004 and 2007, the verbal argument structure database was used in a project which aimed at developing a Hungarian-English machine translation system (Metamorpho: http://webforditas.hu). The lexicon was integrated to the source language analysis and the bilingual module. By the end of the project, the number of different verbal argu-

ment structure patterns exceeded 30.000. The applications above are shortly presented in 5.1.2; for a detailed description, the reader is referred to the bibliography.

IV.3.1 Using the Compositionality Criterion

The two most important considerations at sight were that the lexicon had to be consistent and as domain- and application-independent as possible. However, this independence was seen as an ideal goal and not a criterion, and was not to be favored at the expense of consistency. An average of six people were working on the coding in parallel, therefore, it was important to give them unambiguous instructions. As we have shown in the previous chapter, there are no reliable and generally applicable complement tests for Hungarian. Komlósy (1992) admits that his tests are not sufficient to ensure a clear decision about every particular case. Rákosi (2006) argues that a threefold distinction is minimally needed instead of the argument/adjunct dichotomy, while Kálmán (2006) claims that there is a continuum between productive and lexicalized phenomena, thus, distinct classes cannot be established. The configuration-based syntactic complement tests, as shown in III.1, do not apply to Hungarian, neither they have any equivalent in Hungarian. It can be thus concluded that no purely syntactic tests can be given as criteria to distinguish between complements and adjuncts in Hungarian. This is why we decided to use a method based on compositionality as defined in III.3.

The compositionality criterion relies on the idea that adjuncts have to be defined previously to complements. There are several differences between typical complements and adjuncts in their usage. The typical adjunct is fully productive, has a compositional meaning, and its syntactic-semantic relation to the verb is constant. At the syntactic level, adding a typical adjunct to any verb (phrase) does not change the syntactic distribution of the phrase. Typical complements on the other hand are obligatory, unpredictable (not productive) and non-compositional at the semantic level. However, this is only true in prototypical cases. There is a continuum between the two extremes, and many of the syntactic dependents occupy a place in-between typical adjuncts and typical complements: they are characterized by certain, but not all of these typical syntactic/semantic properties. Among these features, compositionality will be distinguished as the feature which separates adjuncts from complements. A [verb + syntactic dependent] structure

is considered as compositional — and an instance of adjunction, consequently — when the case suffix on the syntactic dependent can be assigned an independent meaning. An independent meaning corresponds to a semantic role which is systematically attributed to the case-marked constituent when used with the members of a semantic verb class. Such structures are generated by adjunction rules which operate on members of verb classes. Complements, on the other hand, were defined as constituents which are syntactic dependents of the verb, attached to it in a way to form a syntactic unit which is semantically non-compositional, i.e. the form+meaning unit they produce is not predictable, cannot be generated by application of productive rules. The meaning of individual verbs differ in their core meaning (see III.3.2), which is an idiosyncratic, lexical property. Thus, semantic arguments of verbs are the constituents whose semantic relation to the verb is defined by the core meaning of the verb.

Since we did not dispose of a complete description of adjunct types and verb classes, we could not simply test whether a given complement candidate belongs to a wider group covered by an adjunction pattern. Our goal was to provide an extensive database of complementation patterns, as coherent as possible with respect to our definition, and further enrich it with relevant lexical semantic information. The compositionality criterion was therefore used as follows. During the first phase of coding (i.e. the coding of the 2.800 verbs chosen by frequency), adjunct classes were dynamically suggested in parallel to the construction of the database. These classes were later to be investigated according to the methodology described in V. The sources of information used by coders were

1. the Concise Dictionary of the Hungarian Language as a sense inventory,
2. the Hungarian National Corpus as an inventory of syntactic contexts,
3. the coders' intuition to decide which context corresponds to which sense.

To decide about the complement status of a given constituent, coders were asked to perform a test for presumed compositionality. The test consisted in replacing synonyms (or, if there were no synonym for the verb, the semantically closest, but morphologically unrelated verbs) to a sentence with the given complement candidate: if the replacement yields a grammatical sentence and the constituent with the same case suffix has the same semantic role, the constituent is a candidate for being productive with a semantic

verb group. As a second test, several semantically unrelated verbs had to be used in a sentence (where the broader context can be different, but the case and the lemma of the potential complement had to be identical), with the constituent having the same semantic role. If the second substitution yielded a correct sentence, the constituent was considered as an adjunct and was not added to the subcategorization frame. Whenever the first substitution suggested that the structure could be partially productive, but the second substitution failed, we examined whether the scope of verbs allowing the syntactic structure and the semantic role could be easily captured. If it was the case, a new verb class was suggested to be added to the verb database designating the syntactic/semantic verb group characterized by the appearance of the given constituent; on the other hand, the adjunct itself was not added to the verbal entry. In practice, adding such lexical semantic/syntactic features during the coding phase of the database was very limited. As our methodology requires verb classes to be anchored to syntactic alternations in order to ensure that we are dealing with both semantically and syntactically relevant properties, class-specific syntactic alternations need to be defined. In case of serious doubt about whether the verbs allowing the constituent form a natural class, we preferred to code the constituent as a complement. The verb groups we added in parallel to the coding were either composed of verbs sharing morphologically explicit features, or were well-known and well-studied verb groups (different types of auxiliaries, verbs of motion). Each of these groups involve one or more implicit presumptions about their complementation patterns, and the explicit description of these features was added to the documentation.

During the coding of the database, we kept in mind that it served two basic functions: first, an extensive, high/coverage and well-structured lexicon was needed for Hungarian NLP applications which require a deeper level of parsing; second, the project aimed at collecting structured data about verbal subcategorization in Hungarian. For the purpose of parsing, we aimed at a reasonable balance between a desired compact representation with well-motivated generalizations and a good coverage of data. However, finding the ideal level of generalization is always a difficult task when semantics is involved in linguistic representation. As to the second goal, the coding of the lexicon corresponds to a first step, i.e. the 'flat' enumeration of subcategorization frames, which, in turn, was followed by two experiments on enriching the structure of the database while maintaining its in-

formational content. The structural changes to be introduced include linking alternations together, defining semantic verb classes based on these alternations and associating semantic roles to class-specific complementation patterns. These projects will be discussed in V and VI.

Our hypothesis is that the different behavior of adjuncts and complements boils down to different semantic role assignment: whereas the role of adjuncts is assigned via the semantics of their case suffix, the semantic role of complements is assigned by the core meaning of the verb in a not transparent way. With respect to both categories of constituents, the case suffix can be considered as a marker of the syntactic/semantic role. However, as we will see later in this chapter, unambiguous identification of syntactic roles by case suffixes poses some difficulties. On the one hand, there are complements without case (because they belong to an uninflected grammatical category): some of them need to be coded in the lexicon, since their syntactic behavior is unpredictable, whereas others can be considered as productive and delegated to adjunction rules. On the other hand, there is a set of verbs which allow or require NPs that can take several different morphological markers. This corresponds to the idea of "thematically bound" complements in the LFG analysis of (Komlósy 1992): the verb obviously requires or allows for a certain semantic role but does not assign a morphosyntactic form to the NP which fills this role. Following our definition, these NPs have to be adjuncts since their case suffix has a meaning on its own: otherwise, it would not be possible to choose between different suffixes. The typical example is given by verbs of motion which can be extended by NPs (or adverbs) to express the content 'where to' or 'from where'?

Finally, one could notice that our test is somewhat similar to Komlósy's third complement test, cited in III.3.2 :

If a word X has an expansion Y, and there is a word Z which can systematically replace X+Y, and can replace X when X is not expanded by Y, but cannot replace X when Y is present, then Y is an optional complement of X.

In our test, we did not apply the criterion "Z can replace X when X is not expanded by Y". The main difference, however, is that KomlÃşsy's test does not define which verbs have to be used in the substitution; if we try to apply it rigorously, it leads to the result

that if in any case a substitution fails, then the constituent is not fully productive and hence is not an adjunct. Our approach is the inverse of what he suggests: if the replacement is possible, it already gives a clue about the adjunct status of the constituent. What is at stake here is not a simple terminological issue but the consideration that adjuncts are rarely totally productive. On a more practical level, including every partially productive phenomenon in the lexicon does not yield a compact and reasonable representation and we gain by generalizing over verb classes.

A basic assumption is implied in the structure of the database: the number of complements was maximized in four. No counter-examples were found during the construction of the database. Although the group of verbs of motion can appear with a frame which includes five syntactic dependents, most of them are productive with this set of verbs:

A szomszéd leküldte a fiát a szobából a kertbe meggyet szedni.
The neighbor sent out his son from the room to the garden to pick cherries.

However, the infinitive appears regularly, hence predictably with the motion verbs which allow extensions of the type from where/to where?. Similarly, being a motion verb implies that the optional from where/to where? extensions can freely appear in its context.

IV.3.2 Feature Set

The primary function of the verb lexicon is to be usable for lexicalizing a sentence parser for Hungarian. The focus is therefore on the correct analysis of any potential input sentence, not the predication of the grammaticality or the likeliness of a given input. Hence, the features in the lexicon serve to annotate verbs and their complement structure preferably in an unambiguous way in the sentence. The principal aspect in the design of the feature set is to minimize ambiguity by allowing the distinction between verb frames (with respect to specificity, syntactic structure (i.e. the type of dependency relation between the verb and other constituents), semantic roles and meaning – as far as possible.) We have to assume that each feature used in the lexicon will be available or can be made available to the parser at the exact phase of the parsing where it will be needed. In order

to insure this availability, some other unpredictable lexical features, though not directly connected to verbal argument structure, were encoded in the database. Typically, these are dependency relations which hold between NPs within the complementation frame.

An important principle behind the coding scheme is clarity: an unambiguous representation has to be achieved for subcategorization frames. What this implies is that we did not aim at making distinctions on the unique basis of verb meanings: if a given subcategorization frame can correspond to several meanings of the verb, it was coded as a unique entry, underspecified for meaning. However, when it was possible to distinguish meanings by using meaningful (morphological, semantic) restrictions, they were integrated into the more specific frame. A large set of morphosyntactic features, based on the morphosyntactic description of the Hungarian National Corpus (Prószéky and Tihanyi, 1992; Oravecz and Dienes, 2002) were used to describe the verb and its complements, as well as a more restricted set of semantic features to encode semantic selection of complements. Morphological features of the verb should not be needed, since the lexical database specifies lexical properties of verbal lemmata, and such properties are supposed to be independent from the specific morphological form of the verb in the sentence. However, some idiomatic syntactic expressions only appear with certain elements of the verbal paradigm. They have to be coded in the lexicon in this specific morphological form, because the syntactic structure only exists with the verb in the specified morphological form:

Ennek a tervnek lőttek.

this+DAT the plan+DAT shoot+past+pl+3

This plan is ruined.

*Én lövök a tervnek.

I+NOM shoot+sg+1 the plan+DAT

I shoot to this plan (no idiomatic interpretation)

IV.3.2.1 Structure

The database is divided into fields, corresponding to the function label of the constituent: VERB, SUBJECT, OBJECT, COMPL1 and COMPL2. These grammatical functions constitute the output labels. Inside each field corresponding to a constituent in the complement structure (i.e. each field except for VERB), there is a set of features subdivided into the following categories: POS (category), CASE (morphosyntactic features), SEM (semantic features) and OPT (optionality). The POS category and the CASE feature(s) unambiguously identify the complement inside the complement structure. In our feature set, we use features interpreted as conditions, and other features interpreted as value assignments. By default, most features correspond to a condition on the input structure: if every condition formulated by these features is met for the complete argument structure, the elements in the structure can be labeled according to the function labels specified in the fields. Some condition features serve to override the default characteristics associated to a given category or structure or even to relax certain conditions (e.g. about the optionality of an element or the underspecification of certain default features). Features interpreted as a value assignment are supplementary output features which are coded in the lexicon because they are not predictable; an example of such value assignments is the +subject_subord= feature that provides information about coreference relations in control structures.[4]

The subcategorization frames are associated to verbs in a flat structure. Unlike e.g. in the Lexicon-Grammar tables, we did not presume the existence of a 'base frame' from which other structures are derived; instead, on a theoretical level we rely on the notion of alternations. In practice, alternating structures were always included whenever this affected the surface syntactic realization of a complementation pattern. The prediction and linking of alternations requires further study. Syntactic operations such as negation, focusing, emphasis by topicalization or contrastive topic are presumed to be fully productive and are not coded.[5]

[4]For a detailed explanation, see the coding of propositional complements later in this chapter.

[5]It is possible, however, that certain complement structures with a highly lexicalized form and semantic content do not allow some/any of these operations. On the other hand, as we focused on parsing and not generation, this was not a concern.

IV.3.2.2 Optionality and Ambiguity

A difficult subtask in parsing is choosing between several parse trees, produced by abundant grammar rules in the system and by structural homonymy in natural language sentences. Humans use their linguistic competence as well as general world knowledge to disambiguate between parses. NLP applications usually handle this task in a separate module, either by using frequency/probability information about the structures or by different heuristics (e.g. always choosing the more specific rule or the one that matches the longest sequence). These approximative methods include the possibility of error. Accordingly, reducing structural homonymy is an important aspect that we decided to take into consideration in the design of the lexical database. Each record of the database was also assigned a unique identifier, and the base vocabulary (the 2.800 verbs chosen on a frequency basis) were assigned meaning descriptions (informal, textual definitions) from the Concise Dictionary of the Hungarian Language. These meaning descriptions were not assigned to verbs but to verb + subcategorization frame entries. The meaning description is not unambiguous: a verb frame can be associated with more than one meaning. In other words, the records were distinguished by syntactic differences and not by meaning. Each record of the same verb had to be distinguished from other records by at least one conditional input feature: it was a declared objective not to create supplementary analyses by splitting syntactically identical complementation frames which only differ semantically. Although we know that in some cases, the same morphosyntactic surface realization can correspond to different semantic structures (in terms of semantic role assignment), and hence they be further expanded by different types of adjuncts [6], the records in our database our underspecified with this respect. In order to be able to choose the matching complementation pattern in an algorithmic way and eliminate ambiguous parses as much as possible, we need to avoid creating records which only differ from each other in their output semantic description.

Multiple analyses are often due to the big amount of optional complements in Hungarian. Subject and direct object are optional in almost every sentential context. Many of the other complements are also optional, it is thus logical to introduce a distinction between obligatory and optional complements. On the other hand, the interpretation of

[6] see V.2. for examples

this information by a parsing algorithm is far from being trivial. Complement structures with one or more optional complements may coincide with shorter structures:

a) A fiú feladta nekünk a szerszámokat.
The boy handed up the tools for us.

In the sentence above the complement frame for the verb (felad, 'to hand up/over') includes the subject (a fiú, 'the boy'), the direct object (a szerszámokat, 'the tools') and an optional complement which is either a beneficiary argument (nekünk, 'for us') or a location argument.

b) A fiú feladta a szerszámokat.
The boy handed up the tools.
The boy gave up the tools. (e.g. the idea/intention of buying, using the tools)

c) A fiú feladta a függetlenségét.
The boy gave up his independence.

The shorter complementation frame in these sentences consists of the subject and the direct object (a szerszámokat, 'the tools' a függetlenségét, 'his independence'). Structural ambiguity is present in sentence b), since it can be interpreted as 1)either corresponding to the same complementation frame as in sentence a), but with the optional beneficiary/location complement omitted, or 2)corresponding to a full, shorter complementation pattern with the meaning 'give up'. The second interpretation is less likely, since felad in the meaning of 'give up' requires an abstract direct object. Nevertheless, a meaning shift can take place and result in a context-dependent abstract interpretation (e.g. 'give up the idea of buying/working with/etc. the tools'). Whereas it is a phenomenon of real structural ambiguity, in any potential application a parser will have to choose the most likely analysis of a sentence, including the identification of the verbal complement frame. This decision can be facilitated not only by stochastic methods but also by heuristic ones, grounded on human judgements. By taking structural ambiguity into account in the de-

sign of the database, we can give a preference to certain syntactic structures over other, less likely ones. The method for doing so was by using the binary obligatory/optional feature: when two subcategorization frames only differ in an optional complement (either two frames with the same length, with the only morphosyntactically different element being optional in both frames, or one shorter frame and one containing an optional further complement: in both cases, the obligatory complements being the same), there is the possibility to give preference to the more likely frame by marking the optional extension of the less likely frame as an obligatory element. This can be interpreted as enforcing the matching of the more likely frame, unless the optional complement appears in the sentence, which indicates unambiguously that we are dealing with the other structure. This method allows to define a "default" interpretation, which is particularly important when we want to associate a meaning (or a translational equivalent) to the structure using its unique identifier. Restrictions on optionality and default interpretations were widely used during the integration of the database in the machine translation system (see VII.2).

IV.3.2.3 Specificity and Multiword Expressions

Besides the optionality of complements, another source of ambiguous interpretations comes from idiomatic complement structures, i.e. multiword expressions (MWEs) with a verb. They usually have a "literal", compositional interpretation, and a different one which is idiomatic. People usually choose the correct interpretation without any difficulty, since a multiword expression "can only be used to convey its idiomatic meaning if there is sufficient thematic incongruence between the topic of the discourse and the compositional meaning of the MWE." (Oravecz et al., 2005). A classification was established for Hungarian MWEs by Oravecz et al. (2005). They differentiate the following types of MWEs:

1. institutionalized structure (e.g. *washing machine*): their meaning and syntax is compositional, but their constituents cannot be replaced by synonyms,

2. MWEs with a semantically empty function verb (e.g. *make a decision, take part*): the meaning is not compositional since the verb is empty,

3. verb+particle structures (e.g. *létre+hoz, 'bring into existence')*,

4. semi-transparent structures which can be modified according to the idiomatic meaning (e.g. *(komoly) bakot lő 'make a (serious) mistake'*, *literal: 'shoot a (serious) billygoat')*,

5. not transparent structures which cannot be modified at all (e.g. *felveszi a kesztyűt 'accept the challenge'*, *literal: 'put on gloves')*,

6. completely fixed (e.g. *lássuk a medvét! 'let's see the bear!'*).

Disambiguation between the lexicalized idiomatic usage and the productive usage of the same syntactic structure usually requires knowledge about the extra-sentential semantic context. In other cases the idiomatic use of a verb frame goes together with syntactic specificities, usually constraints, which help to identify the frame as a multiword expression. Although some expressions can be identified by morphosyntactic constraints, other lexicalized verb frames can have ambiguous interpretations. The coding guidelines we elaborated states the underlying idea that more specific patterns (e.g. a verb frame which includes a lexicalized noun complement) have the priority over more general patterns (where the complement slot can be filled by any case-marked noun). This implies that we have to have recourse to the non-idiomatic analysis of the structure if there is no more specific frame which matches the input. This hierarchy is an important feature of the lexicon, since complement structure descriptions show different degrees of specificity: there is a large interval between very general, underspecified patterns to completely lexicalized collocations.

As shown by the examples to follow, the structure of the patterns and the mode of specification is largely influenced by the targeted application. The following examples illustrate the different kinds of information and their encoding in light of 1) parsing Hungarian sentences and 2) building a machine translation system.

In the most general argument frames, the subcategorization of the verb contains only morphosyntactic identification, i.e. the case suffix of its arguments. There are no further restrictions, besides general rules of complement realization, to constrain the syntactic realization of the complementation pattern. E.g.:

Valaki ad valamit valaminek.

Someone gives something to something/someone.

The range is wide between this and similar patterns and morphosyntactically and/or semantically constrained, lexicalized collocations. Morphosyntactic constraints can be made to the verb or to the complements. For instance, the determiner of the complement NP, the number of its head noun, the tense or mode of the predicate in the subordinate (complement) clause, or the choice between active/passive voice can be constrained in complementation patterns. E.g.:

Constraints in number:

Valaki vitatkozik valakivel. / Valakik vitatkoznak.

Somebody argues with somebody. / Some people are arguing.

The verb vitatkozik (*argue*) has a complement noun in instrumental case (-val), which becomes optional — for obvious semantic reasons — if and only if the subject and the verb itself are in plural. Since subject and verb agree in number, the plural has to be coded only once in the argument frame – preferably in the morphosyntactic description of the verb, as subjects are always optional.

V.vitatkozik+number="PL" SUBJ

V.vitatkozik SUBJ COMPL1+case="INS"+obl

Morphosyntactic constraints in lexicalized structures can refer to negation or passive (formed by an adverbial participle and a copula):

Predicate in passive: V.utal+passive

Valaki valakire van utalva.

Somebody depends on somebody.

Negated predicate: V.nyugszik+negated

Valaki nem nyugszik, amíg ...

Someone will not rest until ...

In a monolingual setting, it is not necessary to encode information about the internal structure of complement NPs as long as they are regular and predictable. However, collocations with lexical restrictions often show specificities which cannot be derived from more general patterns, mostly with respect to the semantic relations between the verb and its arguments in such collocational frames. Syntactically regular and hence seemingly compositional structures, e.g. possessive structure, are assigned with semantically non-compositional roles, as in the sentence below:

Valaki az agyára megy valakinek.

Someone the brain+poss:sg3+SUP go+sg3 someone+DAT

Someone drives someone crazy.

The collocation *agyára megy* forms one semantic unit with the meaning 'drive sy crazy', 'make sy nervous'. This semantic unit, syntactically represented by the verb + NP complement structure, is responsible for assigning the semantic role to the NP possessor of *agyára* ('brain+POSS+SUP'). Although the structure is syntactically regular, it still has to be included as a separate entry since it cannot be derived from the more general *"valaki megy valahová"* (*someone goes somewhere*) argument pattern. Another group of lexicalized patterns is formed by seemingly and intuitively regular possessive structures in which the possessee is a bodypart. The specificity of these complementation patterns, as it turned out during the development of the machine translation system, is that the possessor appears in the English sentence in a different syntactic role, and the possession is not expressed syntactically:

Valaki megveregeti valakinek a vállát/hátát.

Someone pat+sg3 someone+DAT the shoulder/back+ACC+poss:sg3

Somebody pats somebody on the shoulder/back.

Valaki megragadja valakinek a kezét/karját stb. Someone catch+sg3 someone+DAT the hand/arm+ACC+poss:sg3 Somebody catches somebody by the hand/arm.

Both Hungarian complement structures are translated to English by deleting the syntactic encoding of the possessive, while the Hungarian possessor appears as a direct object in the English sentence. Despite the syntactic and semantic regularity of these sentences in Hungarian, these argument structures will have to be lexically coded in a bilingual application, because they are realized differently in the target language. This phenomenon illustrates that we need to consider the influence of the structure and lexical specifities of the target language when building lexical databases for bilingual (or multilingual) NLP applications.

Besides morphosyntactic and lexical restrictions, the scope of argument patterns can be narrowed by using semantic features. A small set of basic semantic features were integrated (e.g. abstract, animate, time, measure). The most important function of semantic features is to promote disambiguation between verb senses without introducing supplementary syntactic analyses for otherwise identical structures. In machine translation, semantic descriptions make up the most efficient tool for differentiating between verb senses. This means that semantic features allow to formalize semantic selection of verbal arguments, and hence to assign different translations to syntactically similar, but semantically different complementation patterns. The example below illustrates how the semantic features of our lexicon were exploited for sense disambiguation in the Metamorpho translation system (see VII.1) :

Valaki elvesz valamit.

Somebody takes something.

HU.VP = SUBJ + TV(:lex="elvesz") + OBJ(pos=N, case=ACC, human=NO)

EN.VP = SUBJ + TV[lex="take"] + OBJ

Valaki elvesz valakit.

Somebody marries somebody.

HU.VP = SUBJ(human=YES)+TV(:lex="elvesz")+OBJ(pos=N, case=ACC, human=YES)

EN.VP = SUBJ + TV[lex="marry"] + OBJ

The most specific complementation pattern type puts lexical restrictions on constituents in their complement positions. Lexical constraints can have multiple functions in the database. First, they can be used for word sense disambiguation, in the same principle as semantic features. The lemma of a complement can help to identify the meaning (and the translation) of the verb and, likewise, the verb can serve as a context feature to choose the meaning of the complement noun. For instance, the noun jelentés has two basic meanings: 1) report, document and 2) meaning. As a direct object of the verbs megír, elolvas (*read, write*), it can only figure with the meaning 1). The same pattern can disambiguate the verb and its complement, e.g. in "Süt a nap" (*The Sun shines*): a verb complementation pattern, specified at a lexical level, could further serve to disambiguate nap (Sun/day) and süt (bak/shine). We can talk about word sense disambiguation when lexical entries have more than one associated meaning (stored a meaning inventory of the given language) from which we choose one according to the context the lemma occurs in. The structures in which the verb and its complement together take a meaning which is completely non/compositional, i.e. independent of the lexical meaning of both constituents, can be considered as collocations. Collocations are usually characterized by bound structure, with reduced possibilities of syntactic variance: not only the head noun, but the whole complement NP is lexicalized. E.g.:

*Valami rossz/*hibás/*negatív fényt vet valamire.*
Something reflects bad/*false/*negative light on something.
Something makes something look bad.

Valami veszendőbe megy.
Something goes into the lost.
Something is wasted.

Verb+complement collocations are often translated by a single verb (the nominal argument does not have an equivalent in translation), which supports the claim that

collocations typically correspond to a single semantic unit which cannot be decomposed.

IV.3.3 Grammatical Functions

IV.3.3.1 Morpohosyntactic Description of Complements

Not only subcategorization frames, but also individual complements inside the frames have to be identified individually. It is thus necessary to provide a unique identifier if we need to provide further specifications, e.g. semantic features or lexical constraints about arguments. The output of a parser that uses this database will have to assign unique grammatical function labels. Hence, complements within the same subcategorization frame have to be distinguished. Moreover, in NLP tasks which require deep parsing and semantic interpretation, semantic argument roles will be associated to NPs on the basis of their specific complement functions. As it was described previously in this chapter, we suppose that syntactic complement functions and, accordingly, semantic roles are encoded in Hungarian syntax by cases or postpositions. It would follow then to consider case suffixes and postpositions as identifiers of complement roles. However, we find that there are other types of subcategorized sentence constituents, considered as complements in our terminology, belonging to an uninflected POS category (e.g. infinitives, adverbs). Case suffixes are thus insufficient, the POS category also needs to be specified in order to provide a unique identification for the elements in the complement frame. In what follows, we will explain the difficulties about identifying complements by grammatical category and case information. According to our basic hypothesis, POS category and case suffix/postposition are the minimal properties defined in verbal subcategorization (further restrictions are possible but not compulsory). Those constituents which are not NPs and hence do not have a case will be identified uniquely by their POS category. For instance, the adverbial complements in the sentences below are syntactically obligatory with the verb *bánik* ('treat', 'handle').

János jól bánik a mostohafiával.
John well treats his stepson+INS
John treats his stepson kindly.

Az igazgató felelőtlenül bánik a pénzzel.

The director irresponsibly handles the money+INS

The director mismanaged the money.

In linguistic theory, it is an open question whether constituents such as *jól* (*well*) or *felelőtlenül*, (*irresponsibly*) receive a thematic/semantic role from the verb and if they do not, whether they can be considered as complements. When creating a computational lexicon, however, it is not a question that every piece of unpredictable information has to be coded.

Most typically, the case suffix in itself is sufficient to identify the constituent with the specific semantic role and complement function. In such cases the only constraint the verb puts on its complement is its morphological case, there are no other formal restrictions on the complement. Constituents of any grammatical category can fill the given function, as long as they can be inflected for case (nouns, adjectives, pronouns and *hogy* (*that*) clauses). For instance, the verb *vár* (*wait*) has a complement in sublative case:

NP:

János a vonatra várt.

John the train+SUB waited

John was waiting for the train.

PRO:

János csak erre várt.

John just that+SUB waited

John was just waiting for that.

ADJ:

János egy jobbra várt.

John a better+SUB waited

John was waiting for a better one.

CLAUSE:

János arra várt, hogy a vonat megérkezzen.

John that+SUB waited that the train arrive

John was waiting for the train to arrive.

The verb only prescribes the case of its complement: any category, as long as it can be inflected for case, can fill the position. It is a general syntactic rule in Hungarian that NPs can be replaced by adjectival phrases or pronouns. Thus, this was considered as the default case, and such complements were coded as NPs (additional records were not added for the other categories). Hogy clauses are also marked by case suffixes: they have a nominal antecedent which displays the suffix. However, they are less free to replace nominal constituents in a complement position, because verbal meaning largely determines whether it is semantically compatible with a propositional semantic content. As the examples illustrate, this does not depend on the case of the complement:

a) A professzor egy viccel fejezte be az előadást.

the professor a joke+INS finished the lecture

The professor finished the lecture with a joke.

b) A professzor azzal fejezte be az előadást, hogy elmesélt egy viccet.

the professor that+INS finished the lecture that told a joke

The professor finished the lecture by telling a joke.

c) Szíŋvesen cserélnék Jánossal.

gladly would-change+sg+1 Jonh+INS

I would gladly change (lives) with John.

d) *Szíŋvesen cserélnék azzal, hogy...

gladly would-change+sg+1 that+INS that

I would gladly change to being ...

It is impossible to finish sentence d) and obtain a grammatical sentence, as any proposition expressing a state of affairs would be semantically incompatible with the verb *cserél* ('change'). It is a lexical property of the predicate whether it allows a complement clause with the conjunction hogy; accordingly, unlike nominal complements, clauses cannot replace NPs in any sentence by default.

IV.3.4 Nominal Complements

Nominal phrases (NPs, AdjPs, pronouns) always have a morphological case in a Hungarian sentence. NPs with postpositions are usually in nominative. The postposition is coded in the lexicon the same way as the case suffix for the NPs without postposition. Some postpositions in turn take NPs with other cases (e.g. vmivel szemben – *opposed to sg*, vkinek a részére – *for sy*, vmin túl – *beyond sg*). When such NPs occur in a subcategorization frame, the case of the head noun has to be stored lexically because it contradicts the default interpretation. However, it is a lexical property of the postposition that it takes an NP with a given case suffix. Hence, this piece of information has to be stored not in the verbal lexicon but in the postpositions' lexical entry. A syntactic analyser would need that information at the stage of the identification of NPs in the sentence, previously to matching verbal complement frames. Whereas in most cases specific case suffixes or postpositions are required to mark the complement relation between the verb and its semantic argument, some verbs have complementation patterns in which a complement function can be filled by morphologically variable constituents. Their case suffix or postposition can be freely chosen from a set of morphological markers. The prototypical example are locative complements (where to / from where?) occurring with verbs of motion. Komlósy (1992) describes these constituents as thematically bound complements, which means that the verb only constrains their thematic role but not their morphosyntactic form. Accordingly, while building the verbal lexicon, we identified such complements by semantically motivated labels instead of listing all the possible morphological markers. In parallel, the correspondence between these semantic labels and morphosyntactic markers was included in the database. Thematic binding is also evidenced by the fact that adverbs can also fill these positions (oda – *there*, direction; messziről – *from far away*). E.g.: dob (*to throw*

sg + from somewhere + to somewhere):

V+dyn

COMPL1.POS="ADV/NP/POSTP"+to

COMPL2.POS="ADV/NP/POSTP"+from

Other constituents often considered as thematically bound complements are those answering the question *where?*. Komlósy (1992) claims that they enter in a complement relation with verbs which require their presence to form a grammatical sentence (except in elliptic structures, where their meaning is implied by textual or extra-linguistic context), e.g.:

Picard kapitány jelenleg a hajón tartózkodik.

Captain Picard is currently located on board of the ship.

Verbs in this group (e.g. lakik, elhelyezkedik, tartózkodik – *live, stay, be located*) typically occur with a locative NP or adverb. However, despite syntactic obligation, we have no reason to include such structures as complementation patterns because they do not meet the compositionality criterion. Almost every verb can take a location adjunct of the same form, and the above cited examples do not differ from productive adjunction neither in surface structure nor in semantic content. Moreover, the database is intended to be used for parsing and not for generating or predicting Hungarian sentences. Hence, syntactic obligation does not compel us to include these entries. We saw this concept confirmed during the creation of the bilingual database within the machine translation project. We did not need to formulate any translation rules which would have referred to these constituents within the verbal argument frame. They are translated to English in the same way when they occur as optional adjuncts as they do in a syntactically obligatory position: otherwise, in case of a non-predictable translation, the source-language side of the translation rule always refers to a complement with one specific case suffix. Other special difficulties raise when we try to identify subjects and objects unambiguously on the sole bases of case suffixes. Copulative verbs (van – *be*, lesz – *become*, marad – *stay* etc.) can take more than one complements in nominative case. Similarly, there are two

nominatives in the structures like Juli elmúlt tizennyolc (éves)/ tizennyolc éves múlt – *Julie is over eighteen (years old).* The subject of these copulative and pseudo-copulative verbs can be identified according to its agreement with the verb in number and person: thus, the constituent which agrees with the verb is taken to be the subject and coded as such in the lexicon. Hungarian verbs have two agreement paradigms, traditionally referred to as subjective and objective. The choice between these paradigms is affected by the nature of the direct object, namely its definite/indefinite character. In very general terms, verbs display objective inflections when they have definit NPs as direct object, whereas indefinite NPs go with the subjective paradigm. A set of indefinite NPs can be used as direct objects, with an adverbial or aspectual meaning e.g. kicsit (*a little*), sokat (*a lot*), egy kiadósat, (*copiously*). The presence of these pseudo-objects depends essentially on the aspectual properties of the verb and those of the intra-sentential context. They were not encoded as complements as they are free to appear even with otherwise intransitive verbs. Moreover, we need to consider the fact that in a big part of Hungarian sentences, the direct object of the verb is absent. Therefore, we cannot encode pseudo-objects as optional objects to every verb which allows for one if we do not want to lose the distinction between transitive, optionally transitive and intransitive verbs. Syntactic analysis need to take into account definiteness agreement between the verb and its object. If there is no accusative NP in the sentence, but the verb is still in its objective paradigm, the transitive complementation pattern is matched into the sentence. However, we also need to encode some lexicalized structures where the verb is inflected according to the objective paradigm, but there is no optional object in the structure and it could not be added either:

A testvérem úgy gondolja, hogy fölöslegesen idegeskedtem.

My brother so thinks-obj that unnecessarily I-worry.

My brother thinks I worry for no reason.

A csapat már nem bírja/győzi erővel.

The team no longer keep-up-obj strength+INS.

The team can't keep up any more.

IV.3.5 Infinitives

IV.3.5.1 Infinitives with Auxiliaries

Infinitives can appear either with auxiliary verbs (the verb in infinitive being the main predicate in these cases), or as a complement or adjunct of a main verb, with a specific grammatical function and semantic role assigned by the latter. The difference between infinitives with auxiliaries and complement infinitives concerns the predicative function (i.e. the assignment of complement functions), and the semantic relation between the infinitive and the finite verb.

Let us briefly recapitulate what we have said in chapter II about the processing of infinitives within the shallow parsing process. The hypothesis was that the difference between auxiliaries and content verbs with an infinitive complement (or adjunct) is that auxiliaries only add a piece of modal, temporal or aspectual information to the meaning of the infinitive. Other top-level constituents in the sentence will thus belong to the subcategorization frame of verb in infinitive. The grammar annotates the infinitives accompanied by auxiliaries as predicates. The relevant morphological properties of the auxiliary (tense, mood, person, number, definiteness) are copied to the predicate's feature structure so that its agreement with the subject (number, person) and the object (definiteness) could be checked at the deep analysis phase. This method ensures that verbs annotated as predicates will correspond to our definition: they will be annotated with tense, number, person and other morphological features, so that the argument structure of the sentence could be correctly matched at the level of deep parsing. The set of auxiliaries is was identified according to Kálmán C. et al. (1989). In the verbal lexicon, they were simply marked by a special feature as a verb class [7]; the shallow parser already includes a grammar to deal with the resulting grammatical consequences.

IV.3.5.2 Infinitives as Complements

Besides adverbs, infinitives are the only grammatical category to appear in verbal complementation patterns without being inflected for case. Other categories were identified in the database entries on the basis of their case suffix in the complementation frame (including clause-type complements, which have a nominal antecedent inflected for case).

[7] More precisely, two classes: one for modal and one for temporal auxiliaries.

Not only infinitives are not marked for case, neither can they be preceded by prepositions or followed by postpositions. Moreover, we find that many occurrences of infinitival complements are not interchangeable with a case-marked NP which would designate the grammatical function. On the other hand, in our grammar we have to decide which grammatical role to assign to the infinitive. We have two different options for encoding infinitives in the argument structure database. First, we can consider the category of infinitive as a supplementary case, i.e. an independent element in the paradigm of morphological markers. This approach is viable but it does not allow to make generalizations over subgroups of infinitival complements. Moreover, this makes it difficult to capture similarities between infinitives and NP complements, especially with respect to semantic roles. As we will see, similarly to propositional complements, infinitives can be interchangeable with NP complements in a set of complementation patterns. The significant difference is that unlike clauses, infinitives cannot be marked for case or modified by postpositions, which necessitates the use of other distributional criteria. Based on distributional analysis, I defined a set of syntactic criteria to determine the grammatical function of infinitives in verbal complement structures.

1) Infinitives coded as subjects

Infinitives will be considered as subjects when a) there cannot be any other potential subject (NP in nominative case) in the sentence, and b) the verb is invariably in the 3rd person of the singular. The verbs occurring in this structure mainly correspond to modal auxiliaries, and the subject complement of the infinitive is embodied by an NP in dative case. E.g.:

Erre nem lehet mérget venni.

This+SUB no can poison take+INF

I wouldn't count on it.

Tán nekem nem szabad taxiba ülni?

Maybe me+DAT no allowed/free taxi+ILLl sit+INF

Am I not allowed to take a taxi?

These verbs are considered as auxiliaries and annotated as such by the shallow parser. The additional information coming from the lexical database concerns the assignment of grammatical functions. The encoding of infinitives as subjects makes it possible to generalize over this group of auxiliaries and state that this group of verbs display the semantic subject of the infinitive as an NP in dative case. This is important because, as we will see, not all the auxiliaries show this syntactic behavior.

2) Infinitives coded as objects

An infinitive has the direct object complement function when there is agreement in definiteness between the main verb and the direct object of the infinitive. This claim is based on the fact that the definiteness feature of the verb always depends on the direct object. E.g.:

Béla olvasni akar.
Béla read+INF want+sg+3+indefinite
'Béla wants to read.'

Béla a könyvet akarja olvasni.
Béla the book+ACC want+sg+3+definite read+INF
'Béla wants to read the book.'

3) Infinitives in other complement functions

The rest of the infinitival complements were classified according to their replaceability. It was systematically examined for each infinitive to be added to the database whether it can be replaced by a case-marked noun or a clause with a case-marked pronominal antecedent expressing the same semantic content. Deverbal nouns or finite versions of the infinitive were used to test these distributional properties, e.g.:

János fél elutazni.
John is afraid+sg+3 leave+INF

'John is afraid to leave.'

János fél az elutazástól.
János is afraid+sg+3 the leaving+ABL
'John is afraid of leaving.'

János fél attól, hogy elutazzon.
János is afraid+sg+3 that+ABL that leave+SUBJ+sg+3
'John is afraid of leaving.'

As shown by the examples, either a deverbal noun with a similar meaning or a clause with the verb in finite form and the antecedent ('that' pronoun) can replace the infinitive. In both cases, the nominal element (either the complement noun or the pronoun antecedent) is inflected for ablative case, which is required by the verb fél ('to be afraid'). A big set of verbs, requiring complements in different cases, are characterized by this distribution. I decided to encode the grammatical role of infinitival complements by marking them with the same case as the corresponding NPs and finite clauses, in order to capture the observation that they fill the same semantic role.

4) Infinitives with light verbs/modal verbs

Finally, a distinct set of infinitives occur with light verbs, i.e. verbs which do not have a full semantic content, they mostly expand the meaning of the verb in infinitive by an aspectual or modal component (e.g. merészkedik 'dare', szíŋveskedik 'please', kegyeskedik 'be so good as', szándékozik 'intend', habozik 'hesitate', enged 'allow', hagy 'let'). With the unique exception of habozik 'hesitate', these verbs require an obligatory infinitival complement, whereas they cannot have an NP in the same semantic position of their complement structure. Hence, the interchangeability test cannot be applied to determine the grammatical and semantic function of these infinitives. Since it is impossible to identify these infinitives to a nominal complement in the same complement structure, this group was not annotated with a grammatical function.

As it was already mentioned, infinitives appearing with verbs of motion were considered as adjuncts and not included in the database.

IV.3.5.3 Infinitives and Control Structures

Another important dimension in the encoding of infinitives is the predictability of the semantic subject of the infinitive. Productively used infinitival adjuncts (e.g. with motion verbs) appear in obligatory subject control structures: their subject refers to the subject of the main verb:

Ez elment halászni.
This one_i go+past+sg+3 $...._i$ fish+INF
This one went fishing.

Infinitives often appear with a subset of verbs of perception and form an accusativus cum infinitivo or object control structure: the semantic subject of the infinitive coincides with the direct object of the matrix verb, i.e. the verb of perception. The infinitive+ can be replaced by NPs or finite clauses with an object function. However, the presence of an explicit grammatical object in the sentence prevents us from applying this criterion.

Láttuk a gyerekeket felnőni.
See+past+pl+2 the children+ACC grow up+INF
'We saw the children grow up.'

Beside verbs of perception, a few transitive verbs allow object control in their complementation frames:

A szülők néha nem hagyják a gyerekeket önállóan dönteni.
The parents sometimes not let the children+ACC independently decide+INF.
'Sometimes parents don't let the children to make decisions on their own.'

While infinitives were encoded as complements without case, a separate referential feature was used in these structures to encode the coreferece relation between the object of the matrix verb and the subject of the infinitive:

verb SUBJ.POS OBJ.POS COMPL.POS
hagy N.NOM N.ACC INF+linked_to="OBJ"

The same reference feature was used in complementation frames with an adverbial participle, whose subject does not coincide with the subject of the main verb but with an other argument:

A szobor ülve ábrázolja a költőt.
The statue sit+ADVPART represent+sg+3 the poet+ACC
'The statue represents the sitting poet.'

By default, adverbial participles in Hungarian appear as optional adjuncts, linked to the subject of the predicate. Some verbs, however, trigger a different interpretation: an NP belonging to the complementation frame of the predicate (the direct object in the example above) is co-indexed with the participle as its subject. We are thus dealing with the same phenomenon of control inside the complementation frame.

IV.3.6 Propositional Complements

IV.3.6.1 Grammatical Function

Complement clauses in Hungarian are introduced by the conjunction *hogy* ('that'). Unlike infinitives, complement clauses are inflected for case: they have a pronominal antecedent in the form of an anaphoric pronoun in the main clause which can fill any syntactic function according to the predicate of the main clause. The anaphoric pronoun is most frequently the demonstrative pronoun *az*, less frequently the 3rd person singular form of the personal pronoun[8]. The pronominal antecedent can be obligatory or optional:

[8]This only occurs when the pronoun is not in nominative

this was encoded by a binary feature. The lemma of the antecedent is presumed to be either the demonstrative or the personal pronoun by default; as we will see later, other antecedent lemmata are possible. In such cases, the lemma of the antecedent pronoun was explicitly encoded. The case of the antecedent pronoun follows from its function (nominative or accusative, corresponding to subject or object, respectively) or is specified as the case of the clause. The form or forms of the pronoun ensue this way from the encoding of the entry. In the case of infinitives, we used a distributional test based on replacing the infinitive with a corresponding nominalization to determine its grammatical function. We do not need to resort to this test with complement clauses: instead, we rely on the case of the pronominal antecedent. CLAUSE was coded as a part of speech, complement clauses were coded in the column which corresponds to their grammatical function (subject/object/other complement, determined by the antecedent). However, pronominal antecedents are optional in some complementation frames, which makes their processing more difficult, since these surface structures may coincide with sentences containing an adjunct clause without an antecedent.

IV.3.6.2 Adjunct Clauses

Adjunct clauses are productive and not included in the database. We limit ourselves to a short presentation of the two types of adjunct clauses with the conjunction hogy (subjunctive and indicative), their importance being that complement clauses have been defined in contrast to them. When the predicate of the subordinate clause is in subjunctive mode (+conj feature), the adjunct *hogy* clause expresses purpose. They come with an optional, typically omitted pronoun antecedent: azért (az demonstrative pronoun in 'causal' case). They can be productively used with agentive verbs.

Taxit híŋvtam, hogy hamar odaérjek.
taxi+ACC called+sg+1+past that on-time get-there+CONJ
I called a taxi (in order) to get there on time.

Adjunct hogy clauses with a verb in indicative are syntactically characterized by an obligatory antecedent, the pronoun az (or, less frequently, the personal pronoun) in in-

strumental case. Such adjunct clauses denote the *mode* of an action, state, relation or property expressed by a verb. Although this adjunct clause is less frequent than the subjunctive one, it is still productive with a large amount of verbs.

Veszélyt jelent az emberekre azzal, hogy rejtegeti a terroristákat.
danger present the people+SUB PRO+INS CONJ hides the terrorists+ACC
They present danger for people by hiding terrorists.

Considering these two types of *hogy* clauses as adjunct implies that any other *hogy* clause not fitting the descriptions above need to be coded in the lexicon as a complement.

IV.3.6.3 Optionality of the Pronominal Antecedent

Pronominal antecedents are always optional (and typically omitted) when the clause fills the subject or direct object function:

Bosszant, hogy Marcsi megint nem találja a kulcsát.
annoy+sg+3 that Mary again not find+sg+3 the key+poss.sg.3
'It annoys me that Mary cannot find her keys again.'

Pronominal antecedents with a case suffix other than nominative or accusative can also be omitted, although it is less frequent:

János fél (attól) / csodálkozik (azon), hogy otthon hagyta a kulcsot.
John is-scared (pro+ABL) / is-surprised (pro+SUP), that home left the key
'John is scared about / is surprised about having left the key at his place. '

The grammatical function of the complement clause in these verbal complement patterns can be identified in the database according to the syntactic function of the optional antecedent. The resolution of structural ambiguity between a complement clause without antecedent and an adjunct clause which always come without an antecedent, on the other

hand, is delegated to parsing.

IV.3.6.4 Adverbial Antecedents

Clauses with an adverbial antecedent constitute a different category. On the one hand, adverbs usually have an adjunct function. This would lead to the implication that co-referent clauses, as they expand the meaning of the adverb, are also adjuncts. This is the case of hogy clauses introduced by the antecedent Ãžgy ('in that manner'): they can be substituted productively for mode adverbs:

Az előadó kedvesen / mosolyogva válaszolt a kérdésre.
the lecturer kindly / smiling answered the question+SUB.
The lecturer answered the question kindly / similing.

Az előadó úgy válaszolt a kérdésre, hogy közben kedves maradt.
the lecturer so answered the question+SUB, that meanwhile nice remain+past+sg+3
The lecturer answered the question, still staying nice.

However, the same adverb can also introduce complement clauses. In these cases, the clause is obligatory and has a semantic role different from mode adjuncts. This characterized a limited subset of verbs of thinking with a complement clause introduced by *úgy*:

A feleség úgy okoskodott/vélte/látta,hogy ez már válóok.
the wife so think/consider/see+past, that this already reason-for-divorce
The wife considered it as a reason for divorce.

Neither the appearance of the complement clause with these verbs, nor the semantic link between the matrix verb and the subordinate clause are predictable, since the phenomenon is not productive among verbs of thinking. A supplementary feature which underlines the lexicalized nature of these structures is the seemingly unmotivated definite conjugation of the verbs lát ('see') and vél ('consider'), despite the fact that they never

have a direct object in these complementation frames.

An even more special subset of complement clauses occur with adverbs which in general would not allow productive use of hogy clauses as extensions. These adverbs occur with hogy clauses only in lexicalized patterns; otherwise, i.e. in productive structures, the clauses linked to them are introduced by adverbial pronouns (pronouns with an adverb-like distribution). For instance, the adverb ott ('there') can be extended productively with a clause introduced by the locative adverbial pronoun ahol ('where'):

a) ott, ahol azelőtt laktunk
(there) where we lived earlier

b) *ott, hogy azelőtt laktunk
*(there) that we lived earlier

Construction a) can be used productively in sentences in the distributional context of a locative adjunct (or complement). Within the construction, the pronouns ahol also functions as a locative adjunct/complement in the embedded clause. However, the conjunction hogy cannot be substituted productively (example b)). Only a limited set of verbs allow a clausal complement in which the pronominal antecedent is an adverbial pronoun and the complement clause starts with the conjunction hogy; evidently, the semantic link between the clause and the verb is not to be interpreted as a locative adjunct:

A férj ott rontotta el, hogy nem mosogatott soha.
the husband there spoil+past prefix, that not do-dishes never
The husband ruined it by never washing the dishes.

Complement clauses with adverbial antecedents are therefore considered as lexicalized if

1. they have a non-compositional meaning and unpredictable semantic role which overwrites the compositional one,

2. they trigger other lexical specificities, e.g. the unmotivated definite conjugation of

the matrix verb,

3. they occur with an antecedent which otherwise does not allow hogy clauses.

IV.3.6.5 Complement Clauses without Antecedent

As we have seen above, antecedents are optional when the hogy clause has a subject or direct object function, and can sometimes also be omitted with other functions. On the other hand, certain predicate groups come with hogy clauses which cannot be preceded by any antecedent. The most typical examples is verbs of communication:

János szólt, hogy menjünk át.
John say+past, that go+subj+pl+1 prefix
'John told (us) to come over.'

János szólt, hogy elkészült az ebéd.
John say+past, that is-ready the lunch
'John told (us) that lunch was ready.'

Az autós rám kiáltott, hogy álljak meg.
The driver me+sub shout, that stop+subj prefix
The driver shouted at me to stop.

Similarly to other verbs of communication which often come with a propositional direct object either in indicative or in subjunctive, the clauses above express the content of the communication. However, propositions without antecedent do not behave like direct objects. First, the matrix verb in these examples is conjugated in the indefinite paradigm, while direct object clauses induce definite conjugation. Second, these verbs are not transitive in that they cannot have an NP direct object. On the other hand, it is semantically motivated to encode these clauses as direct objects, as it makes possible to capture the parallelism with communication verbs with object clauses of the same semantic role. Clauses without antecedent are in a complementary distribution with direct object clauses, as they only occur with strictly intransitive verbs. Hence, for the sake of

generalization capacity, we decided to categorize these clauses as direct objects to ensure a uniform labeling of propositional content with communication verbs. At the same time, a new morphological feature was introduced to mark the exceptional indefinite conjugation (+object_type=indefinite).

Verbs of emotion can equally appear with propositional complements without antecedents. These propositions express the cause of the emotional state:

A húgom nyafogott/panaszkodott/bőgőtt, hogy iskolába kell mennie.
My sisster whined/complained/cried, that school+ill need go+INF
My little sister was whining/complaining/crying because she had to go to school.

A kislány szomorkodik, hogy nem kapott ajándékot.
The girl being-sad, that not get+past present+ACC
The little girl is sad because she did not get any present.

The difference is that these predicates allow to put an antecedent in the matrix clause, although this is not the default case, as the resulting sentences are not as neutral as the ones without antecedent:

A húgom nyafogott/panaszkodott/bőgőtt azért/amiatt, hogy iskolába kell mennie.
My syster whined/complained/cried, that+CAU, that school+ILL need go+INF
My little sister was whining/complaining/crying because she had to go to school.

A kislány szomorkodik azért/amiatt, hogy nem kapott ajándékot.
The little girl being_sad+sg+3, that+cau/that_for, that not get+past present+ACC
The little girl is sad because she did not get any present.

Adjunct clauses expressing cause with the same antecedents can modify a wide range of predicates: this is a productive adjunction possibility in Hungarian syntax. However, the structures above are lexicalized with respect to the optionality of the antecedent and

hence considered as complement clauses.

IV.3.6.6 Syntactic Features

IV.3.6.7 Mode of the Subordinate Predicate

The predicate of the subordinate *hogy* clause can be in predicative or in subjunctive (imperative) mood. Subjunctive and imperative mood only differ when the verb has a prefix: in subjunctive it remains attached to the verb, whereas in imperative it is detached and follows the verb. Since there is no difference of a strictly morphological nature between these moods, and they completely coincide when the verb does not have a prefix, they were not differentiated in the database. By default, the mood of the subordinate clause is underspecified, since many verbs (e.g. verbs of communication or thinking) can govern both predicative and imperative clauses. The +conj distinctive feature is used to express that subjunctive mood is compulsory. Interrogative clauses are characterized by the presence of a question word, detached verb prefixes (similarly to questions occurring in main clauses), the conjugation of the verb being either predicative or subjunctive/imperative. By default, *hogy* clauses are interpreted as being affirmative, while interrogative clauses are marked by the +questionword feature.

IV.3.6.8 Pronominal and Nominal Antecedents

The pronominal antecedent, as we stated above, is taken to be optional by default. A unary feature is added to the entry when the pronoun is obligatory in the main clause. Even the conjunction *hogy* can be optional with a few verbs – these items are labelled by the +hogy_opt feature. Certain predicates allow a clause with a nominal antecedent, or the same clause with an optional pronoun antecedent, or a nominal complement with the same case suffix:

Az elnök cáfolta az (újabb utazásáról szóló) híŋrt.
The president denied the news (about preparing a new trip).
Az elnök cáfolta, hogy újabb utazásra készül.
The president denied that he would be preparing a new trip.
Az elnök cáfolta a híŋrt, hogy újabb utazásra készül.

The president denied the news that he was preparing a new trip.

It is a question whether we have to consider the clause as an optional extension of the nominal complement, or as a propositional complement with a nominal antecedent. In the first case, we do not need to encode the structure as a separate entry. Indeed, we consider it as a property of the noun that it can be extended by a hogy clause. In fact, in every position where we can have a noun optionally extended by a clause, we can also have a clause without nominal antecedent, we cannot decide independently which structure to consider as basic. Let us consider a verb with a complement structure in which propositional content is allowed, but its syntactic expression is not constrained, beside subcategorization for a given case:

a) A szomszéd megértette a problémát.

The neighbor understood the problem.

V + complement1=N+case=sup

b) A szomszéd megértette, hogy nem működik a szoftver.

The neighbor understood that the software did not work.

V + complement2 =CLAUSE+case=sup

c) A szomszéd megértette a problémát, hogy nem működik a szoftver.

The neighbor understood the problem that the software did not work.

The question is about what kind of category we should attribute to structure c). The main verb itself requires a complement which can be expressed by propositional content and accepts nouns that themselves can be extended with a syntactic proposition, due to their semantics. Either way, both the verb (example b)) and the noun can appear independently with a proposition: 'a probléma, hogy nem műdik a szoftver' is a well-formed NP which does not need to syntactically depend on any verb. The way we encoded these structures in the verb database was by creating two separate entries: one corresponding to b) (which is a regular clause complement) and another one corresponding to a) and c).

In the latter entry, the NP is coded as a complement with the corresponding case, and the optional clausal extension is identified separately as a different constituent. Instead of being inflected for case, the clause complement has a unary feature that marks it as being linked to the NP antecedent (i.e. being an extension to it).

The advantage of this approach is that it enables us to directly encode coreference relations between the arguments of the two predicates according to the guidelines specified in the following section, while still maintaining the presumption that the internal structure of complements is not 'visible' at the level of argument structure [9].

IV.3.6.9 Coreference and Missing Subjects

Subject can be omitted from any Hungarian sentence. When the subject of the subordinate clause is not present, the subordinate predicate still agrees with it in number and person. Moreover, semantic properties of the subject can sometimes be (partially) inferred from the semantic selection constraints of the verb. Besides these clues, we can also presume that in most cases, the dropped subordinate subject co-refers with an argument present in the matrix clause. This hypothesis does not hold for every context and depends largely on the semantics of the main verb. As lexical properties of matrix verbs often determine coreference relations, is logical to encode this type of information in the corresponding entries. Verbs with a propositional complement were therefore categorized according to the coreference constraints they put on subordinate clauses. The strength of such constraints varies from verb to verb and from relatively strict constraints to rather probabilistically interpretable features. In any case, the constraint is always pragmatic in nature; violating it do not result in syntactically ill-formed sentences. Using them could potentially complement a probabilistic co-reference resolution system. The +subject_subord= feature was used to identify the absent subordinated subject. The value of the feature is a reference to the identifier of an argument in the same lexical entry (direct_object|complement1|complement2). By default, if the feature is absent, the target argument is presumed to be the subject of the main verb. It is important to note that this feature is interpreted inside an argument structure, i.e. it is not linked to a verbal lemma

[9]The relevant pieces of information for matching the argument structure are properties of the head of the complement phrase and systematically 'percolate' to the phrase in our shallow parser.

in general but to a particular entry of this verb, since this property can change as a function of the complete argument structure of a given verb. [10] The following constraints have to be met in order for the subordinate subject to be identifiable to the target argument: the inferred number and person of the absent subject has to coincide with those of the target argument, the eventual semantic selection constraints of the subordinate predicate on its subject have to be met by the target argument. If these requirements are met, the absent argument can be linked to the target argument. Let us illustrate the coreference schema of one lexical entry of the verb mond ('tell'):

Subject: N+case=nom
Object: CLAUSE+subject_subord=complement1+conj
Complement1: N+case=dat

Azt mondta neki Jani, hogy vásároljon be.
PRO.ACC tell+past PRO+DAT Johnny that shop+subj prefix
Jani told him to go shopping.

By using the +subject_subord=complement1 reference, we identify the empty subject of the subordinate clause with the dative argument of the matrix verb:

Azt mondta $neki_i$ Jani, hogy $..._i$ vásároljon be.

If he matrix verb does not have a syntactic subject (e.g. *úgy adódik, hogy* '(it) turns out that') or if it has a syntactically present but semantically empty subject (e.g. *úgy alakult a helyzet, hogy* 'the situation turned out that'), we need to indicate that the default coreference interpretation does not hold: this is done by the +subject_subord='intern' feature. The constituent that co-refers with the subordinate subject is not to be looked for in the main clause. The same notation applies when semantic factors prevent a coreference interpretation between the NPs in the argument structure of the matrix verb

[10] E.g. different coreference may be interpreted differently in the subjunctive and in the indicative complement clause of the same verb.

and the subordinate subject (e.g. *'vki útját állja annak, hogy* 'someone prevents that...' [11])

Illustration: default features A 'default' pattern for a verb with a propositional complement is the one that contains the least amount of explicit specifications. It is also understood to be the most common type of hogy complements: the one that simply replaces an NP complement with the same case suffix and potentially the same semantic content. However, the structure of the lexicon is such that a certain amount of information gets specified via default values: the default *hogy* clause is not underspecified with respect to such properties. The following example illustrates the use of default values.

Lexical enrty:

csodálkozik ('to be surprised')
Subject: N+case=nom
COMPL1: CLAUSE+case="sup"

Syntactic Structure:

János nem csodálkozott (azon), hogy megint megbukott a számvitel-vizsgán.
John NEG was-surprised PRO+SUB that again failed the accountancy exam.
John was not surprised to fail the accountancy exam again.

The features assigned to the entry by default are:

- optionality of pronominal antecedent:
  János nem csodálkozott, hogy megint megbukott a számvitel-vizsgán.
  John NEG was-surprised that again failed.sg.3 the accountancy exam.

- obligatory conjunction *hogy*,

[11] Unlike with the English verb, the subordinate subject of the Hungarian verb cannot be raised to the main clause.

- indicative mode in the subordinate clause,
- presumed co-reference between the main subject and the omitted subordinate subject:

  János$_i$ nem csodálkozott, hogy ...$_i$ megint megbukott a számvitel-vizsgán.

IV.3.7 Semantic Description

Beside syntactic subcategorization, verbs also put semantic selection constraints on the elements in their argument structure. The strength of semantic selection is highly variable among verbs: violating a semantic constraint in some cases results in uninterpretable or ungrammatical sentences, while in other structures it simply produces sentences which are pragmatically unlikely. The strength of a selection constraint can be inferred from corpus (provided that we have sufficient amounts of data, because such high-level generalizations require very high coverage), either automatically (Resnik, 1993) or intuitively by consulting the corpus while coding the lexicon. In any case, the crucial difficulty about encoding semantic selection constraints is that these constraints can be violated, even systematically, e.g. in metaphors which are common, mostly, but not exclusively in literary texts. E.g.:

... titokzatosan intenzíŋv olasz életek szenderegnek
... mysteriously intense Italian lives are slumbering

It is for this consideration that we decided not to use semantic descriptions in a restrictive way. As our presumption was that basically any semantic constraint can be violated, we used semantic restrictions exclusively for discriminating between different senses. An additional semantic constraint on one of the (otherwise identical) frames would ensure that a different meaning can be associated to this structure. E.g.:

unspecified:
bele|vág SUBJECT: N+case=nom COMPLEMENT1: N +case=ill
meaning: to punch, to cut (into)

Meglepetten azon kapta magát, hogy ököllel belevág az ablaküvegbe.

He surprised himself by suddenly punching his hand through the window

with semantic specification:

meaning: to start, to embark upon

bele|vág SUBJECT: N+case=nom COMPLEMENT1: N +case=ill+abstract

A meg nem nevezett angol ügynök elhagyta az országot és belevágott következö izgalmas küldetésébe.

The unnamed English agent has left the country and has already embarked on his next challenging mission.

On the other hand, when the input itself presents semantic ambiguity, both analyses can be provided (according to the structure of the parser):

unspecified:

fel|ad SUBJECT: N+case=nom OBJECT: N +case=acc

meaning: to hand over, to hand up

Feladtam az ïŋrást.

I handed up the document.

With semantic specification:

fel|ad SUBJECT: N+case=nom OBJECT: N +case=acc+abstract

meaning: to give up

Feladtam az ïŋrást.

I gave up writing.

A big set of deverbal nouns show an ambiguity in Hungarian: they can refer to an action (writing) or the (concrete, materialized) result of this action (written document). Depending on the annotation of these nouns and on parsing strategy, we have different

possibilities to deal with this question while matching argument structures, or to preserve the ambiguity and delegate the problem to a later step of the analysis.

The following binary semantic features were used for semantic selection restrictions on complements, all of them characterizing only NP complements:

1. abstract (as opposed to material)
2. animate (humans, animals, plants)
3. human
4. mass (mass noun)
5. bodypart (parts of human (animal) body, including organs)
6. time (units of time, dates)
7. weather (any vocabulary related to weather; corresponding to natural force in a thematic role terminlogy)
8. measure (measurement units)
9. dynamic (machines (computer, vehicles etc.); nouns that can function as subject of the verb "működik" (to work))
10. proper noun

This set of semantic features was also established and finalized during the first phase of coding. Subsequently, the nominal vocabulary used in the project (13.325 nouns) were annotated with these features in the noun dictionary. It is important to note that adding semantic features to the verb database is a very time-consuming task as it implies that the annotation of the nominal part of the vocabulary will have to be extended accordingly. On the other hand, when using the verb lexicon for an application that requires fine-grained meaning distinctions (e.g. for machine translation), narrower semantic classes can be needed. The development of the Hungarian WordNet (Kuti et al., 2007), subsequently to the verb lexicon project, allowed to extract other relevant semantic classes, which were

used for more precise meaning descriptions in the machine translation application. E.g.:

unspecified:
meg|állapïŋt SUBJECT: N+case=nom OBJECT: N +case=acc
meaning: to assert, to specify

E fejezet megállapíŋt néhányat e követelmények közül.
Some of these requirements are specified in this chapter.

With narrow semantic specification:
meg|állapïŋt SUBJECT: N+case=nom OBJECT: N+case=acc+semgroup=disease
meaning: to diagnose

Az orvos tüdőgyulladást állapíŋtott meg.
The doctor diagnosed pneumonia.

IV.3.8 Conclusion

In this section, lexical representation of Hungarian verbs and verbal subcategorization frames were presented from a pragmatic, data-oriented point of view. A computational lexicon for Hungarian verbs has been constructed according to the guidelines detailed above. The lexicon is meant to be a general-purpose machine readable database, primarily to be used for parsing. We presented the most important aspects of the design of our lexical database of verbal subcategorization frames. We described the structure of the lexicon, especially the role of POS categories and case/postposition information with respect to the unambiguous identification of complements within a subcategorization frame. Uninflected grammatical categories (adverbs, infinitives) and propositional complements without a case-marked antecedent were categorized by category-specific distributional tests. The advantage of this categorization is to capture relevant similarities between NP complements and complements belonging to uninflected categories but filling the same function in the argument structure.

During the construction of the verb lexicon, the compositionality criterion was applied to separate productive and predictable structures from truly idiosyncratic complements. However, the definition of lexical semantic verb classes is a complex task: we found that very few semantic class could be satisfactorily defined in parallel to the construction of the database. We believe that many phenomena, currently encoded in the lexicon, boil down to lexical semantic properties of verbs and as such, can be captured in terms of class-specific diathesis alternations or class-specific adjunction possibilities. In the following two sections, we will present two different experiments for enriching our database with well-motivated generalization through lexical semantic properties.

Chapter V

Manual Definition of Verb Classes by a Typology of Adjuncts

(The ideas presented in this section result from a tight cooperation with Enikő Héja and contain overlap with (Gábor and Héja, 2005))

V.1 Introduction

In this chapter we present an experiment which aims to enrich the verb lexicon with lexical semantic features. This experiment can also be considered as a first attempt to put the theoretical issues discussed in chapter III into practice: namely, to define verb classes on the basis of a typology of adjuncts they allow, and to find class-specific syntactic or morphosyntactic alternations. The alternations serve as formal tests to decide on whether a given verb belongs to the class in question; the hypothesis being that alternations designate verb classes which allow for the same type of adjuncts to appear in their context. We also investigate the potential of such a verb classification for a parallel syntactic and semantic role labeling.

To recapitulate, our reflection in chapter III aimed at identifying criteria to tell apart complements from adjuncts in order to 1) build a database of complement structures for Hungarian and 2) provide a consistent method for the description of adjunction in Hungarian. These two steps are necessary (though not sufficient) for a deep parsing, conceived in terms of labelled dependency relations between top-level sentence constituents.

In chapter IV, we presented the coding guidelines of our verb lexicon. This lexicon relies on the notion of compositionality and productivity among verb classes; however, a consistent definition and enumeration of verb classes needs to be preceded by a detailed linguistic investigation. The scope of the work presented in this chapter is to define predicate classes manually, and to predict the distribution of adjuncts by class-specific adjunct grammars. We believe that syntactic function and semantic role go hand in hand: hence, the grammatical function labels we associate to adjunct types are interpreted both at the semantic and at the syntactic level.

Grammatical function, as described in chapter III, is interpreted in a wide sense and includes semantic role and syntactic complement/adjunct status. It follows from the claims presented earlier that semantic and syntactic role labeling cannot be done in two independent steps, since complement or adjunct status cannot be defined without reference to semantics. On the one hand, this is implied by the fact that, as it was shown previously, there are no systematic differences between complements and adjuncts at the syntactic level in Hungarian. On the other hand, we will see in this section that some distributional phenomena can only be accounted for using a finer grained and semantically motivated categorization of adjuncts.

Another important benefit of including semantic roles in our definition of grammatical function is that it makes possible to build a parallel syntactic and semantic role labeling system. Up to the present, no semantic role inventory (such as FrameNet (Baker et al., 1998) or VerbNet(Kipper-Schuler, 2005)) or corpora annotated with semantic roles have been constructed for Hungarian; the experiment presented in this chapter is therefore the first one aiming to provide a (partial) inventory of semantic roles. The semantic roles used in our system are entirely grounded on syntactic or syntactically relevant lexical semantic distributional properties. A crucial difference of our role inventory compared to the ones cited above is that we only deal with categorizing adjuncts[1]. We claim that adjuncts' semantic role is generalizable (the semantic relation between a predicate and an adjunct is productive over verb classes) . On the other hand, we are not associating semantic roles to complements, since their roles can only be described precisely with reference to the meaning of the predicate. With this important distinction in mind, let us note that

[1]Though an extended scope of adjuncts, see III/3

we are not challenging FrameNet or VerbNet or mainstream SRL systems using these resources in that we are not aiming to produce an exhaustive, general-purpose inventory of semantic roles. Indeed, we are trying to disambiguate top-level NPs with respect to their dependency relation with the predicate. As a side effect, we are also producing semantically motivated labels to the extent that the content of the semantic label bears syntactic relevance. In what follows, we describe an experiment on building grammars for disambiguating morphosyntactic categories according to the grammatical functions they embody.

We start from the evidence that rather than configurational positions, morphological markers encode grammatical functions in Hungarian. Such morphological markers include case suffixes and postpositions. For the sake of our current experiment, we limit ourselves to case suffixes, as they are more ambiguous: most of them can encode complement roles as well as more than one adjunct role. We assume that case suffixes as morphological markers of syntactic functions are not always devoid of meaning: many syntactic structures which would normally be delegated to the verbal lexicon by concurrent theories can in turn be analyzed compositionally, by joining the meaning of the case suffix (i. e. the semantic role of the case-marked constituent), the meaning of the case- marked NP and that of the verb. Therefore, a systematic description of verbal complementation can be viewed as an enumeration and disambiguation of the different functions these case suffixes (and postpositions) can encode in sentences. A parallel syntactic and semantic role labeling system can be elaborated on the basis of case suffixes, by the specification of the contexts they occur in and the roles they encode in the given context. In what follows, we will present arguments that reinforce the conclusion drawn in chapter III about the syntactic relevance of semantic features, the methodology of constructing rules for disambiguating case suffixes according to the grammatical function they denote, and a case study for the specific suffix -val (instrumental) as an implementation of the parallel syntactic and semantic role labeling method and its linguistic background. The grammars have been developed on a corpus of two contemporary novels and evaluated on another corpus of literary texts.

V.2 Syntactic Relevance of Semantic Roles

The choice of semantic roles in a semantic role labeling system can be motivated either by linguistic evidence or, in language technology, by domain-specific needs. In language theory, it is considered as a guarantee of validity to build our categorization on natural classes; this is also what we argued for in 3.3. Dowty's highly influent article on thematic roles (1991) suggests a procedure that we completely endorse and try to apply to what we think to be relevant generalizable semantic roles in Hungarian:

"A useful strategy for ensuring that we are examining a single semantic phenomenon under the rubric of 'thematic role' may be to determine what role types are motivated by the argument-selection problem, and then see whether this same set of role types is also significant elsewhere in grammar."[2]

Moreover, in our case, syntactic relevance of semantic roles is extremely important since our system lies on the supposed parallelism between semantic and syntactic structure. In what follows, we provide evidence for the syntactic relevance of the semantic roles we used.

Most linguistic theories hold the assumption that only constituents of the same category can be coordinated (Chomsky, 1957). When talking about case-marked NPs governed by a verb (either as complements or as adjuncts), besides category, grammatical function/morphological case has to be taken into account. For instance:

Félek [a kígyóktól és a vihartól].
I am afraid of [snakes and storms].
*Félek [a kígyóktól és a sötétben].
*I am afraid of [snakes and in the dark].

On the other hand, if two NPs in the same sentence have the same case and the same grammatical function, they have to be coordinated:

[2]With the important distinction that we take adjuncts' distribution as a starting point and examine their selection and realization first.

*Félek [a kígyóktól] [a vihartól].

*I am afraid of snakes of storms.

Consider now the following example:

*János beszennyezte a szőnyeget sárral és a cipőjével.

John stained the carpet+ACC mud+INS and the his-shoe+INS

*John stained the carpet with mud and with his shoes.

The sentence above shows that the constituents sárral 'with mud' and cipőjével 'with his shoes' cannot be coordinated, therefore they play different roles in the Hungarian sentence in spite of the fact that they are both optional extensions with the same case suffix, and both of them could be interpreted semantically as instruments of the staining event. They can appear together in the sentence without being coordinated (although they should not be adjacent):

János a cipőjével beszennyezte a szőnyeget sárral.

János sárral szennyezte be a szőnyeget a cipőjével.

If only constituents of the same category/same grammatical function can be coordinated, it follows that the NPs sárral and cipőjével do not share the same syntactic function. In a binary scale this would imply that one of them has to be analyzed as a complement, while the other one is an adjunct. However, standard descriptions of prototypical complements and adjuncts cannot help to determine which role should be assigned to which NP, since they are positioned at the same level of productivity and optionality. The problem becomes deeper when we take a look at the following examples:

Párizsban még bíztam az apámban.

Paris+INE still trust+past+sg+1 the father+ps+sg+1+INE

'In Paris I still trusted my father.'

In the example above the Hungarian counterparts of 'Paris+INE' 'father+INE' share the same case suffix in a well-formed sentence. This is unproblematic, since the two NPs fill different roles, 'Paris+INE' is clearly a (location) adjunct, while 'father+INE' is a complement. Now consider the following sentence:

A tavalyi évben Párizsban még bíztam az apámban.

The last year+INE Paris+INE still trust+past+sg+1 the father+INE

'The last year in Paris I still trusted my father.'

In the sentence above three different constituents have the same case (inessive) to encode three different functions: a complement, a time adjunct and a location adjunct. Since none of them can be coordinated, the question arises how it is possible to account for the grammaticality of the sentence using only two categories for grammatical functions? If the two adjuncts in the sentence above can appear together in a sentence without being coordinated, this is because their grammatical function is different. If we find the solution in dividing the category of adjuncts (time and location adjuncts in the example above), we admit that there is a semantic motivation behind the distributional behavior of sentence constituents attached to the verb. Therefore, the difference between the adjunction rules that operate in this sentence can be described in semantic terms. This corresponds to the point we developed above: adjuncts are attached to the verb by syntactic rules which assign different semantic roles to the NPs in parallel with their diverse syntactical functions. The rules take as input the case suffix and the NP, and, optionally, refer to the semantic class of the verb, and the output is a syntactic/semantic function label on the NP appearing as adjunct. The function label encodes a relation: adjunct of a verb on the syntactic level, time or location of an event at the semantic level. The semantic content is calculated compositionally from the semantics of the case suffix, that of the NP and of the verb. This can be extended to the other examples, hitherto considered as complements: a more elaborated lexical semantic representation of verbs allows to separate those meaning components which play a role in the syntactic operation of adjunction, and at the same time allow to construct compositional representations of the semantics of adjunction.

In what follows we will show how a gradual notion of adjunction can contribute to

a better description of Hungarian syntax and to semantic role labeling. This approach does not aim at eliminating the complement-adjunct distinction – on the other hand, it aims at complementing it by introducing a gradual scale. The advantage of this method is to be able to explain the appearance and syntactic behavior of certain optional NPs in verbal complementation patterns and account for the rules of adjunction in Hungarian.

V.3 Case Suffixes

The systematic exploration of adjunction in Hungarian we propose is based on the study of case suffixes. The case suffix can be defined as the rightmost inflectional suffix on Hungarian nouns: it cannot be followed by any other suffix, and a noun can have only one case suffix. According to these criteria, we can distinguish 19 cases in Hungarian. At the first stage of our research, we hypothesized that the two functional/grammatical cases (nominative and accusative) have a special status with respect to syntactic and semantic role assignment. The principal syntactic specificity of these two complement types is the obligatory agreement with the verb (in number and person for the subject; in definiteness for the object, see IV.1 Nominal Complements). We presume that due to their specific syntax, they do not represent typical items in the case paradigm; our strong presumption is that subjects can only occur as a verbal complement (subject or nominative complement of copulative verbs), i.e. with a non-transparent semantic role assignment. Accusative on the other hand, as we have seen in IV.3.4, can be a marker for the pseudo-object adjuncts, which were not considered as being subcategorized for by the verb due to their very productive nature. Pseudo-objects are always indefinite; the question of agreement is therefore irrelevant for this category. Pseudo-objects can also appear with otherwise intransitive verbs, i.e. verbs without a definite paradigm. We can thus say that accusative case can be a marker for adjuncts under some specific conditions, but accusative is still not a typical case suffix with a variety of potential transparent meanings. Obviously, linking theories have been investigating generalizations and predictions about the semantic nature and the realization of subjects and objects. However, these functions and the corresponding suffixes were excluded from the scope of our investigation hereby due to their atypical behavior. When investigating the distribution of other cases, we supposed

that they have their own syntactic and semantic properties that make them compatible with certain verbs and thus enable them to act as morphosyntactic markers of adjunct functions. Our task is to examine case suffixes and enumerate their possible syntactic and semantic functions and to formalize a set of rules which disambiguate between the functions encoded by the same suffix and associate the corresponding syntactic/semantic label to the case-marked NP in each context. We therefore try to answer the following questions with respect to the distribution of case suffixes:

1. Which are the sentential contexts an NP with the given case suffix can appear in?
2. What is the syntactic function of the case-marked NP (studied especially with respect to coordination)?
3. What is the semantic relation between the NP and the predicate?
4. Are there any restrictions on the scope of NPs or predicates in context of which they can express the same function?

If we find that for a given distributional context specified in step 1), the case suffix fills a function specified in step 2), while having a semantic role specified in 3), the following task is to formulate the input side of the rule according to the findings of step 4). If there are no distributional restrictions on the nature of verbs the adjunct can occur with, we can say that the case suffix has a default meaning and is able to express this meaning in any context. In this case, we do not need to specify the scope of compatible predicates in terms of verb classes. Since we define case suffix meanings in terms of relations, it is extremely rare to find general meaning components that would be compatible with any potential verbal meaning. In practice, it is more frequent for case suffixes to have more than one 'default' meaning but with restrictions on the semantic nature of the NP they occur with.

V.4 Types of Rules. Informal Test of Compositionality

Since there may be more than one 'default meaning' associated with a case suffix, the input side of the default rules we create will have to refer to the semantic (or, eventually,

syntactic) features of the NP the suffix appears in. They are still considered as default rules, since they cannot refer to the predicate: the case suffix encodes the same grammatical functions independently of the semantics of the predicate. For instance, the case suffix -ban (inessive case) productively denotes time adjuncts if it appears on a constituent expressing time; otherwise, it denotes location.

Whilst the adjunction structures captured by default rules represent the maximal degree of productivity and compositionality, on the other edge of the scale we find verb + complement structures which are neither compositional nor productive. Compositionality can be tested informally by trying to paraphrase the role of the NP in relation to the verb: in the case of a complement structure, this relation cannot be paraphrased without including the meaning of the verb. For example, consider the sentence:

1) A közönség elhalmozta az előadót kérdésekkel.

The audience overwhelmed the lecturer+ACC questions+INS.

The audience overwhelmed the lecturer with questions.

If the structure verb + NP.INS ('overwhelming with questions') was compositional, we could associate an abstract label to the NP which would capture its role without making reference to the predicate of the sentence (similarly e.g. to location, manner, goal etc.). This seems impossible, as is shown by the fact that we cannot formulate natural language paraphrases for expressing the relation between the predicate and the NP without including the verb itself or a synonym of it. Correspondingly, the NP constituent cannot be productively used with other verbs with the same semantic role and syntactic form. As the semantic relation between the NP and the verb is specific to the given predicate and cannot be generalized, it is obvious that the structure is not productive. This informal test gives us a first hint of the complement status of an NP.

Finally, the rest of the verb+NP structures seem to be mid-way between regular adjunction phenomena and lexical encoding. In these cases, the function the case suffix has in the structure depends on the semantic class of the predicate. To take a very simple example, let us consider the different uses of ablative case (*–tól*). Used with motion verbs, it denotes the starting point:

2) János odébbment az oldalvonaltól.

John walked away the sideline+ABL.

John walked away from the sideline.

The same case suffix appears systematically on NPs denoting the cause of the event; the scope of predicates allowing this syntactic encoding of the said semantic role could be identified as sharing the meaning component 'change of state'. In fact, a wider verb group, that of non-agentive verbs allow the cause to be expressed by an ablative NP (see our argument on the event type shift test in III/2); 'change of state' verbs are specific in that they allow alternating expressions for the *cause* NP involving instrumental and ablative (see V.5.2.3).

3) János elszomorodott a hírtől.

John got sad the news+ABL.

John got sad after hearing the news.

The two features that differentiate these instances of grammatical functions from the ones exemplified by the sentence 1), and therefore provide motivation for analyzing them as adjuncts, are the generalizability of the semantic role and the productive use of the case suffix with a semantically defined verb group. We therefore construct non-default adjunct rules to associate the semantic roles above to the NPs occurring with any verb member of the verb class the role is specific to. Thus, we can associate these functions to the case suffix by two distinct rules, each of them referring to the semantic class of the verbs they occur with in their input side. The next section clarifies the details about the nature of adjunct rules. An important consequence of our method is that having a wider range of grammatical functions (i.e., positing as many separate grammatical functions as there are syntactically motivated semantic roles in our system) provides us a tool to account for the co-occurrence constraints that apply to NPs having the same case suffix. Concretely, if we affirm that a well-formed Hungarian clause cannot contain more than one NPs with the same grammatical function expressed with the same case suffix unless

they are coordinated, then we have difficulties explaining the correctness of sentence [3], which has two adjuncts of the same structure (NP+inessive case). This observation led us to state as many roles for case suffix bearing NPs as there are rules for the given case suffix: each of our rules outputs a different role label. Hence, we can simply state that no clause can contain more NPs with the same role and the same suffix: in other words, each rule may apply only once in a clause[3]. In sentence [3], one of the NPs in inessive case is analyzed by a rule which outputs 'location adjunct' as a function label, while the other one is labelled as a 'time adjunct' by a different rule: this is why they are allowed to appear together in a clause without being coordinated.

V.4.1 Note on Alternations

By studying the semantically distinct functions of case suffixes, we believe to be able to explore syntactic-semantic verb groups allowing the same alternations and attributing the same semantic role to the given suffix. The key to guarantee coherence are, however, syntactic alternations, as they constitute the accessible/testable part of our practice. At this point we have to consider that syntactic alternations have not been studied for Hungarian as systematically as it has been done for English by Levin (1993). Essentially, the notion of syntactic alternation or, more specifically, "diathesis alternation" has not been given an exact definition for Hungarian.[4] Therefore, we have to define what we consider as syntactic alternation and what alternations are relevant for linking and semantic role attribution in Hungarian. Due to the characteristics of Hungarian as a non-configurational language and its rich morphology, we expect the semantic components of a predicate to be reflected in syntax in a different way. The choices we needed to make were for instance: do we include syntactic operations induced by morphological derivation (e.g. passive-like structures) or by verb prefixation? Or the possibility to change the case suffix on a given element of the complementation pattern, without changing the truth conditions of the sentence (if described by truth conditions, e.g. meglepődött/kiakadt/megdöbbent a polgármesteren (SUP) / polgármestertől (ABL) : *(s)he was supprised/shocked by th mayor*).

[3]NPs added by the same rule are necessarily coordinated

[4]Note that such a definition cannot be formulated without a reliable complement test.

We will argue for extending the notion of diathesis alternation and include morphological derivations as long as they induce a systematic change in the complementation pattern of the predicate. The reason for this choice with respect to SRL is that verb groups defined by morphosyntactic alternations allow for stating relevant generalizations and therefore extending the lexicon using unambiguous morphosyntactic criteria. From a theorietical point of view, diathesis alternations in Hungarian require more detailed studying. On the other hand, the generalizations yielded by our rules may reveal important correlations about new aspects of the syntax-semantic interface and, more generally, about linking properties of predicates in morphologically rich, non-configurational languages. We hereby emphasize that the present work aims to 1) explore syntactically relevant semantic properties and the verb classes defined on the basis of these properties, 2) annotate NPs belonging to the complementation pattern of the predicate in Hungarian texts. In contrast, we do not aspire to build a representation which predicts every piece of syntactic information about the structure formed by the predicate and its syntactic dependents (such as the obligatory character of the complement/adjunct, the extensive description of semantic selection constraints, application of syntactic transformations/participation in syntactic constructions, description of control relations and the like).

V.5 Showcase: Typology of Adjuncts in Instrumental Case

After having described the general framework, we will now present an experiment through which we can follow the concrete realization of the definition of verb classes by a typology of adjuncts. Th experiment aims to produce a set of grammars to annotate occurrences of verb+complement/adjunct structures with syntactic/semantic roles, relying on lexical semantic verb classes. The present work draws from the previous experiment by (Gábor and Héja, 2005) and especially relies on the semantic roles and alternations defined by E. Héja within this experiment.

The exploration of verb classes follows from a categorization of adjuncts with a given case suffix according to their semantic role. In this study, we take the distribution of case- marked NPs as a starting point and attempt to formulate precise criteria about the

contexts in which the suffix fills a certain function. We subsequently formulate annotation rules on the basis of these contexts and annotate a corpus with semantic roles. The interest of this work is twofold: on the one hand we test whether it is possible to categorize predicates into lexical semantic classes according to the criteria presented above; on the other hand, we test whether this method can be a valid basis for developing a semantic role labeling (SRL) system for Hungarian. As this is the first piece of manual work completed according to the guidelines of the compositionality test defined previously, a significant number of methodological questions had to be dealt with while confronting the principles to the data. The most important one concerns the nature of syntactic alternations in Hungarian, which, as we have seen, had not been thoroughly defined previously to the exploration to adjunction rules and verb classes. To implement the syntactic and semantic role labeling system, a set of grammars have been constructed in NooJ to annotate NPs with this case suffix according to their function. The rule system has been developed on a short literary corpus. A NooJ corpus was created from these texts (coming from the Hungarian National Corpus). The decision to develop our rules on contemporary literature texts was motivated by the hypothesis that these texts contain a bigger variety of semantic roles (as opposed to more restricted text types present in the HNC, such as scientific texts/news articles/transcripts of parliamentary debates). Moreover, the corpus texts were analyzed using the NP chunker presented in II., which does not include a named entity recognizer. As literary texts have less complex named entities than the other text types in HNC, the chunker produces the highest level of precision on literary texts. In this section, we would like to evaluate the precision of our syntactic/semantic role labeling grammars independently of the precision of the input and thus we opt for the less noisy input possible. Subsequently, a third contemporary novel was used to evaluate the results. It is important to note that we aimed to construct a rule set which outputs unambiguous annotation. We aimed at minimizing semantic ambiguity by linking semantic properties to directly observable syntactic phenomena. The input of our rules is almost completely unambiguous, both with respect to constituent chunking (see chapter II) and with respect to semantic features. This means that semantic features encoded in the dictionary were always assigned to matching lexical heads. If a dictionary entry is provided with a semantic feature, its occurrences as a lexical head are always given the

said semantic feature, whereas any lexical entry not having the given semantic feature in the dictionary are considered by default as having a negative value for this feature.

The suffix chosen for study is instrumental (-'vAl' 'with'). This suffix, as the name suggests, has an intuitively strong default meaning, but also occurs in several different contexts, exemplifying the gradual nature of productivity.

V.5.1 Default rules

We started by defining the default meanings of the case suffix. One of these meanings will be associated to the occurrences of the case suffix within the scope of the complementation pattern of any predicate (where the case suffix appears on a top-level NP syntactic dependent of the predicate), unless stated otherwise; i.e., unless a more specific rule applies to the same structure.

V.5.1.1 Instrument and Comitative

The first two default meanings we define for the suffix -val are comitative and instrument. Comitative, as a semantic role, refers to an additional participant of the action or event denoted by the predicate: this participant takes part in the action/event referred to by the predicate by accomplishing the same action as the agent or undergoing the same event as the patient of the predicate. The comitative NP, therefore, has to denote a human being. Instruments, on the other hand, are not humans but objects that are being used to carry out an intentional or unintentional action. It is with the +human semantic feature that we distinguish between the two default roles: comitatives are +human, instruments are -human:

V + NP+human+INS → comitative
V + NP-human+INS → instrument

A kapuőr időszerűnek találta [INSTRUMENT kettős lakattal] bezárni a kaput.
The concierge had seen fit to fasten the gate [INSTRUMENT with a double lock].

The adjuncts these two rules recognize are prototypical instances of adjunction in that

the input side of the rule does not refer to any specific property of the verb: these adjuncts are therefore productively used with a very wide range of predicates. The fact that comitative and instrument are different semantic relations, and not only differ with respect to the semantic nature (+/-human) of the adjunct NP, can be proven by coordination tests:

A portás a takarítónővel bezárta a kaput lakattal.
The concierge, with the cleaner, closed the door with a lock.
but:
*A portás a takarítónővel és lakattal bezárta a kaput.
*The concierge closed the door with the cleaner and with a lock.

V.5.1.2 Mode

The third default meaning we defined is mode adjunct. It is quite common that NPs with instrumental suffix have and adverb-like interpretation and distribution:

Kedvesen, széles mosollyal adtak útbaigazítást.
They gave us indications kindly, with a big smile.

A diákok gyorsan és lelkesedéssel végezték el a feladatot.
The students performed the task quickly and with enthusiasm.

but:
*A diákok lelkesedéssel és a tanárokkal végezték el a feladatot.
The students performed the task with enthusiasm and with the teacher.

*A diákok lelkesedéssel és egy számítógéppel végezték el a feladatot.
The students performed the task with enthusiasm and with a computer.

As the coordination tests above show, we are dealing with a new type of adjunction. This semantic role, like the previous ones, is productively assigned to a class instrumental NPs. The role is called 'mode' in our SRL system, and implies that if an NP with this role

occurs in the sentence, the NP describes a characteristic or a state of the agent/subject NP while participating in the event referred to by the predicate. In several cases it seems to be difficult to tell apart adverbial NPs from instrumental ones. This is because adverbial NPs express the mode of carrying out an action, while instrumental NPs refer to the instrument of that action – which, in turn, influences significantly the mode of carrying out the said action. To illustrate this, let us consider the following example:

János autóval ment moziba.
John car+INS went cinema+ILL
John went to the cinema by car.

Mivel megy János moziba?
what+INS go John cinema+ILL
What does John go to the cinema with?

Hogy megy János mozi - ba?
how goes John cinema+ILL
How does John go to the cinema?

The choice of the wh-expressions in Hungarian show that the two semantic roles (instrument and mode) are interchangeable: when using "mivel" ('with what') we enforce instrumental reading, while using "hogy" ('how'), the modal/adverbial one is supported (while this is the only option in English). The difficulty lies in the correct disambiguation, especially considering the fact that instruments, by extension, can be interpreted as modal adverbs, where 'using instrument x' corresponds to the mode of carrying out an action. However, non-ambiguous mode adjuncts do not allow the paraphrase 'using instrument x': they contain abstract nouns and the denoted notion (e.g. 'enthusiasm') is not 'used' for accomplishing the action. We limited the scope of our third default rule to non-ambiguous mode adjuncts. In order to identify non-ambiguous mode adjuncts, we used the approximation that it is most likely that nouns derived from adjectives or verbs take an adverbial position. We used a morphological feature present in our NooJ

dictionary which indicates that the noun in question is derived from an adjective (with the suffix '-ság') or from a verb (with suffix '-ás'), which are the most common Hungarian deadjectival and deverbal derivational suffixes, respectively. Besides this approximation, the study of the development/training corpus allowed us to extend the input side of this rule by semantically related, but morphologically simple nouns (not produced by derivation). Concretely, a set of NPs (mostly consisting of a bare noun) are used with the instrumental suffix and with an adverb-like distribution, e.g. örömmel ('with joy', happily). As we gradually came across such phenomena while studying the distribution of instrumental NPs, we indicated this semantic feature (NPs expressing a characteristic or a state and therefore allowing adverbial use) in the nominal dictionary.

V.5.1.3 Measure

The last semantic role we observed in the corpus and defined as a default rule was labeled 'measure'. The input of this feature is an NP which expresses a measurement unit, annotated as +measure or +time in our dictionary. An NP which expresses a time period or any measurement unit in instrumental case is annotated as expressing the rate of difference or time period between two states or two events. For example: 'három százalékkal nőtt' ('increased with (by) three percents') or 'három méterrel odébbment' ('moved-away with (by) three meters' – *moved three meters away*). Associating this function to a measure or time NP can only be overridden by a more specific (non-default or complement) rule. However, this does not imply that every predicate would allow this type of extension. What the rule states is that any NP in instrumental case meeting the semantic criteria has to be annotated as measure, unless stated otherwise by a previous, and thus more specific rule. In order to provide an extensive description of adjunction, however, one would have to define the scope of each adjunct rule in terms of predicate classes, including even as board notions as aspectual classes (for time adjuncts). On the other hand, for the time being and for the sake of the present experiment, we opted for this solution, yielding the same result in our role labeling system.

To recapitulate, the semantic roles described so far are not specific to natural classes of predicates.[5] They correspond to lexical meanings of the suffix '-val'. In the following

[5]They can be specific to certain, morphologically or semantically defined subclasses of case-marked

section, we present class-specific semantic roles, and the process of defining verb classes based on syntactic alternations.

V.5.2 Non-Default (Class-Specific) Rules

V.5.2.1 Non-default mass instrument

As we have seen in 5.5.1, there are two distinct, although semantically close instrument-type functions encoded by the suffix -val. The coordination test indicates that they have to be treated as two different syntactic functions:

János a cipőjével beszennyezte a szőnyeget sárral.
John his shoes+INS stained the carpet mud+INS.

This second type of instrument will be treated by a non-default rule which attributes the semantic role defined as non-default mass instrument. With respect to its semantic content, the NP is close to the NP complement of the well-known load/spray verb group (Levin, 1993):

Jack sprayed paint on the wall.
Jack sprayed the wall with paint.
Jack loaded hay on the truck.
Jack loaded the truck with hay.

The semantic content can be captured and identified using the implications that 1) the adjunct instrument has to denote a mass material, 2) as a result of the event, the material (or at least some amount of it) is transferred to another location, namely to the surface of the object denoted by the object of the predicate (as opposed to the default instrument role when the transfer is not implied). Predicates allowing this type of adjunct are for example 'beszennyez', 'bemocskol' or 'összemocskol' (to 'stain'), 'bepiszkít', ('to smear'), but also different prefixed variants of the verb 'ken', e.g. 'beken' or 'összeken'('to rub'/'to smear'/'to anoint',...), 'befest' ('to paint'). Semantically, these predicates form a

constituents, however.

sub-class of the change of state verbs. In order for us to consider them as a coherent and relevant subclass, they need to meet the following criteria:

1. the semantic role and the morphosyntactic encoding of the adjunct NP needs to be consistent among class members,

2. the predicates have to share at least one syntactic alternation specific to the class,

3. they also have to share a semantic meaning component (in order not to contradict the Semantic Basis Hypothesis).

We can first lay down that the verbs 'beszennyez', 'bemocskol' ('to stain' or 'to dirt'), 'bepiszkít/bepiszkol', morphological variants ('to smear') are close synonyms. We are dealing with morphologically complex predicates: they are composed of a verb prefix ('be' or 'össze' and a stem derived from the noun 'szenny' (dirt, trash), 'mocskos', 'piszkos' ('dirty'). [6] The semantics of the class can be described by two shared meaning components: change of state by material transfer.

What we find when we examine the syntactic characteristics of these verbs is that they have a very simple complementation pattern, allowing a subject, a direct object, the above mentioned 'mass material' NP and, at a less specific level a set of other optional adjuncts (such as instrument, manner, time, location, which we will not take into consideration in what follows, being too general to be used as distinctive features). Thus, any of the verbs listed above will accept the structure exemplified in the following sentence:

1) A diákok megint bepiszkolták a falat festékkel.
the students again smeared the wall paint+INS.
The students smeared the wall with paint.

The verbs 'beszennyez', 'bemocskol', 'bepiszkít/bepiszkol' ('to stain' or 'to dirt') can undergo a morphological passive-like alternation with the suffix -Odik:

[6]The adjectives, in turn, are derived from two synonyms for dirt: piszok, mocsok. We will not discuss the possibility that the verb could in fact be derived from the nouns, as it is not relevant to our classification experiment.

2) Az ősz, a hó, mind beszennyeződik/bepiszkolódik/bemocskolódik ennek a fájdalomnak a hangulatával.

The autumn, the snow, everything get-stained the feeling of this agony+INS.

The autumn, the snow, everything gets stained by the feeling of this agony.

On the other hand, the same structure yields questionable sentences with 'beken'/'összeken' ('to rub'/'to smear'/'to anoint',...), or 'befest' ('to paint'):

3) ? A szőnyeg bekenődik sárral.

The carpet gets smeared with mud.

4) ? A fal befestődik piros festékkel.

The wall gets painted with red paint.

The passive forms of these predicates (3 and 4) are dubious at the best and do not occur in the Hungarian National Corpus – therefore, we did not consider them as part of our verb group (despite the semantic similarity).

It is important to note that in passive-like forms (sentence 2), the transferred material can be expressed with instrumental suffix just as in sentences 1), but it can be replaced with ablative case and a different semantic role (ablative, as we have seen earlier, is a productive marker of the CAUSE semantic role with change of state verbs):

4) A szőnyeg beszennyeződik/bepiszkolódik/bemocskolódik a sártól.

The carpet gets-stained the mud.ABL.

The carper gets stained with mud/from the mud.

At the same time, the ablative NP is ungrammatical with the verbs 'beken'/'összeken' ('to rub'/'to smear'/'to anoint',...), or 'befest' ('to paint'), which underlines that we are dealing with two distinct verb classes:

5) * A szőnyeg bekenődik a sártól.

The carpet gets smeared from the mud.

6) *A fal befestődik a piros festéktől.

The wall gets painted from the red paint.

We therefore distinguish 'beszennyez', 'bemocskol', 'bepiszkít/bepiszkol' as well as any verb sharing the alternations above, as a separate predicate class, while 'beken' and 'befest' do not belong to this class. Neither they form a coherent class apart, as they do not share any alternations.[7]

V.5.2.2 Non-Default Associate

The second class-specific semantic role we are dealing with is a restricted variant of the default comitative. The comitative semantic role was defined as an additional participant of the event denoted by the predicate who assists the agent of the verb to accomplish the said action or undergoes the same event as the patient of the predicate.[8] Therefore, sentences with a comitative allow to be transformed the following way:

János moziba megy Marival.

John cinema+ILL goes Mary+INS.

John goes to the cinema with Mary.

János és Mari moziba mennek.

John and Mary cinema+ILL go+pl+3

John and Mary go to the cinema.

[7]We found that the morphological alternations of the verb 'ken' ('to smear') and its prefixed derivates — as it could be inferred from its English equivalent — show a certain similarity with the English load/spray class. However, we did not find other Hungarian verbs sharing its characteristics and therefore did not consider to include the load/spray class in our classification.

[8]This is actually an additional argument against the Semantic Specificity criterion: the default comitative, strictly speaking, has the same role as the subject with the same semantic implications. Therefore, its semantic role is specific to the verb, without behaving like an argument at the semantic level or like a complement at the syntactic level.

These structures are equally allowed by the predicates with a non-default associate:

János találkozott Marival.
John met Mary+INS.
John met Mary.

János és Mari találkoztak.
John and Mary met+pl+3
John and Mary met.

János veszekedett Marival.
John argued Mary.INS.
John argued with Mary.

János és Mari veszekedtek.
John and Mary argued+pl+3.
John and Mary argued.

Semantically, the non-default associate differs from the default comitative by 1) its semantic/pragmatic obligatoriness (the event cannot take place without the 'associate'), 2) the semantic restriction +HUMAN does not apply to non-defaults (while the comitative becomes instrument if it is -HUMAN):

János veszekszik a számítógéppel.
John argues the computer.INS
John argues with the computer.

The status of this sentence is pragmatically dubious, as the verb forces a non-default associate participant role to the NP 'with the computer', although the computer pragmatically cannot participate in this event. However, unlike in the case of predicates with a comitative, it is incorrect to annotate the non-human NP as an instrument.

Moreover, the basic implication for the default comitative semantic role (namely, if B is a comitative of A in action C, B accomplishes the same action as A) does not necessarily hold for non-default associates: if we consider the Hungarian sentence "János veszekedett Marival." ('John argued with Mary'), it does not imply that the associate Mary also did argue with John, though it implies that she was present and even passively participated in the event. These semantic properties provide us enough motivation to distinguish the two roles and try to explore the verb class which takes this kind of adjunct (which will be considered as a complement if we do not find a coherent predicate class). The verbs examined included *találkozik* (to meet), *csókolózik* (to kiss), *konzultál* (to consult), *játszadozik* (to be playing), *kötekedik* (to tease), *társalog* (to chat), *veszekedik* (to argue). This verb class cannot be delimited by Levin-type diathesis alternations; however, other distributional properties suggest that this verb group forms a natural class of predicates. Along the lines of these arguments one would favor representing the non-default associate as a complement. On the other hand, there exists a syntactic test on the basis of which we can distinguish default and non-default associates syntactically: only the latter allow to be substituted by the reflexive pronoun "egymással" ('with each other'):

*János és Mari moziba mennek egymással.
*John and Mary cinema go+pl+3 each-other+INS.
John and Mary go to the cinema with each oher.

János és Mari veszekednek egymással.
John and Mary quarrel+pl+3 each-other.INS.
John and Mary quarrel with each other.

Conversely, "együtt" ('together') can only be used with the default comitative:

János és Mari együtt mennek moziba.
John and Mary together go+pl+3 to cinema.
John and Mary go to the cinema together.

*János és Mari együtt veszekednek.[9]

John and Mary together quarrel.PL3.

*John and Mary quarrel together.

The pronominal test above raises a theoretical and methodological question. We have presumed so far that the distribution of adjuncts correlate with diathesis alternations; this presumption is based on Levin's work on English verb classes (1993). On the other hand, we found (III.3) that on the semantic level, there is a parallelism between semantic verb classes and the notion of event type used by Komlósy (1993), and Komlósy defines verbs denoting the same event type at the syntactic level by the scope of adjuncts they allow. We have seen, moreover, that adjuncts can change the distribution of predicates. Therefore, we included adjuncts in our notion of syntactic alternation. However, in the tests defined above we are dealing with lexically bound adjunct-type structures, i.e. the basis for delimiting a verb class would be constituted uniquely by the possibility of adding either 'együtt' or 'egymással' to the complementation pattern. One could argue that this is not a syntactic but a lexical phenomenon and as such, cannot serve to distinguish syntactically relevant lexical semantic verb classes. On the basis of this argument, one would be constrained to qualify the non-default associate as a complement of the verbs listed above. This point can also be underlined by the semantically obligatory character of these NPs, although this property cannot be tested on the syntactic level, as they are not syntactically obligatory. Let us remind here that the main stake of our experiment is not to argue that the individual NPs in the examples above are not complements; conversely, we accept that there is a gradual distinction to be made between typical complements, typical adjuncts and adjuncts with more or less reduced productivity. Nevertheless, we aim to discover to what extent and how lexical semantic features affect the syntactic realization of NPs in verbal complementation patterns. In order to examine this question, we attempt to make as many syntactically relevant generalizations in our lexicon as possible. From this perspective, it seemed highly relevant to include this test in our inventory of verb class definitions.

[9]The grammaticality of this sentence is dubious. The only possible semantic interpretation would be that János and Mari quarrel with a third person, which would indicate that we are dealing with a non-default comitative interpretation.

V.5.2.3 Cause (Change in Mental State)

The next class we identified as syntactically relevant is characterized by the meaning components we named 'cause change in mental state'.[10] Verbs belonging to this group are *megdöbbent* ('surprise'), *felidegesít* ('make nervous'), *megrémít* ('horrify'), e.g.:

János megdöbbentette Marit a hírrel.
John surprised Mary.ACC the news.INS
John surprised Mary by telling her the news.

The implications of this semantic predicate are the following: 1) a change occurs in the mental state of the person (+human argument), 2) this change is caused by a situation brought into existence by the entity denoted by the subject. The syntactic alternation characterizing this verb group involves the subject, the case of the other NPs in the complementation structure and the morphology of the verb:

a) János megdöbbentette Marit a hírrel.
John surprised Mary+ACC the news+INS
John surprised Mary by telling her the news.

b) A hír megdöbbentette Marit.
The news surprised Mary+ACC
The news surprised Mary.

c) Mari megdöbbent a hírtől.
Mary was-surprised the news+ABL
Mary was surprised by the news.

[10]We have not examined the internal structure of metapredicates as components of verb meaning. There seems to be motivation to analyze e.g. 'change in mental state' as a complex metapredicate since, as we will see, verbs belonging to this group share some syntactic properties with the broader 'cause change' group. However, giving a complete description of the structural properties of metapredicates falls beyond the scope of the present thesis.

Verbs belonging to this class systematically undergo the alternation illustrated by sentences a-c). The alternating structures b-c) are discussed by (Komlósy, 1992). In the broader context of agentivity and causation, the English equivalent of this alternation is also discussed, among many others, by Fillmore (1968); Van Valin and Wilkins (1996); Levin and Hovav (2005). It is demonstrated (Van Valin and Wilkins, 1996) that non-agent subjects can be realized as oblique complements in this structure, while agents cannot:

Pneumonia killed his uncle.

His uncle died from/of pneumonia.

Brutus killed Caesar.

*Caesar died from/of Brutus.

Van Valin and Wilkins (1996) argue that agentivity is not implied, though a possible interpretation, in the argument structure of the verb kill (as opposed to murder), and attribute the semantic role "effector" to the argument in question. Komlósy, on the other hand, attributes the thematic role "natural force" to the given argument in the alternating structures b-c). Croft (1998) propose an account based on a ranking of participants in terms of their force-dynamic relations to each other, instead of a thematic role hierarchy: "one participant outranks another if it is antecedent to the other in the causal chain (in terms of transmission of force)." (Croft, 1998) In any case, these analyses explain the phenomenon by a distinction in the nature of different types of causations (agentive/volitional vs non-volitional activity, direct on indirect cause): they state that these semantic components vehicle relevant syntactic distinctions, namely in terms of diathesis alternations. If we examine the Hungarian data, what we find is that verbs accepting the contexts in b) and c) share the meaning component 'causation of change' and must have a non-agentive interpretation:

János megdöbbentette Marit. *agentive or non-agentive*

John surprised Mary.

A hír megdöbbentette Marit. *non-agentive*

The news surprised Mary.

A szél becsapta az ablakot. *non-agentive*

The wind shut the window.

Az ablak becsapódott a széltől. *non-agentive*

The window (was) shut by the wind (wind+ABL).

On the other hand, the alternation set completed by sentence a) only applies to verbs indicating 'causation of change in mental state':

János megdöbbentette Marit a hírrel.

John surprised Mary by telling her the news.

*János becsapta az ablakot a széllel.

*John shut the window with the wind.

It seems that if the sentence already has a 'natural force' argument, it is impossible to add a second 'causer' (of any kind); however, if we replace it with a simple instrument, sentence type a) becomes acceptable, but c) goes wrong:

a) János betörte az ajtót a kalapáccsal.

John broke the door+ACC the hammer+INS

John broke the door with the hammer.

b) A kalapács betörte az ajtót.

The hammer broke the door+ACC

The hammer broke the door.

c) *Az ajtó betört a kalapácstól.

The door broke the hammer+ABL.

*The door broke from/by the hammer.

In accordance with the fact that *betör* does not participate in the complete set of

alternations, our system correctly predicts the instrument semantic role to its argument. Verbs in the cause change in mental state group accept all the alternations and are assigned the 'cause' semantic role to their instrumental NP adjunct.

V.5.3 Complements

Our definition of complements is based on full lexicalization: the logic of the system implies that any instance of verb+NP dependency relation which are not covered by the above listed rules are considered as complements. Using our verb frequency list and the development corpus, we created a list of verbs which occur with an NP in instrumental case but 1) the semantic role of the NP does not correspond to any of the default meanings of instrumental case, 2) the verb does not belong to any of the verb classes defined above (based on its alternation patterns or on the semantic role attributed to the NP).

However, our scope of interest remains at the level of *structures*: we did not include nouns lexicalized *with* an instrumental case suffix (e.g. tavasszal (*during the spring*), bizonnyal (*certainly, with certitude*)). Such lexicalized nouns have a lexical meaning which is stable among verbs and is not in relation with other meanings of the case suffix; moreover, their relation to the predicate (if any) does not vary among verbs. Neither did we include lexicalized verb+NP.INS items in which the head noun of the NP is lexically bounded.

Let us hereby develop an important issue about the nature of verbal complementation. As we have seen in 3.2, phrase structure grammars claim that complements, unlike adjuncts, can change the syntactic distribution of the constituent they are added to.[11] We subsequently showed a counter-example where an adjunct modifies the complementation options of a predicate structure. We therefore conclude that both complements and adjuncts can alter the complementation of a predicate. In the case of adjuncts, we presume that such modifications in complementation are predictable, while in the case of complements we are dealing with lexicalized idiosyncratic phenomena. Since adding a complement to a predicate can change its subcategorization in an unpredictable way, this implies that a precise annotation of complements has to rely on deep parsing. Instead

[11]This is only implied in phrase structure grammars assuming X-bar theory; in practice, each of the grammars discussed in III/1 and thereafter belongs to this category.

of individual complements, full subcategorization frames have to be matched against the sentence structure. Our implementation does not follow this method: instead, we try to label individual complements/adjuncts according to their function when we only have information about the predicate they attach to. The obvious reason for this is that our work aims at discovering relevant patterns of complementation and build a database of complement structures based on the findings of this phase of work. Therefore, when evaluating the implementation as a SRL system one should consider that it does not dispose of a deep parsing module.

V.5.4 Implementation as Semantic Role Labeling

During the process of creating our SRL grammars we used a development corpus and a frequency list of verbs from which the first 2.800 entries were retained[12]. The corpus was created from two volumes of 20th century literature: Abigél (1978, novel) by Magda Szabó and Leírás ('Description', 1979, short stories) by Péter Nádas. The corpus was parsed with NooJ using the shallow parser presented in chapter II. The verb list and the corpus were used for defining case suffix functions, forming the predicate groups and coding valence. The rule system was continuously being refined and the dictionary completed with subcategorization information and lexical semantic features (defining predicate groups) during the development period. Concretely, we used the verb list to launch queries on the corpus and retrieve occurrences of the verbs on the list with instrumental NPs. First, these occurrences were categorized, forming intuitively coherent semantic classes according to the semantic role of the instrumental NP. This intuitive grouping served as a basis for our predicate classes. [13] This step was followed by the validation of verb classes, involving the description of class-specific alternations. The definition of the input side of the rules (the criteria for a given rule to apply, e.g. semantic constraints), as well as the design of the rule system was performed in parallel to this step. At last, structures which were not matched by any of the rules were coded as complements in verbal dictionary entries. The rule system was continuously being refined

[12] i.e. all the verbs among the 20.000 most frequent words in the Hungarian National Corpus (Váradi, 2002)

[13] as we expected, we found that many of the predicate classes are relevant for more than one case suffix.

and the dictionary completed during the development period.

SRL grammars were implemented in NooJ, as a part of the complex linguistic analysis module and especially relying on the chunker and the annotated corpus described in chapter II. Lexical and morphological analysis was fully based on dictionaries. The dictionaries we used for text processing contained the 900.000 most frequent word forms of the Hungarian National Corpus, analyzed morphologically by the Humor analyzer (Prószéky and Tihanyi, 1992). A dictionary entry is composed of a word form, its corresponding lemma and morphological code, enhanced by additional semantic and syntactic information. The dictionary was completed by a set of semantic features for nouns and adjectives by merging the morphosyntactic dictionary with the NooJ dictionary of simple lexical units, containing the semantic features from the lexical database of the RIL HAS (see section IV/1). This semantic information is percolated up to the level of NP as described in chapter II. The verbal lexicon was enhanced with the information on verb classes, i.e. predicates belonging to the classes listed above were marked by features referring to the class.

The rules apply to NPs in the text in the order corresponding to their degree of specificity. Note that this is the exact opposite of the order the rules were created (starting from the most productive rules and leading towards verb-specific subcategorization). The three specificity degrees we distinguished were 1) complements (complete lack of generalization), 2) non-default rules (generalization at the predicate-class level), 3) default rules (full generalization). The input is constituted by clause-segmented chunked text (making use of the heuristic clause delimiting grammars presented in II.2.5). Constituent order within the clause is not taken into account and adjacency is not required. The grammars are applied in shortest match mode, i.e. when there is more than one matching NP in the same clause, the one nearest to the predicate is presumed to have the most specific relation to it. The SRL rules themselves are unambiguous, as is the phrase chunking output. However, ambiguous outputs may be produced from semantically ambiguous input, e.g. when an input word has two entries in the dictionary: one annotated as +human and another one annotated as -human. If the headword of an NP is ambiguous, the SRL rules output all the possible interpretations.

The output created in NooJ is an XML annotation. The annotated texts were indexed

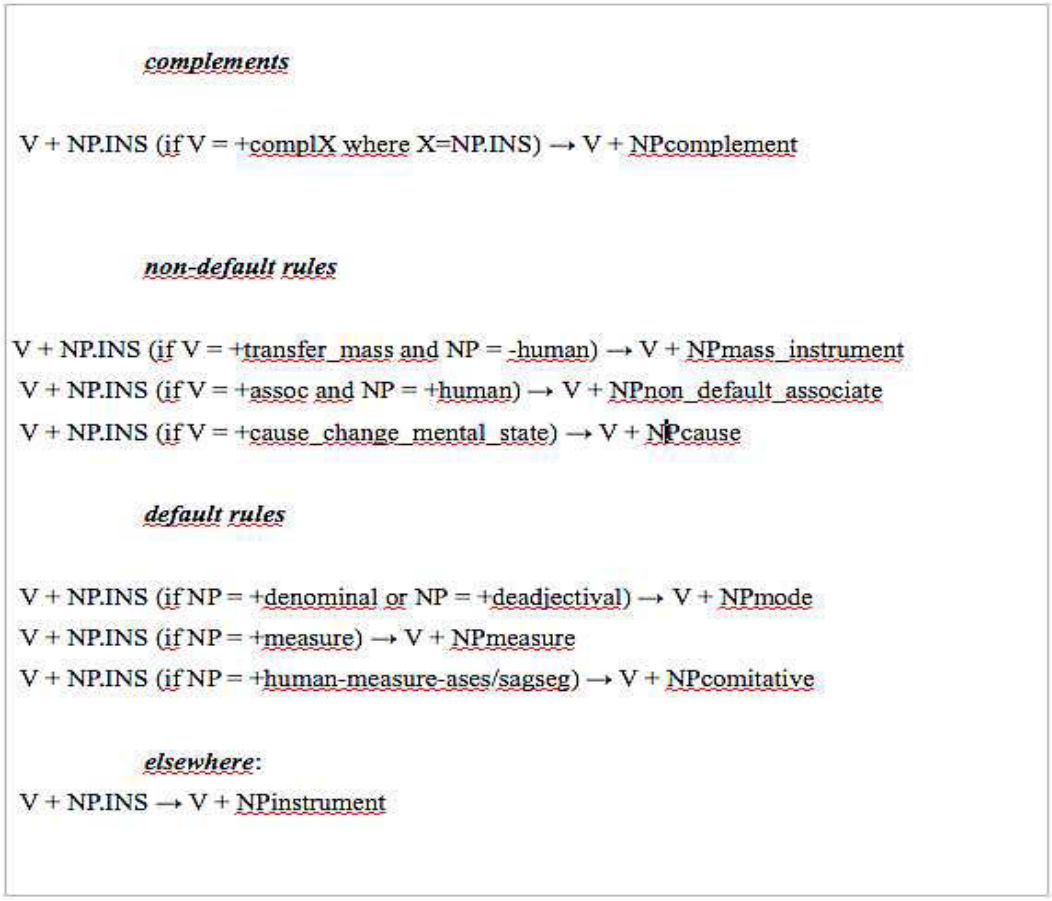

Figure V.1 : Semantic Role Labels and the Corresponding Rules

by the corpus query tool Xaira (XML Aware Indexing and Retrieval Architecture)[14]. Xaira was also used for queries and for displaying annotations during the development period.

V.5.5 Evaluation

A test corpus was created from a distinct literary sub-corpus (several 20th century novels) of our shallow parsed corpus (see II./3). It contains neither direct overlap with the development/training corpus nor extracts from the same literary texts. The test corpus consists of 32.180 sentences and contains 3.849 top-level instrumental NPs (as annotated by the chunker and presumed to have a direct dependency relation with the predicate). Classification accuracy was evaluated as follows. Each annotated predicate with an instrumental NP in its context (top-level NP with a head in instrumental case and within

[14]Xaira was developed by Lou Burnard and Tony Dodd and it is distributed by the Research Technologies Service at Oxford University Computing Services: http://www.oucs.ox.ac.uk/rts/xaira/

the clausal context of a verbal predicate) was classified. Precision was measured for individual roles as follows. The sentential concordance was displayed and the accuracy of the labelling was evaluated a) according to the correctness of the semantic label, in terms of fitting the specific semantic/pragmatic implications defined above; b) according to the applicability of syntactic alternations specific to the verb group, if any. The label was considered as accurate if it met both the semantic and the syntactic criteria. The default instrument label made up 58% of the labels assigned to instrumental NPs with 2244 hits – only a random sample of 1070 occurrences were manually evaluated among them.

In cases where more than one semantic role was appropriate, the more specific one was considered as correct, as only the more specific one can eventually be considered as providing an extensive description. In the case of systematic ambiguity, which occurs between default mode and default instrument roles when the instrument is associated with a specific mode of accomplishment of an action (e.g. *késsel eszik*, to eat with a knife), it is systematically the instrumental role that has been assigned and it has been considered as a precise label. Evaluation was carried out on the complete set of NP candidates to be assigned a label, including both correctly recognized and false candidates (i.e. incorrectly parsed NPs). Therefore, the table below summarizes results on the toolchain performing constituent segmentation and classification accuracy for semantic roles:

Semantic Role	Label Frequency	Correct Labels	Precision
complement	371	271	0.73
non-default associate	33	33	1
transfer mass	1	1	1
mental change	6	5	0.83
measure	7	3	0.42
mode	553	378	0.68
instrument	2244	760	0.36
ALL	3646	1681	0.46

Table V.1 : Precision of our classification for the task of SRL

While evaluating the precision of the labels assigned to NPs, erroneous assignments were categorized according to which other class the example belongs to (if any); false NP

candidates were marked as parse errors. This allowed us to calculate a precision only on correct input: among 3136 relevant candidates, 1757 were assigned a correct label, which corresponds to a precision of 55%. The tables below illustrate assignment error types: colums correspond to the correct label for the given entry, rows indicate the assigned labels. "Err" refers to parse errors.

Matched by	Compl	Assoc	Mass	Mental	Meas	Mode	Comit	Instr	Err
compl (371)	**271**	0	0	0	0	41	3	7	49
assoc (33)	0	**33**	0	0	0	0	0	0	0
mass (1)	0	0	**1**	0	0	0	0	0	0
mental (6)	0	0	0	**5**	0	1	0	0	0
measure (7)	0	1	0	0	**3**	2	0	0	1
mode (378)	101	1	0	3	0	**378**	9	3	58
comit (431)	91	25	0	0	0	12	**230**	7	66
instr (1070)	172	55	6	8	0	183	49	**386**	160

Table V.2 : Precision for individual semantic roles

V.5.6 Position with respect to current SRL methods

Due to the fact that our work is the first one in Hungarian semantic annotation, direct comparison with other systems on the same dataset is not possible. Moreover, our evaluation scores can be interpreted as SRL results with several distinctions with state of the art SRL in mind. First, the systems cited hereafter use pre-defined sets of semantic roles and usually dispose of corpora annotated with these roles. The importance of this condition is twofold: first, unlike in our experiment, the semantic label set has already been tested, confronted to data and validated for use in SRL systems (though these aspects remain questioned). Second, a certain amount of probably relevant information, such as the interaction/correlation between the distribution of individual roles, can be extracted from such annotated resources, while our system at its current degree of completion does not allow for such investigations. Correspondingly, when semantically annotated corpora are available for a given language, SRL is almost exclusively done by supervised machine learning methods. Researchers also often have the choice to use the tag set (set of semantic roles) which is best adapted to the learnability criteria and generalization

capacity of their system (on an independent evaluation of the learnability of semantic role sets, see (Merlo and van der Plas, 2009)). Moreover, most of these systems rely on high-precision outputs of full parsers, which we do not dispose of for our experiments. As concluded by Carreras and Marquez (2005) from the CONLL shared task results on SRL, systems relying on full parsing achieve 10% higher F-measures in this task compared to systems using partial parsing. For instance, the feature set used in the supervised SRL experiment of Gildea and Jurafsky (2002) include the grammatical function of the constituent, its position in the sentence and the complete parse tree path, while we have to limit ourselves to information from partial parsing. Finally, as noted in the introduction of the present chapter, our system does not entirely correspond to an SRL system in its functionalities. We do not annotate subjects, objects and oblique complements according to their semantic role: we limit ourselves to productive structures [15], while also undertake the task of delimiting those from fully lexicalized dependencies.

The seminal article by Guildea and Jurafsky (2002) presents a supervised learning experiment for SRL based on statistical classifiers trained on roughly 50,000 sentences hand-annotated with FrameNet semantic roles. Each training sentence is parsed into a syntactic tree and various lexical and syntactic features are extracted and used for the semantic classification. Information on the prior probabilities of various combinations of semantic roles was equally taken into account. The system achieves 82% accuracy in identifying the semantic role of pre-segmented constituents, while it achieves 65% precision and 61% recall on the task of simultaneous segmenting and classification. Several other methods were proposed along this line of research (i.e., using different classification or sequential labeling methods and a combination of lexical and syntactic features extracted from semantically annotated corpora). There has been significant evolution in SRL results since (Guildea and Jurafsky, 2002). As Marquez (2009) notes, the choice of the learning algorithm actually has less influence on the results than the choice of the features. However, our task is more similar in nature to unsupervised or cross-lingual semantic role induction, as their methods do not rely on manually annotated training data. Actually, the only alternative to our experiments would have been to resort to either an unsupervised sense induction method or to attempt to transpose semantic role annotation from

[15] although we use a broadened notion of productivity

another language where labelled data was available. However, cross-lingual semantic role induction experiments have only been tested for relatively close language pairs: Padó and Lapata (2009) project role-semantic annotations from English to German corpora, while the English corpus is semantically annotated. Unsupervised approaches are, in turn, far less frequently used for SRL. The experience of Swier and Stevenson (2004) is based on a bootstrapping method for semantic role induction, in which the probability model is iteratively updated by adding previously labeled constituents to the training set. Unfortunately, their results are not directly comparable to those obtained by supervised ML methods.

V.5.7 Discussion and Error Analysis

Let us remind that the goal of our experiment was twofold. First, it was an attempt to verify our hypothesis that semantic verb classes underlie class-specific alternations in Hungarian by allowing semantically compatible adjunct-type NPs to appear with a given case suffix. Second, this hypothesis was confronted to the data by virtue of implementing a set of grammars for syntactic and semantic function labeling. The aspects of evaluation include: a) the usability of the role set, i. e. whether it is possible to assign a role unambiguously to the major part of NPs in the corpus, b) quality and relevance of the predicate classification and the corresponding syntactic features, i. e. the possibility to assign more specific and more precise semantic roles to predicate class-specific adjuncts on the basis of syntactic alternations, c)the accuracy of classification, i. e. the precision of our grammars and the coverage and precision of its input features (besides predicate classes, semantic features for nouns).

The most prominent part of the errors are due to incomplete lexical coding. This is revealed on the one hand by the fact that errors are concentrated in the default area, i.e. the rules that use the least of lexical information contain the most errors. This probably indicates that those structures should have been recognized by more specific grammars (consider e.g. the amount of complements erroneously annotated as instrument adjuncts). On the other hand, it can be globally concluded that most of the errors lean in the direction from the less specified towards the more specified annotation and not the opposite. This comes from the fact that although the most frequent verbs were confronted to the test

corpus, lexical encoding only covers the test corpus vocabulary, which is apparently not satisfying yet for new corpora. Moreover, in the case of mode adjuncts, our solution was overtly approximative and involves a particular problem: namely, certain nouns produced by the suffix -ás/és denote objects resulting from an action (e.g.*leírás*, description) rather than the action itself and are therefore unable to fill the function of mode adjunct. On the other hand, several frequent lexicalized items express mode systematically, while not being produced by the said derivation (e.g. *szándékkal*, with the intention of); hence the low recall of the mode adjunct rule.

Errors in complement detection (besides parse errors) are caused by two factors: either the verb can have an instrumental complement, but it is not obligatory and thus the instrumental NP in the sentence is an adjunct, or the complement is present in the clause but is not the instrumental NP which is closest to the predicate. Both reasons are possible, and the grammars in their current form do not allow other types of disambiguation than by the relative distance from the predicate.

Two important structural lacks of the system are the inaccurate treatment of factitive verbs and the problem of coreference resolution. Factitive verbs are produced by a productive morphological operation; sentences with a factitive predicate should ideally be treated as a syntactic construction on its own, as e.g. negation or focusing. NPs with instrumental case appear occasionally, though not very frequently in this construction, since the syntactic operation of factitive produces a structure in which instrumental case encodes the subject function of the base verb. As our shallow parser currently does not support factitive as a productive syntactic construction on its own, for now, these structures were treated as 'parse errors'.

Information about coreference resolution is not included in our shallow parser outputs. This means that a certain amount of pronouns, which can either refer to objects or abstract notions are not provided any semantic features, since these features entirely depend on the reference of the pronoun in context. This produces higher error rates mostly with respect to the default instrument rule, as this rule applies to "default" cases where semantic features are typically absent.

The SRL grammars are applied to the output of a shallow parser. Lexical features are treated as binary and unambiguous specifications: if a lemma has a feature in the

dictionary, it is assigned to every occurrence of this lemma while annotating the text. This means that semantic ambiguity is not taken into account, neither for verbs nor for NPs. As we do not dispose of a WSD system, we cannot satisfactorily deal with this problem at present. It is important to add, however, that any WSD system presupposes the existence of a sense inventory, adapted or adaptable for NLP tasks; currently, as noted by Héja et al. (2009), we lack such inventories for Hungarian.

The very encouraging aspect of the evaluation is that the verb classes and the semantic roles provide good coverage to the data, since it is possible to manually annotate the instrumental NPs in the corpus with respect to their semantic roles and according to our guidelines. The fact that specific rules work with a high precision compared to default rules shows that the rules we stated are grammatically reliable; on the other hand, errors are due to the fact that a) the input is often left unrecognized as such (incomplete lexical encoding), b) shallow parsing does not provide sufficient information on a set of structures (such as coreference resolution; see above).

The most striking fact about evaluation scores is the low presence of some non-default structures. It is not exclusively due to recall issues: the manual annotation shows that non-default structures like change in mental state and material transfer occur sporadically in the literature corpus. On the other hand, the two most frequent groups, complement NPs and instrument NPs make up the biggest part of the structures to be classified. It is notable, however, that we can find additional patterns inside the group of complements: there seem to be other, unexplored verb classes with potentially coherent semantic role assignment with respect to instrumental case. One typical example is made up by verbs sharing the same prefix, e.g. *szembe/kerül, szembe/száll, szembe/megy, szembeszegül* (to oppose, to rise against, to face, to defy) – all of them having an instrumental NP expressing the opposite force. These verbs were not classified into a lexical semantic group since we did not have a corresponding syntactic test. The large number of complement NPs in the test set suggests that some additional verb groups may have been left unanalyzed and have been classified as a complement.

With respect to the SRL task, we conclude from the results that our future work has to concentrate on the improvement of the vocabulary on one hand. On the other hand, since our rule system is a relatively high level application, we have to exclude the mistakes

originating from the lower-level rules.

V.6 Conclusion

In this section, we have presented a methodology for manually exploring lexical semantic verb classes, class-specific syntactic structures and adjunct types. Our method was data-oriented, partly due to the fact that lexical semantic verb classes have not been studied yet in detail in Hungarian, and partly because we wanted to ensure immediate feedback on the accuracy and the coverage of the classification provided by the method. We took the distribution of case suffixes as a starting point and classified top-level NPs according to their syntactic/semantic function encoded by the suffix. We hypothesized that natural classes of predicates allow for the same types of adjuncts. When defining a predicate class, the hypothesis was verified by syntactic tests specific to the given group. We found that the notion of alternation needed to be extended and a broader set of syntactic tests has been used in order to account for systematic adjunct role assignment phenomena. Subsequently, we implemented the adjunct and predicate classification model as a syntactic/semantic role labeling system for the instrumental case suffix. The system consists of an ordered list of grammars that assign syntactic/semantic function labels to top-level NPs in instrumental case according to their relation to the predicate, using the information from shallow parsing and a set of lexical semantic features. We evaluated the grammar output on a corpus of literary texts, manually pre-annotated with semantic roles and performed an error analysis. We concluded that the criteria established for verb classification and adjunct typology is usable for manual annotation, thus, the function labels and the syntactic tests are adequate. It is a positive aspect of the results that several subcategories of NPs in verbal complementation pattern could be re-analyzed as adjuncts, which means that we increased the predictive capacity and the generalization power of our lexical representation. Moreover, the systematic study of verb classes showed that the nature of alternations in Hungarian is not purely syntactic but also morphological, and extends beyond the scope of diathesis alternations as defined by Levin (1993, 2005). Although the available resources are scarce (lack of a reliable sense inventory, a deep parser, WSD or anaphora resolution module), the results on the automatization of

semantic role annotation are encouraging. However, the manual definition of verb classes and corresponding semantic role assignment is a very demanding task, both in terms of time and in terms of linguistic expertise. Moreover, the evaluation of the automatic annotation work revealed that significant amount of effort has been invested in the study of certain verb classes which proved to be less relevant for the annotation task, due to the low frequency of verbs with the given complementation pattern from to these classes.

On the one hand, we believe that thorough manual work is inevitable in order to provide a Levin-type description of lexical semantic verb classes for Hungarian and to integrate the results to our verb lexicon. On the other hand, this work can be facilitated by unsupervised learning from existing resources. Automatic extraction of lexical semantic information from corpora has become widely popular in the last years. In the following section, we present an experiment on the acquisition of lexical semantic verb classes by unsupervised distributional clustering.

Chapter VI

Automatic Acquisition of Verb Classes by Unsupervised Clustering

VI.1 Lexical Acquisition and Semantic Features

The present chapter deals with lexical acquisition, i.e. the study of machine learning methods aiming to acquire lexical properties of individual words from text corpora. Our investigation builds on the hypothesis that semantic similarity underlies similar syntactic behavior (Semantic Basis Hypothesis, SBH) and therefore, it is possible to cluster words by their syntactic contexts in a corpus and obtain semantic word classes. We focus on verbal subcategorization and semantic meaning components as defined in (Levin, 1993). Throughout our experiments, we are seeking the answer to two questions:

1. Are the resulting clusters semantically coherent (thus reinforcing the Semantic Basis Hypothesis)?

2. If so, what kind of syntactic information can be correlated to the semantic properties, i.e. what are the alternations responsible for their similar behavior?

VI.1.1 Definition of verb classes

"Generally, the definition of a verb's semantic class can be considered as part of its lexical entry, next to idiosyncratic information: the semantic class generalises as a type definition over a range of syntactic and semantic properties, to support Natural Language

Processing in various areas..." (Schulte im Walde, 2000)

Part of speech categories constitute the basic generalization over word classes: words belonging to the same category behave similarly when they are put in a sentence. Syntactic rules describe how to build up sentences from in terms of POS categories. However, semantic subclasses of words show similar syntactic behavior, whereas semantically different members of the same POS category do not appear in identical syntactic contexts (Harris, 1954). Starting from this observation, a big amount of research has been carried out aiming to give a systematic description of the relation between lexical semantics and syntactic subcategorization. The present work is inspired by (Levin, 1993) and Levin and Hovav (2005) who suggest to represent verbal meaning by a set of meaning components. Verbal meaning can be described by a core meaning which is specific to each lexical item, and meaning components which represent generalisations over semantic properties. Verbs sharing one or more meaning components can be classified in the same semantic group. On the other hand, meaning components determine verbal syntax to a certain extent: they affect verbs' participation in diathesis alternations. Thus, semantic verb groups can be inferred from the set of diathesis alternations which characterise the given verbs. Her hypothesis is that classifying verbs according to their participation in diathesis alternations leads to classes of verbs sharing at least one relevant meaning component.

VI.1.2 State of the Art in Lexical Acquisition

For over a decade, automatic construction of wide coverage structured lexicons has been in the center of interest in the natural language processing community. On the one hand, structured lexical databases are easier to handle and to expand because they allow making generalizations over classes of words. On the other hand, interest in the automatic acquisition of lexical information from corpora is due to the fact that manual construction of such resources is time-consuming, and the resulting database is difficult to update. Moreover, the flourishing interest in the lexical module of the grammar, as well as the increasing quantity of available corpora have provided motivation to apply the distributional hypothesis in order to automatically acquire lexical properties from corpora. As demonstrated by Kipper et al. (2008), lexical acquisition methods are useful for constructing and extending lexical resources by relevant information. Current research

directions in the automatic acquisition of lexical properties concentrate on the acquisition of verbal subcategorization frames (Brent, 1991; Manning, 1993; Briscoe and Carroll, 1997; Sass, 2009; Korhonen, 2002), selectional restrictions (Resnik, 1993), and (syntactically relevant) semantic properties of lexical items. In what follows, we are going to focus on the latter field.

Semantic properties are usually understood as relations of synonymy, hypo- and hypernymy or the semantic verb classes defined in the previous section. Among sense induction experiments for nouns based on distributional similarity, Pantel and Lin (2002); Véronis (2004) focus on unsupervised word sense induction from monolingual corpora, whereas van der Plas et al. (2008); de Cruys and Apidianaki (2011) extend the method towards using multilingual aligned corpora. Hearst (1992) presents experiments on the acquisition of hyponymy and hypernymy relations. Fabre and Bourigault (2006) use distributional methods to extract morphosemantic derivational realtions between words of different grammatical categories from corpora.

Most of the work in the field of acquisition of verbal lexical properties aims at learning subcategorization frames from corpora. However, semantic grouping of verbs on the basis of their syntactic distribution or other quantifiable features has also gained attention. The goal of these investigations is either the validation of verb classes based on (Levin, 1993) (Schulte im Walde, 2000; Korhonen et al., 2003, 2000), finding algorithms for the categorization of new verbs (Korhonen and Briscoe, 2004), or identifying alternations(Lapata, 1999). Supervised and unsupervised approaches are equally popular amongst the experiments on the acquisition of verb classes. However, supervised approaches require considerable linguistic expertise as they are based on linguistically motivated statistical indicators, i.e. quantifiable approximations of diathesis alternations. It is not surprising that supervised approaches are mostly used for experiments on English, since other languages usually do not dispose of such an extensive inventory of diathesis alternations. Since alternations are language-specific, these methods have the inconvenient of requiring repeated effort in order to be extended to a new language. The approximations are mapped into a feature space representation which can be extracted from corpora and quantified. Merlo and Stevenson (2001) approximate diathesis alternations by hand-selected grammatical features. They try to classify verbs in the three major classes of optionally

intransitive verbs in English: unergative, unaccusative, and object-drop verbs. The property distinguishing among these three classes as are the thematic roles assigned by the verbs. On the other hand, verbs across these three classes allow the same subcategorization frames; thus, classification based on subcategorization alone would not distinguish them. The method thus largely relies on the heuristic approximation of information about thematic roles. The authors assume that a thematic role is a label taken from a fixed inventory of grammaticalized semantic relations. However, as they admit in the paper, notions as agent or theme lack formal definitions– even though they are assumed to be universal across languages, their definitions diverge across linguistic theories. This makes it even more difficult to extend the usability of the approach to new verb classes or languages. To overcome this drawback, Joanis and Stevenson (2003) developed a general feature space for supervised verb classification. Stevenson and Joanis (2003) investigate the applicability of this general feature space to unsupervised verb clustering tasks. As unsupervised methods are more sensitive to noisy features, the key issue is to filter out the large number of probably irrelevant features. They propose a semi-supervised feature selection method which outperforms both hand-selection of features and usage of the full feature set.

The method described by Brent (1993) focuses on alternation extraction rather then verb classification. It is based on local morphosyntactic cues instead of complete parses of entire sentences. Although the morphosyntactic patterns used in this experiment are easier to define and hence require less effort from linguists, it still remains language-dependent. Both Brent (1993) and Merlo's supervised approach (Merlo and Stevenson, 2001) have the advantage that they do not require the corpus to be parsed. Moreover, Brent applies regular patterns to raw corpora, i.e. his method does not require POS-tagging – however, this constitutes a limitation of his approach as probability estimations are biased by noun-verb ambiguities.

Unsupervised methods for alternation detection and verb clustering use syntactic features extracted from parsed corpora Schulte im Walde (2000); Schulte im Walde and Brew (2002); Briscoe and Carroll (1997); Korhonen et al. (2003). Subcategorization frame frequencies are potentially completed by semantic selection information. Verbs are represented as relative frequencies of subcategorization frames found in the context of the

verb. Despite the usually very high number of different subcategorization frames that constitute the feature space, these experiments have shown that using fine-grained syntactic distinctions and including adjuncts in subcategorization frames improve clustering results. However, as Schulte im Walde (2000) notes, adding information on semantic selection leads to sparse data problem and lower performance: feature spaces limited to syntactic information yield better results. On the other hand, Alishahi and Stevenson (2007); Li and Brew (2008) have recently reported success on introducing lexical semantic features in the feature space in an efficient way.

The experiments described above mostly or exclusively concentrate on English. However, in recent years state-of-the-art results have been reported for other languages. There are two directions in the work on the extraction of subcategorization frames for French. Since several extensive lexical-syntactic resources have been created manually or semi-automatically for French, such as the Lexicon-Grammar (Gross 1975, 1992), the DicoValence (van den Eynde and Mertens, 2006), the Lefff (Sagot et al., 2006), or the Dictionnaire Syntaxique des Verbes Français (Dubois and Dubois-Charlier, 1997), the first line of research tries to use these resources to extract information about verbs' lexical properties. Gardent et al. (2006) extracts subcategorization frames from the Lexion-Grammar. Falk (2008) uses three subcategorization lexicons to identify relevant semantic verb classes in French with the Formal Concept Analysis method. These approaches do not use frequency information but rely on manual work. Like the approaches above, (Kupsc and Abeillée, 2008) rely on manually constructed resources: they use the French Treebank (Abeillée et al., 2003) to extract Treelex, a verbal subcategorization frame lexicon. Although the method is corpus-based, it exploits a manually built treebank and hence remains limited in coverage. (Chesley and Salmon-Alt, 2006) present an exploratory study on subcategorization extraction from a corpus. They use an automatically parsed literary corpus and obtain results for 104 verbs. The first large-scale subcategorization extraction experiment is described in (Messiant, 2008 and Messiant et al., 2008): he uses the subcategorization frame filtering method described in (Korhonen et al., 2000), and the resulting subcategorization lexicon LexSchem contains syntactic frames for over 4000 verbs. (Gardent and Lorenzo, 2010) report on an ongoing project which also aims at extracting subcategorization frames from corpora. As to semantic verb classification, (Falk, 2008) describes

the first attempt to automatically create semantic verb classes for French. She uses three manually created dictionaries to infer verb classification by Formal Concept Analysis.

Due to the limited quantity of syntactically annotated corpora, as well as the long-lasting lack of an available large-scale parser, only a few studies concentrated on lexical acquisition for Hungarian. Simon et al. (2010) compare three different methods of subcategorization extraction applied to Hungarian data. Sass (2007) experiments with learning verb classes on the basis of lexical context features. The work reported hereafter is the first attempt to automatically learn Hungarian and French semantic verb classes from corpora.

VI.2 Unsupervised Acquisition of Hungarian Verb Classes

(The work presented in this section results was carried out in collaboration with Enikö Héja) and overlaps with (Gábor and Héja, 2007)

This section reports an attempt to apply an unsupervised clustering algorithm to a Hungarian treebank in order to obtain semantic verb classes. Our hypothesis is that semantic meaning components underlie syntactic realization of verbs. We investigate how one can obtain semantically motivated verb classes by syntactic clustering. This approach is in line with current research directions to define semantic properties based on distributional evidence rather than manually built and thus intuition-based resources such as WordNets or traditional dictionaries.

The 150 most frequent Hungarian verbs were clustered on the basis of their complementation patterns, yielding a set of basic classes and hints about the features that determine verbal subcategorization. The resulting classes serve as a basis for the subsequent analysis of their alternation behavior. Since we do not dispose of a Levin-type classification for Hungarian, we do not know the classes to obtain. Therefore, we need to apply unsupervised methods. This most related work to this experiment are those of Schulte im Walde and Brew (2002); Joanis and Stevenson (2003). However, unlike these projects, we report an attempt to cluster verbs on the basis of their syntactic properties with the further goal of finding the semantic classes relevant for the description of Hungarian verbs'

alternation behavior. Another significant differenc is that while in English semantic argument roles are mapped to configurational positions in the tree structure, Hungarian codes complement structure in its highly rich nominal inflection system. We are thus looking for a machine learning method which works reliably on non-configurational distributional data. Therefore, we start from the examination of case-marked NPs in the context of verbs. As we do not have presuppositions about which classes have to be used, we chose an unsupervised clustering method described in (Schulte im Walde, 2000). The 150 most frequent Hungarian verbs were categorized according to their complementation structures in a syntactically annotated corpus, the Szeged Treebank.

VI.2.1 Feature Space

As of present, no full parsers are available for Hungarian. Automated syntactic analysis cannot be used satisfactorily for extracting verbal argument structures from Hungarian corpora, hence, the experiment was carried out using a manually annotated Hungarian corpus, the Szeged Treebank (Csendes et al., 2004). Texts of the corpus come from different topic areas such as business news, daily news, fiction, law, and compositions of students. It currently comprises 1.2 million words with POS tagging and syntactic annotation which extends to top level sentence constituents but does not differentiate between complements and adjuncts. When applying a classification or clustering algorithm to a corpus, a crucial question is which quantifiable features reflect the most precisely the linguistic properties underlying word classes. Schulte im Walde (2000); Schulte im Walde and Brew (2002); Briscoe and Carroll (1997) use subcategorization frame frequencies obtained from parsed corpora, potentially completed by semantic selection information. Merlo and Stevenson (2001) approximate diathesis alternations by hand-selected grammatical features. While this method has the advantage of working on POS-tagged, unparsed corpora, it is costly with respect to time and linguistic expertise. As in our experiment we do not have a pre-defined set of semantic classes, we need to apply unsupervised methods. Neither have we manually defined grammatical cues, not knowing which alternations should be approximated. Hence, similarly to (Schulte im Walde, 2000), we represent verbs by their subcategorization frames. In accordance with the annotation of the treebank, we included both complements and adjuncts in subcategorization patterns. It is important to note,

however, that not only practical considerations lead us to this decision. First, there are no reliable syntactic tests for differentiating complements from adjuncts. This is due to the fact that Hungarian is a highly inflective, non-configurational language, where constituent order does not reveal dependency relations. Indeed, complements and adjuncts of verbs tend to mingle. In parallel, Hungarian presents a very rich nominal inflection system. Second, we believe that adjuncts can be at least as revealing of verbal meaning as complements are: many of them are not productive (in the sense that they cannot be added to any verb), they can only appear with predicates the meaning of which is compatible with the semantic role of the adjunct. For these considerations we chose to include both complements and adjuncts in subcategorization patterns.

Subcategorization frames to be extracted from the treebank are composed of case-marked NPs and infinitives that belong to a children node of the verb's maximal projection. As Hungarian is a non-configurational language, this operation simply yields a non-ordered list of the verb's syntactic dependents. The order in which syntactic dependents appear in the sentence was not taken into account. There was no upper bound on the number of syntactic dependents to be included in the frame. Frame types were obtained from individual frames by omitting lexical information as well as every piece of morphosyntactic description except for the POS tag and the case suffix. The generalization yielded 839 frame types altogether.

VI.2.2 Clustering Method

In accordance with our goal to set up a basis for a semantic classification, we chose to perform the first clustering trial on the 150 most frequent verbs in the Szeged Treebank. The representation of verbs and the clustering process were carried out based on (Schulte im Walde, 2000). The data to be compared were the maximum likelihood estimates of the probability distribution of verbs over the possible frame types:

$$p(t|v) = \frac{f(v,t)}{f(v)}$$

with $f(v)$ being the frequency of the verb, and f(v, t) being the frequency of the verb in the frame. These values have been calculated for each of the 150 verbs and 839 frame types. Probability distributions were compared using Kullback-Leibler divergence as a

distance measure:

$$D(x\|y) = \sum_{i=1}^{n} x_i \cdot \log \frac{x_i}{y_i}$$

Due to the large number of subcategorization frame types, verbs' representation comprise a lot of zero probability figures. Using relative entropy as a distance measure compels us to apply a smoothing technique to be able to deal with these figures. However, we do not want to lose the information coded in zero frequencies - namely, the presumable incompatibility of the verb with certain semantic roles associated with specific case suffixes. Since we work with the 150 most frequent verbs, we wish to use a method which is apt to reflect that a gap in the case of a high-frequency lemma is more likely to be an impossible event than in the case of a relatively less frequent lemma (where it might as well be accidental). That is why we have chosen the smoothing technique below:

$$\begin{aligned} f_{\mathrm{e}} &= \frac{0,001}{f(v)} \quad \text{if} \\ f_{\mathrm{c}}(t,v) &= 0 \end{aligned} \tag{VI.1}$$

where f_e is the estimated and f_c is the observed frequency.Two alternative bottom-up clustering algorithms were then applied to the data:

1.First we employed an agglomerative clustering method, starting from 150 singleton clusters. At every iteration we merged the two most similar clusters and re-counted the distance measures. The problem with this approach, as Schulte im Walde notes about her experiment, is that verbs tend to gather in a small number of big classes after a few iterations. To avoid this, we followed her in setting to four the maximum number of elements occuring in a cluster. This method — and the size of the corpus — allowed us to categorize 120 out of 150 verbs into 38 clusters, as going on with the process would have led us to considerably less coherent clusters. However, the results confronted us with the chaining effect, i.e. some of the clusters had a relatively big distance between their least similar members.

2. In the second experiment we put a restriction on the distance between each pair of verbs belonging to the same cluster. That is, in order for a new verb to be added to

a cluster, its distance from all of the current cluster members had to be smaller than the maximum distance stated based on test runs. In this experiment we could categorize 71 verbs into 23 clusters. The convenience of this method over the first one is its ability to produce popular yet coherent clusters, which is a particularly valuable feature given that our goal at this stage is to establish basic verb classes for Hungarian. However, we are also planning to run a top-down clustering algorithm on the data to get a probably more precise overview of their structure.

VI.2.3 Results

With both methods we describe in Section 3, a big part of the verbs showed a tendency to gather together in a few but popular clusters, while the rest of them were typically paired with their nearest synonym (e.g.: zár (close) with végez (finish) or antonym (e.g.: ül (sit) with áll (stand)). Naturally, method 1 (i.e. placing an upper limit on the number of verbs within a cluster) produced more clusters and gave more valuable results on the least frequent verbs. On the other hand, method 2 (i.e. placing an upper limit on the distance between each pair of verbs within the class) is more efficient for identifying basic verb classes with a lot of members. Given our objective to provide a Levin-type classification for Hungarian, we need to examine whether the clusters are semantically coherent, and if so, what kind of semantic properties are shared among class members. The three most popular verb clusters were investigated first, because they contain many of the most frequent verbs and yet are characterized by strong inter-cluster coherence due to the method used. The three clusters absorbed one third of the 71 categorized verbs. The clusters are the following:

C-1 VERBS OF BEING: *marad* (remain), *van* (be), *lesz* (become), *nincs* (not being)

C-2 MODALS: *megpróbál* (try out), *próbál* (try), *szokik* (used to), *szeret* (like), *akar* (want), *elkezd* (start), *fog* (will), *kíván* (wish), *kell* (must)

C-3 VERBS OF MOTION: *indul* (leave), *jön* (come), *elindul* (depart), *megy* (go), *kimegy* (go out), *elmegy* (go away)

Verb clusters C-1 and C-3 exhibit intuitively strong semantic coherence, whereas C-2 is best defined along syntactic lines as "modals". A subclass of C-2 is composed of verbs

which express some mental attitude towards undertaking an action, e.g. (szeret (like), akar (want), kíŋván (wish)), but for the rest of the verbs it is hard to capture shared meaning components. It can be said in general about the clusters obtained that many of them can be anchored to general semantic metapredicates or one of the arguments' semantic role, e.g.: CHANGE OF STATE VERBS (erősödik (get stronger), gyengül (intransitive weaken), emelkedik intransitive rise)), verbs with a beneficiary role (biztosíŋt (guarantee), ad (give), nyújt (provide), készíŋt (make)), VERBS OF ABILITY (sikerül (succeed), lehet (be possible), tud (be able, can)). Some clusters seem to result from a tighter semantic relation, e.g. VERBS OF APPEARANCE or VERBS OF JUDGEMENT were put together. In other cases the relation is broader as verbs belonging to the class seem to share only aspectual characteristics, e.g. AGENTIVE VERBS OF CONTINUOS ACTIVITIES (ül (be sitting), áll (be standing), lakik (live somewhere), dolgozik (work)). At the other end of the scale we find one group of verbs which "accidentally" share the same syntactic patterns without being semantically related (foglalkozik (deal with sg), találkozik (meet sy), rendelkezik (dispose of sg)).

VI.2.4 Evaluation and Discussion

As Sabine Schulte im Walde (2009) notes, there is no widely accepted practice of evaluating semantic verb classes. She divides the methods into two major classes. The first type of methods assess whether the resulting clusters are coherent enough, i. e. elements belonging to the same cluster are closer to each other than to elements outside the class, according to an independent similarity/distance measure. However, relying on such a method would not help us evaluating the semantic coherence of our classes. The second type of methods use gold standards. Widely accepted gold standards in this field are Levin's verb classes or verbal WordNets. As we do not dispose of a Hungarian equivalent of Levin's classification — that is exactly why we experiment with automatic clustering — we cannot use it directly. We also run across difficulties when considering Hungarian verbal WordNet (Kuti et al., 2007) as the standard for evaluation. Mapping verb clusters to the net would require to state semantic relatedness in terms of WordNet-type hierarchy relations. However, if we try to capture the distance between verbal meanings by the number of intermediary nodes in the WordNet, we face the problem that the semantic

	acc	ins	abl	ela
indul	-	ins/com	source	source
jön	-	ins/com	source	source
elindul	-	ins/com	source	source
megy	-	ins/com	source	source
kimegy	-	ins/com	source	source
elmegy	-	ins/com	source	source

Table VI.1 : The semantic roles of cases beside C-3 verb cluster

distance between mother-children nodes is not uniform. As our work is about obtaining a Levin-type verb classification, it could be an obvious choice to evaluate semantic classes by collecting alternations specific to the given class. Hungarian language hardly lends itself to this method because of its peculiar syntactic features. The large number of subcategorization frames and the optionality of most complements and adjuncts yield too much possible alternations. Hence, we decided to narrow down the scope of investigation. We start from verb clusters and the meaning components their members share. Then we attempt to discover which semantic roles can be licenced by these meaning components. If verbs in the same cluster agree both in being compatible with the same semantic roles and in the syntactic encoding of these roles, we consider that they form a correct cluster. To put it somewhat more formally, we represent verb classes by matrices with a) nominal case suffixes in columns and b) individual verb lemmata in rows. The first step of the evaluation process is to fill in the cells with the semantic roles the given suffix can code in the context of the verb. We consider the clusters correct, if the corresponding matrices meet two requirements:

- They have to be specific to the cluster.
- Cells in the same column have to contain the same semantic role.

According to Table VI.1. ablative case, just as elative, encodes a physical source in the context of verbs of motion. Both cases having the same semantic role, the decision between them is determined by the semantics of the corresponding NP. These cases encode an other semantic role in the case of verbs of existence (Table VI.2).

	acc	ins	abl	ela
marad	-	com	cause	material
van	-	com	cause	material
lesz	-	com	cause	material
nincs	-	com	cause	material

Table VI.2 : The semantic roles of cases beside C-1 verb cluster

VI.2.5 Future work

There are two major directions regarding our future work. With respect to the automatic clustering process, we have the intention of widening the scope of the grammatical features to be compared by enriching subcategorization frames by other morphological properties. We are also planning to test top-down clustering methods such as the one described in (Pereira et al., 1993). On the long run, it will be inevitable to make experiments on larger corpora. The obvious choice is the 180 million words Hungarian National Corpus (Váradi, 2002). It is a POS-tagged corpus but does not contain any syntactic annotation; hence its use would require at least some partial parsing such as NP analysis to be employable for our purposes. The other future direction concerns evaluation and linguistic analysis of verb clusters. We define well-founded verb classes on the basis of semantic role matrices. These semantic roles can be filled in a sentence by casemarked NPs. Therefore, evaluation of automatically obtained clusters presupposes the definition of such matrices, which is our major linguistic task in the future. When we have the supposed matrices at our disposal, we can start evaluating the clusters via example sentences which illustrate case suffix alternations or roles characteristic to specific classes.

No corpus can be big enough to avoid the problem of data sparseness. Since it is impossible to construct an extensive corpus which covers all linguistic and lexical phenomena at a given moment in time, we cannot assume that what is not in the corpus is not in language. Besides, a common problem with statistical methods is the difficulty to model phenomena for which a limited amount of data is available. However, a lexical acquisition method is efficient only if it can be applied to less frequent words, and being able to estimate the behavior of unseen or rare data is essential for the portability of the method. When designing the feature space, we need to take into account the extent of

data sparseness to be handled. Rare data are usually treated by smoothing; therefore, choosing the adequate smoothing method is important when designing the algorithm. Since the experiments presented above were dealing with frequent verbs, smoothing was seen as only a technical problem: the absence of a subcategorization frame with a high frequency verb is most likely due to incompatibility between the verb and the frame. However, when extending our method to deal with low frequency data, we will need to integrate a predictive smoothing in the algorithm.

VI.3 Unsupervised Acquisition of French Verb Classes

VI.3.1 Clustering French Verbs - Method

In our second experiment, we tried to cluster French verbs into lexical semantic classes. We based our approach on the two hypotheses described in the previous section, namely, that semantically similar verbs tend to occur in similar syntactic contexts, and consequently, this kind of semantic information can be obtained from syntactically analysed corpora. We would like to confirm that 1) verbs can be clustered in semantic groups by comparing their distribution in corpora, 2) verbs' distribution can be modeled by subcategorization frames extracted from corpora. Therefore, the results of the experiment can also be an indicator of the quality of the subcategorization information, which in our experiment is provided by an automatic subcategorization extraction method (instead of raw corpus data). We aim to create a general algorithm applicable to new verbs if they are sufficiently represented in the corpus. We use the same unsupervised clustering algorithm as in the experiments on Hungarian. The input data is provided by the subcategorization extraction algorithm of Messiant (2010). We decided to use an unfiltered version of his lexicon for two reasons. First, filtering can be a source of errors. Second, longer subcategorization frames with adjuncts can be very informative and relevant for calculating semantic similarity. A light filtering is applied though in order to reduce the feature space: frames with a frequency lower than 5 are excluded. With this filtering, the final size of the feature space (i.e. the number of different subcategorization frames) depends on the verbs in the experiment. In the experiments to be presented, the feature space varied between 433 and 1095 frames. As in the experiment on Hungarian, verbal

representation corresponds to their distribution over the frames, expressed as a maximum likelihood estimate calculated from the LexSchem lexicon. Three different distance measures were then used to calculate similarities between individual verbs:

- Kullback-Leibler divergence:

$$D_{KL}(x\|y) = \sum_{i=1}^{n} x_i \cdot \log \frac{x_i}{y_i} \tag{VI.2}$$

- Jensen-Shannon divergence:

$$D_{JS}(x\|y) = \frac{1}{2} D_{KL}\left(x\|M\right) + \frac{1}{2} D_{KL}\left(y\|M\right) \tag{VI.3}$$

  where

$$M = \frac{1}{2}\left(x + y\right) \tag{VI.4}$$

- and the skew divergence:

$$D_{\alpha}(x\|y) = D_{KL}\left(x\|\alpha y + (1-\alpha)\, x\right) \tag{VI.5}$$

When using Kullbach-Leibler divergence, the smoothing method described above have been applied to the relative frequency data. The Jensen-Shannon divergence, as well as skew divergence, are variants of the Kullbach-Leibler divergence, they do not require smoothing. The Jensen-Shannon divergence is a symmetrical measure. When using the other two distances, which are not symmetrical, the minimum of the distance was considered for each verb pair. Skew divergence is a weighted variant of Kullbach-Leibler divergence, proposed by (Lee, 2001). Weighting is done by the free parameter α, its optimal value is close to 1 – we set it to 0.99. The stop of the clustering process depends on two parameters: the cardinality of the clusters and the maximal distance between the centroids of the two clusters to be unified. The optimal values of the parameters have been calculated during test runs.

VI.3.2 Experiments with evaluation by synonym classes

In an first experiment, we attempted to evalute the quality of our results with a gold standard which is not based on Levin's classification method. The reference classification used in this experiment is purely semantic in nature, which allows us to measure the semantic coherence of our verb clusters independently of the syntactic similarities.

The semantic verb groups in this gold standard were obtained using the Dictionnaire de synonymes de Caen (Manguin, 2004) as a basis. Since the structure of the original dictionary is organised at the level of individual words instead of groups of synonyms, it was impossible to establish a direct mapping between the two resources. Moreover, the synonym dictionary suggests for each word a set of synonyms which can replace them in one of their meanings, in a subset of their distributional contexts. This implies that the relation of synonymy does not hold unconditionally between all the pairs of words proposed as synonyms for a given entry. Hence, we decided to apply a different classification method to obtain semantic verb groups from the synonym dictionary. Our synoym classification method is based on the work presented in (Manguin, 2003). We used the verbs at the inetrsection of the synonym dictionary and the LexSchem lexicon. The similarity between words have been calculated by the Jaccard index. It is a similarity measure inspired by set theory, based on the proportion of the synonyms the words share and the union of all of their synonyms in the dictionary:

$$S_J(x\|y) = \frac{|X \bigcap Y\|}{\|X \bigcup Y\|} \tag{VI.6}$$

Verbs have been classified with the nearest neighbour clustering method. Since we aimed to produce a reliable reference classification, a limited manual inetrvention has been included: the clustering process was stopped after three iterations, and the resulting groups have been verified manually. In order to be in line with our hard clustering method, polysemy was not included in the standard at this phase of the classification: i.e., each verb belongs to one unique class in the standard.

However, certain problems related to semantic classifications, be they automatic or manual, lead us to the conclusion that it was necessary to integrate a manual validation of the reference classes. Namely, we have to accept that the notion of symonymy between two verbs depends largely on their contexts: it is difficult, if not impossible, to establish

strict criteria to determine the primary lexical meaning of a verb and thus, to decide which semantic group it belongs to, without any reference to their distributional contexts. Therefore, even if the belonging of a verb to a certain group can be justified — since groups have been extracted from a manually built resource, the grouping is always justified — we are once again confronted to the chaining effect: groups are made up of correctly associated verb pairs, but lack a central semantic coherence. We tried to eliminate this undesirable effect by manual correction, aiming to build homogeneous classes coherently organized around a central semantic concept. However, it is difficult to define the notion of synonymy and to test whether individual verbs belong to the same semantic concept without using syntactic criteria. As shown by the experiment of Véronis (2003); Héja et al. (2009), semantic concepts not supported by distributional evidence tend to be blurry. Therefore, subjectivity and intuition cannot be completely ruled out. The second problem concerns the nature of the semantic link between distributionally similar verbs. Namely, several verbs clustered together show strong semantic coherence which can be captured by a common core meaning, nevertheless, the semantic relation between them cannot be considered as synonymy. Antonyms constitute a special subtype of this category, e.g. in our first sample verb set we found verb pairs grouped together such as 'restaurer'/'fausser', or 'surgir'/'disparaître'. In the terminology of (Levin, 1993), the semantic link between verbs sharing the same group can be captured in terms of metapredicates, i.e. meaning components. Therefore, she does not state that verb classification should be limited to grouping synonyms together. When creating a reference classification, these aspects have to be taken into consideration.

We created two sample sets from the gold standard. The first one is made up of 155 verbs, belonging to 26 classes in the standard, whereas the second one contains 68 verbs and 14 classes (the smaller sample is a subset of the bigger one). This allows us to compare the effect of the sample size and the number of the classes on the clustering performance. The complexity of the task depends on the number of classes. The baseline for a classification task with n classes is 1/n, i.e. 0.038 and 0.071 for the bigger and the smaller sample, respectively. Different evaluation metrics can be used for optimizing the parameters and for evaluating the results. The metric called Adjusted Pairwise Precision (Schulte im Walde and Brew, 2002) calculates the precision of clusters in terms of verb

pairs. APP is the average proportion of all within-cluster pairs that are correctly co-assigned. It compensates for the bias towards small clusters by a weighting factor that increases with cluster size:

$$APP\,(C) = \frac{1}{|C|}\sum_{i=1}^{C}\frac{paires\_correctes\_dans\_c_i}{paires\_dans\_c_i} \times \frac{c_i - 1}{c_i + 1}$$

This measure was used for setting the parameters: chosing cluster size and maximal distance measure. The best scores were obtained by using skew divergence and limiting cluster size to five elements. Therefore, detailed quantitative and qualitative evaluation has been carried on using the results obtained with these parameters.

Distance measure	Sample	Cardinality	APP
Kullbach-Leibler	155	6	0.123
Jensen-Shannon	155	6	0.125
Skew	155	6	0.149
Skew	68	5	0.233

Table VI.3 : Clustering results with different parameters

For a quantitative evaluation, it is possible to associate the clusters produced by an unsupervised algorithm to classes of the gold standard by associating each cluster to the semantic class which is prevalent (i.e. has the most members) inside the cluster. This allows us to calculate the *modified purity* and the *weighted class accuracy* of our clusters (Sun et al., 2008).

Modified purity is a global measure which evaluates the mean precision of clusters. Each cluster is associated with its prevalent class. For each cluster, the proportion of verbs that belong to the prevalent class are calculated – verbs which no dot belong to the prevalent class, as well as singleton clusters, are considered as errors.

$$mPurity\,(C) = \frac{\sum_{n_{prevalent}(k_i)\geq 2} n_{prevalent}\,(k_i)}{|C|} \tag{VI.7}$$

Weighted class accuracy can be considered as a measure of recall: for each class in the gold standard, we calculate the proportion of verbs assigned to the prevalent cluster associated to that class. By definition, this quantity cannot be bigger than the maximal cardinality of our clusters.

$$Acc\,(C) = \frac{\sum_{i=1}^{C} verbes\_dans\_GRP.DOM_i}{|C|} \tag{VI.8}$$

Finally, we defined the F-mesure as the harmonic mean of modified purity and weighted class accuracy:

$$F = \frac{2 \times mPurity \times Acc}{mPurity + Acc} \tag{VI.9}$$

Distance measure	Sample	Cardinality	mPurity	Accuracy	F-measure
Skew	155	5	0.49	0.40	0.44
Skew	68	5	0.60	0.58	0.59

Table VI.4 : Evalution with the best parameters

In our second experiment, we used a gold standard created by T. Poibeau as a Levin-type classification for French. The resource is composed of 176 verbs, classified into 16 semantic classes. The size of the classes in the gold standard vary between 8 and 17. In order to be able to test the robustness and the extensibility of the method, we chose verbs from different frequency groups. The classes correspond to a subset of Levin's classification, the first candidates have been obtained by translating the English verbs. In order to make sure that the French classes are coherent with Levin's classification, we deleted the verbs which do not share the same basic syntactic constructions. Therefore, the classes in the gold standard are characterized by a shared semantic component as well as (at least) one syntactic construction. Finally, we validated the verb classes by comparing them to the Lexicon-Grammar and found that the verbs belonging to the same class generally belong to the same LG table, with a few exeptions. (A detailed study on the comparison of the two approaches and the resulting resources was taken out in parallel by Messiant (Messiant et al., 2009).

The same clustering method was applied to the verbs from the gold standard. We compared clustering results to the gold standard using evaluation metrics described in the previous section. The baseline for this experiment is 1/16= 0.0625.

If we optimize the parameters of the algorithm for the measure APP, we obtain the best results with the cluster cardinality limited to 4 elements. As illutrated by the following examples, the strong precision seems to confirm the hypothetical link between semantic properties and the syntactic distribution of verbs as observed in the corpus:

Distance Mesure	Cardinality	APP	mPURITY	ACC	F-measure
KL	6	0.13	0.48	0.30	0.36
KL	5	0.13	0.51	0.27	0.35
JS	4	0.21	0.60	0.28	0.39
JS	5	0.18	0.54	0.30	0.38
skew	4	0.22	0.62	0.27	0.37
skew	5	0.18	0.55	0.29	0.37
skew	9	0.16	0.47	0.35	0.40

Table VI.5 : Results evaluated on Levin-type verb classes

clusterA: errer (wander), voyager (travel), circuler (circulate,run), naviguer (navigate)

clusterB: dire (say), indiquer (indicate), affirmer (claim), déclarer (declare)

clusterC: signaler (report), révéler (show), montrer (show), annoncer (announce)

clusterD: rouspéter (groumble), ronchonner (groumble), grogner (groan), râler (groan)

However, recall values penalize more strictly the structural difference between clustering results and the gold standard, namely with respect to the cardinality of clusters. If we optimize the parameters for F-measure, bigger clusters yield better results. At the same time, weighted class accuracy shows how the inter-cluster coherence decreases while increasing the cardinality of the clusters above 4 elements. By looking at the clusters, it becomes evident that the chaining effect leads to weaker results: instead of coherent classes centered around a core semantic component, we obtain classes with a series of verb pairs with a semantic link which is lightly modified ever time a new sub-cluster is added. For example, the clusters below show a certain level of semantic coherence, but this semantic relation is different from that in the equivalent class of the gold standard: (square brackets indicate classes in the gold standard):

clusterE: [resplendir pétiller scintiller] [vibrer]

clusterF: [consterner ennuyer] [dévisager] [rosser]

clusterG: [bougonner gémir] [trembler vaciller]

clusterH: [grésiller geindre] [trembloter] [flamboyer]

clusterI: [consolider renforcer] [réintégrer] [maintenir]

Authors	Data	Nb of verbs	Nb of classes	APP
Schulte im Walde and Brew	German	57	17	0.15
Korhonen et al. 2003	English mono	110	33	0.21
Korhonen et al. 2003	English poly	110	33	0.29
our experiment1	French-syonym	68	14	0.23
our experiment2	French-Levin classes	176	17	0.22

Table VI.6 : Comparison of clustering results

VI.3.3 Comparison and Discussion

It is not surprising that quantitative evaluation results decrease when increasing the number of clusters in the reference classification (as it is reflected by the baseline of the task), as well as when increasing the number of verbs to be classified. Quantitative evaluation of such results have to be interpreted accordingly: when comparing numerical values, the baseline and the number of verbs have to be taken in consideration. A meaningful direct comparison can only be accomplished between clustering experiments on the same set of words and using the same reference. Moreover, as the table in the previous section shows, different quantitative evaluation metrics produce different rankings. Namely, APP and mPurity, although both of them are metrics of precision, do not always correlate, since they are biased by the cardinality of clusters. This bias cannot be ruled out in an unsupervised setting, as we do not have any a priori information on the number of semantic classes.

For these considerations, it is not possible to map our results directly to those obtained for other languages. However, an indicative comparison can be given with experiments evaluated with the same metrics on similar datasets. Among the earliest experiments, we need to mention the work of (Schulte im Walde and Brew, 2002), who clustered German verbs using a manually created, Levin-type gold standard. The semantic aspects of their verb groups are closely related to Levin's English classes. The dataset contains 57 verbs in 17 groups. Korhonen et al. (2003) use an extended version of Levin's classification to produce two types of evaluation datasets: one of them containing each verb classified according to its predominant sence, and the other one containing each verb in as many groups as it has senses according to the classification. This polysemous dataset is intended

Authors	Method	F-measure	Baseline
Li et al., 2008	supervised	66.3	0.066
Joanis et al., 2008	supervised	58.4	0.066
Sun and Korhonen, 2009	unsupervised	57.5	0.066
our experiment1	unsupervised	40	0.062
our experiment2	unsupervised	59	0.071

Table VI.7 : Comparison by F-measure and baseline

for the evaluation of their soft clustering method applied to polysemous verbs. Since we always use hard clustering in our experiments, the monosemous reference classification is conceptually closer to our approach, but we included both results in the table.

Recently, Joanis et al. (2008) proposed a gold standard of 205 verbs, classified into 15 Levin classes. Korhonen (2009) provides a comparative overview on the performance of recent approaches using this gold standard. (F-measures are calculated as a mean of the weighted class accuracy and the modified purity, as defined in this section). The table below shows the comparison between our results and these experiments on English data.

For a better understanding of the scores, we have to note that Li and Brew (2008) and Joanis et al. (2008) obtained more precise results since they used supervised approaches, unlike Sun and Korhonen (2009). Unsupervised clustering is more sensitive to noise and is constrained to use general feature spaces, not optimized for the given classes. Surprisingly, better evaluation scores are obtained when using the synonym-based gold standard than with the Levin-type verb classes, although Levin classes are always supported by syntactic similarities. An important consequence of this finding is that French diathesis alternations need to be reflected upon. It is likely that a reliable classification cannot be constructed by translation: even though some syntactic constructions may be shared among class members, the resulting reference classes may not reflect the classification organized around meaning components. In other words, the aim of automatic distributional clustering is to discover meaning components which govern syntactic alternations. A closer look at the clustering errors shows that the same typical errors occur in every test set we used: namely, false chaining of synonym pairs and lack of central coherence therefore; semantic links other than synonymy (which in some cases correspond to an error in the reference classification), and grouping of verbs with an accidental syntactic similarity but

no semantic link. Direct manual validation of classes is a complex linguistic task, leading us to create an inventory of syntactically relevant meaning components with the method presented in section VI/2 for Hungarian verb classes.

VI.3.4 Future work

It is important to note that the level of accuracy reported above was obtained by using a fully automatized text processing tool chain from the annotation of to corpus to the extraction of relevant syntactic features that compose the feature space. Moreover, our feature space is conceived to be completely general and thus, it does not build on any a priori knowledge about the semantic verb classes. The results correspond to the state of the art in verb clustering, although the performance of current systems does not allow results to be used directly in NLP applications. However, these experiments confirm the Semantic Base Hypothesis and show that automatic classification is a promising line of research.

Several improvements can be foreseen at the level of clustering methodology. Switching to a supervised classification method would allow to specify a set of meaning components and to adapt the feature space to the specific classification task (by weighting or filtering of features). The most class-specific subcategorization frames would provide valuable indication on the nature of diathesis alternations in French. Extending the representation of verbs to other features (namely, semantic selectional restrictions) would yield a different classification. However, it is not evident how to incorporate semantic information without having to face date sparseness. Another possible improvement would be to switch to a soft clustering method. On the long run, this is even necessary since many verbs are truly polysemous and thus, can belong to more than one classes. This representation also reflects the idea that meaning components form a network, and verbs are disambiguated among them by the choice of syntactic construction they take in each given context.

Chapter VII

Conclusion and Future Work

VII.1 Applications Using the Resources Presented in this Thesis

Certain instances of the work presented in this thesis were accomplished by the author as a part of a research and development project in partnership between the RIL HAS and external academic and (or) industrial institutions. The resources developed by the author of the thesis have been utilized in these projects, as well as in subsequent projects carried out independently by other research groups. The creation of the verbal argument structure database was first initiated within the European project MATCHPAD (Machine Translation for the Use of Czech, Polish and Hungarian Public Administration) lead by Systran SA. The first phase of the coding took place during this project. The second phase of coding was accomplished in parallel to the development of Szeged Treebank (Csendes et al., 2004). Subsequently, the database was integrated into the information extraction system developed in the project Information Extraction from Short Business News (Prószéky, 2003). The information extraction system is based on a pattern matching process centered around lexicalized verbal complement structures. A set of abstract semantic schemata were manually defined or extracted from domain-specific resources and put in correspondence with verbal complement frames. The lexical database was extended to cover the vocabulary of a 150.000 words corpus of short business news. During pattern matching, these complement frames assign domain-specific semantic roles to sequences in input sentences. Between 2004 and 2007, the verbal argument structure

database was used in a project which aimed at developing a Hungarian-English machine translation system (Metamorpho: http://webforditas.hu, http://itranslate4.eu). The MT system (Prószéky and Tihanyi, 2002), combines the advantages of example-based (EBMT) and rule-based machine translation (RBMT), to create a new paradigm, pattern-based translation. Within this paradigm, productive (general) syntactic rules and lexically specified patterns are encoded in the same formalism: there is a continuum between typically productive phenomena and patterns with different degrees of specification. The lexicon was integrated to the source language analysis and the bilingual module (Tihanyi, 2006). In order to do so, each entry in the lexicon was translated to English. The database has a twofold function within the MT system: Hungarian subcategorization descriptions support syntactic analysis, while their corresponding English equivalents are used to produce target language translations. By the end of the project, the number of different verbal argument structure patterns in the MT system exceeded 30.000.

The development of the Hungarian shallow parser is part of a bigger project to create Hungarian resources in NooJ for the use of the linguist/NLP community as well as for other researchers in the humanities area. The shallow parser was used between 2005 and 2007 in the project Realization of Near-natural Human-Machine Speech Interaction in Information Systems (Tamm et al., 2008) lead by the Department of Phonetics of RIL HAS. The project aimed at providing a tool for the automatic annotation of the main prosodic elements in Hungarian sentences, based on syntactic analysis. The major linguistic challenge of the project was to identify focus, topic, contrastive topic and other syntactic operations having a crucial effect on prosody in Hungarian. The identification of these structures was based on the input provided by the shallow parser. At a later phase of the project, the relevant vocabulary of the verbal complement structure database was also added to the syntactic module.

The Hungarian NooJ module is extensively used in narrative psychological studies by the research groups in Narrative Psychology at the Institute of Cognitive Neuroscience and Psychology of the MTA Research Centre for Natural Sciences and at the University of Pécs (http://narrativpszichologia.pte.hu/). Launched by the initiative of János László, these studies aim to show that automatized psychological content analysis is more efficient

when it relies on a deeper linguistic analysis, as opposed to the more widespread, purely statistical methods of content analysis.

Several projects were carried out within the thematics of psychological content analysis, covering the fields of life stories and personality, study of national and ethnic identity in historical narratives, emotion detection, and the study of the correlation between emotional structure and narrative structure. The research methodology is based on an automatized content analysis. The Hungarian NooJ module was used to perform a linguistic analysis on corpora (e.g. books on Hungarian history or transcriptions of life stories narrated by individuals) as a first step. Subsequently, a set of specific grammars have been created in NooJ by the research team on narrative psychology (Liu and László, 2007; Vincze et al., 2010; Ehmann and Garami, 2010) for the automatized processing of different aspects of narrative structure, e.g. agentivity, intentionality, emotional states, mental predicates. The first step towards integrating external resources relevant for narrative psychology were taken in 2009 by the elaboration of a new type of grammar, capable to process external XML annotation enhanced with thematic roles and other psychologicallly relevant annotations (Vincze et al., 2010). Following this line of work, Ehmann (2010) successfully integrated a series of external resources in her text processing module for psychological status monitoring of crews in isolated, confined and extreme settings. (Pólya and Gábor, 2010) created a corpus of personal narratives annotated at three distinc levels: emotional intensity (as measured by a PROCOMP5 biofeedback system), narrative structure (annotated semi-automatically) and linguistic structure (annotated automatically using the Hungarian NooJ module). Quantifiable correlations between these structures were studied in the corpus: the experiments proved that there is an inverse correlation between emotional intensity and the elaboration of narrative structure (Pólya and Gábor, 2010).

VII.2 Conclusion

We presented the creation of two resources for Hungarian NLP applications: a rule-based NP chunker and shallow parser, and a database of verbal subcategorization frames. Hungarian, as a non-configurational language with a rich morphology, presents specific

challenges for NLP at the level of morphological and syntactic processing. While efficient and precise morphological analyzers are already available, Hungarian was less resourced with respect to syntactic processing. Our work aimed at overcoming this problem by providing resources for syntactic processing.

In chapter II., we presented a shallow parser developed manually as a set of cascaded grammars in NooJ. We presented how the specific features of NooJ as a grammar development environment and a corpus processing tool can be exploited to model language phenomena such as agreement or self-recursion inside phrases with a bound constituent order. The NP chunker function of the shallow parser was evaluated on an extract from a manually annotated Hungarian treebank. We also described the construction of an automatically annotated 10 million words corpus using this shallow parser.

We have seen that while constituent chunking can be done efficiently for Hungarian by a set of grammars operating on POS categories as input, a correct processing of verbal dependencies has to rely on a lexical database of verbal complementation frames. In chapter III, we argued that current definitions of the complement-adjunct dichotomy present contradictions due to the fact that semantic definitions and the corresponding syntactic tests do not designate the same set of constituents as complements. We also show by a counter-example that adjuncts, similarly to complements, can provoke event type shift and hence change the syntactic distribution of the verb phrase they are attached to. We further propose a methodology for adapting complement/argument definitions and tests to a non-configurational language where complement tests based on surface syntax systematically fail. This methodology is based on the observation that contrary to a widely shared presumption, adjuncts are often not fully productive. We therefore propose a gradual notion of productivity, defined in relation to lexical semantic verb classes. After presenting the coding guidelines and the structure of our verbal subcategorization database (IV), we present and discuss two experiments on enhancing the database with relevant lexical semantic informations (V and VI). The goal of these experiments is to adapt state of the art results in argument realization research to Hungarian and to elaborate a compact lexical representation for verbs and their subcategorization frames, as well

as for predicting the scope and the semantic role of adjuncts a verb can take. We first present a method to manually define lexical semantic verb classes, their syntactic specificities and their semantic role assignment with respect to adjuncts. A case study, the categorization of adjuncts and adjunct-taking predicates with the instrumental case suffix, was described in detail. In parallel to the definition of verb classes, the categorization of adjunct with the suffix -val was implemented as a semantic role labeling system. The evaluation and error analysis let us conclude that the semantic role set and the syntactic tests used for verb classification are valid and usable for the annotation task. However, due to the demanding and time-consuming nature of the manual work required for this task, we moved towards experimenting with machine learning methods. We applied an unsupervised clustering method to syntactic data extracted from the Szeged Treebank to obtain clusters of semantically related and distributionally similar verbs. The method was then applied to French, using an automatically created subcategorization lexicon. The experiments confirmed that distributional similarities can be exploited to induce semantically coherent verb classes both in Hungarian and in French. We demonstrated that a feature space with purely syntactic information and including complements as well as adjuncts can be used to model efficiently the distributional properties of verbs. The verb classes we obtain can be used to support linguistic analysis of the relevant lexical semantic properties of verbs, as well as their syntactic implications specific to verb classes.

VII.3 Future Work

VII.3.1 Improving the acquisition of lexical semantic features

We have seen that relevant information can be extracted from corpora regarding lexical semantic verb classes. The manual evaluation and error analysis of these classes can help to move towards a more complete description of verb classes and corresponding syntactic properties in Hungarian. We plan to work in parallel on the amelioration of the clustering technique and on the integration of the results in the lexical database. With respect to the clustering technique, an important direction is to define a more precise and less noisy feature space. Our current algorithm works with a large number of features, due to the fact that every top-level sentence constituent is considered to be in the complementation

frame of the verb. In order to reduce the feature space in a meaningful way, we can envisage several possibilities:

- Using an automated method for feature space reduction, e.g. the one proposed by de Cruys (2010) based on non-negative tensor factorization.

- Applying an automatic alternation detection method prior to verb clustering (McCarthy, 2001). It is likely that our notion of alternations has to be adapted to Hungarian, and therefore, the alternation detection method will rely on different syntactic presumptions, e.g. we will need to take morphological information into account when adapting the method to Hungarian. The advantage of using alternation detection before verb clustering is to select the most important (i.e. alternating) features and be able to give them a higher weight when calculating the distance/similarity measure, even in an unsupervised setting.

- Applying the algorithm put forward by Sass (2009) for subcategorization acquisition. His method has the advantage to dynamically calculate equivalences between basic subcategorization frames and their longer variants and proposing a relevance measure for a subcategorization frame and a predicate based on the updated co-occurrence information. The equivalences as output by his algorithm could be used for feature space reduction.

Before integrating the results from automated lexical acquisition into the verb lexicon, the syntactic relevance of verb classes needs to be stated in terms of class-specific alternations. A lexical semantic feature is conceived as a generalization over a set of syntactic properties and a set of implications about semantic role assignment. These generalizations make it possible to reduce the number of lexicon entries for verbs belonging to a certain class, as the lexical semantic features automatically associate the corresponding alternating entries to every member of the class. Moreover, a number of predictable elements will be moved from the database as their class-specific prediction will make it possible to process them as adjuncts. The explicit formulations of such implications will have to be based on a linguistic study similar to the procedure defined in V.

VII.3.2 Recognizing complement structures using the verb lexicon

The verbal subcategorization database together with the shallow parser makes it possible to move towards deep parsing and annotate verbal complement structure in Hungarian sentences. We plan to provide a solution to this task using NooJ, and to make it available within the Hungarian module of the tool. The major issue to resolve is to find an algorithm to efficiently match subcategorization patterns to input text, while

- the constituent order is almost completely free in Hungarian sentences,
- any number of optional adjuncts or sentence adverbs can be inserted between elements in the argument structure.

The grammars constructed for integrating external argument structure information into NooJ's annotation structure for psychological processing (see VII.1) can be considered as a model for matching subcategorization patterns. These grammars rely on the use of variables: they are composed of a recursive loop which runs across the sentence to collect complement candidates, and a set of lexical constraints applied at the end to check lexical properties of the candidates and verify whether they correspond to the properties expected by our grammar. The next step would be to check not only the properties of the individual constituents, but also the compatibility between these constituents, to verify whether they form a correct subcategorization frame together. The development of such grammars and their application to large corpora will provide a basis for a subsequent semantic processing, e.g.by exploiting our lexical semantic verb classes and the associated semantic role assignment properties.

List of Tables

List of Figures

Bibliography

Abeillée, A. and Candito, M.-H. (2000). FTAG: A Lexicalized Tree Adjoining Grammar for French. In *Tree Adjoining Grammars: Formal, Computational and Linguistic Aspects*, pages 305–329. CSLI publications.

Abney, S. (1996). Partial Parsing via Finite–State Cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, Prague, Czech Republic.

Alberti, G. (2006). *Matematika a természetes nyelvek leírásában [Mathematics in the Description of Natural Languages]*. Tinta Kiadó, Budapest, Hungary.

Alishahi, A. and Stevenson, S. (2007). A Cognitive Model for the Representation and Acquisition of Verb Selectional Preferences. In *ACL Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 41–48, Prague, Czech Republic.

Babarczy, A., Gábor, B., Hamp, G., and Rung, A. (2005). Hunpars: a Rule-based Sentence Parser for Hungarian. In *Proceedings of the 6th International Symposium on Computational Intelligence*.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the COLING-ACL Conferences*, pages 86–90.

Baker, M. C. (1988). *Incorporation: A Theory of Grammatical Function Changing*. The University of Chicago Press, Chicago.

Beesley, K. R. and Karttunen, L. (2000). Finite-State Non-Concatenative Morphotactics. In *38th Annual Meeting of the Association for Computational Linguistics*.

Booth, T. L. (1969). Probabilistic Representation of Formal Languages. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:74–81.

Borsley, R. D. (1989). An HPSG approach to Welsh. *Journal of Linguistics*, 25:333–354.

Brent, M. R. (1991). Automatic Acquisition of Subcategorization Frames from Untagged Text. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 209–214, Berkeley, CA.

Brent, M. R. (1993). From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19:203–222.

Bresnan, J. (1978). A Realistic Transformational Grammar. In Halle, M., Bresnan, J., and Miller, G. A., editors, *Linguistic theory and psychological reality*, pages 1–59. MIT Press, Cambridge.

Bresnan, J. (1982). *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, MA.

Bresnan, J. and Zaenen, A. (1990). Deep Unaccusativity in LFG. In Dziwirek, K., editor, *Grammatical Relations. A Cross-Theoretical Perspective*. Center for the Study of Language and Information, Stanford University.

Briscoe, T. and Carroll, J. (1997). Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC.

Briscoe, T., Copestake, A., and Boguraev, B. (1990). Enjoy the Paper: Lexicology. In *Proceedings of the COLING Conference*, pages 42–47.

Butt, M. (2006). *Theories of Case*. Cambridge University Press, Cambridge, UK.

Carreras, X. and Marquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling.

Chomsky, N. (1956). Three Models for the Description of Language. *IRE Transactions on Information Theory*, 2(3):113–124.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris, Dordrecht.

Copestake, A. (1992). *The Representation of Lexical Semantic Information*. PhD thesis, University of Sussex.

Croft, W. (1998). Event Structure in Argument Linking. In *The Projection of Arguments: Lexical and Compositional Factors*. Stanford: Center for the Study of Language and Information.

Csendes, D., Csirik, J., and Gyimóthy, T. (2004). The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora at COLING 2004*, pages 19 –23, Geneva, Switzerland.

Cunningham, H. (2002). GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36(2):223–254.

de Cruys, T. V. (2010). A Non-negative Tensor Factorization Model for Selectional Preference Induction. *Natural Language Engineering*, 16(4):417–437.

de Cruys, T. V. and Apidianaki, M. (2011). Latent Semantic Word Sense Induction and Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1476–1485, Portland OR, US.

Doran, C., Egedi, D., Hockey, A. B., Srinivas, B., and Zaidel, M. (1994). XTAG System: A Wide Coverage Grammar for English. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67(3).

Earley, J. (1983). An Efficient Context-free Parsing Algorithm. *Commun. ACM*, 26(1):57–61.

Ehmann, B. and Garami, V. (2010). Narrative Psychological Content Analysis with NooJ: Linguistic Markers of Time Experience in Self-Reports. In *Selected Papers from the 2008 International NooJ Conference*, pages 186–196. Cambridge Scholars Publishing.

É. Kiss, K. (2002). *The Syntax of Hungarian*. Cambridge University Press, The Hague.

Elekfi, L. (1997). *Magyar ragozási szótár. [Dictionary of Hungarian Inflections]*. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest.

Erk, K., Kowalski, A., Padó, S., and Pinkal, M. (2003). Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of the ACL 2003 Conference*.

Evans, E. G. (1997). Approximating Context–Free Grammars with a Finite–State Calculus. In *35th Annual Meeting of the Association for Computational Linguistics*, pages 452–459, Madrid.

Fabre, C. and Bourigault, D. (2006). Extraction de relations sémantiques entre noms et verbes au-delà des liens morphologiques. In *Proceedings of the 13th TALN Conference*, pages 121–129, Leuven, Belgique.

Fabre, C. and Bourigault, D. (2008). Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18(1).

Falk, I. (2008). Création automatique de classes sémantiques verbales pour le français. Master's thesis, Ecole Doctorale IAEM Lorraine, Nancy.

Fillmore, C. J. (1968). The Case for Case. In *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston.

Fillmore, C. J. (1982). *Frame Semantics*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Gábor, K. (2007). Syntactic Parsing and Named Entity Recognition for Hungarian with Intex. In *Formaliser les langues avec l'ordinateur: De Intex á Nooj*. Presses Universitaires de Franche-Comté.

Gábor, K. and Héja, E. (2005). A Rule-based Analysis of Complements and Adjuncts. In *Proceedings of the Third International Seminar on the Computer Treatment of Slavic and East European Languages*.

Gábor, K. and Héja, E. (2006). Predikátumok és szabad határozók. [Predicates and Adjuncts]. In Kálmán, L., editor, *KB 120 A titkos kötet. [KB 120: The. Secret Volume]*, pages 135–152. Research Institute for Linguistics, HAS.

Gábor, K. and Héja, E. (2007). Clustering Hungarian Verbs on the Basis of Complementation Patterns. In *Proceedings of the ACL Conference Student Research Workshop*.

Gábor, K., Héja, E., and Mészaros, A. (2003). Kötõszók korpusz-alapú vizsgálata. In *Proceedings of the Hungarian Computational Linguistics Conference 2003*, pages 305–306, Szeged, Hungary.

Gardent, C., Guillaume, B., Perrier, G., and Falk, I. (2006). Extraction d'information de sous-catégorisation à partir des tables du LADL. In *Actes de la conférence Traitement Automatique des Langues Naturelles*, Louvain, Belgique.

Gardent, C. and Parmentier, Y. (2007). Semtag: a Platform for Specifying Tree Adjoining Grammars and Performing TAG-based Semantic Construction. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar*. Basil Blackwell.

Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. The University of Chicago Press, Chicago.

Grimshaw, J. (1990). *Argument Structure*. MIT Press, Cambridge, MA.

Gross, M. (1975). *Méthodes en syntaxe*. Hermann, Paris.

Gross, M. (1991). Linguistic Representations and Text Analysis. In *Linguistic Unity and Linguistic Diversity in Europe*, pages 31–61. Academia Europaea, London.

Gross, M. (1997). The Construction of Local Grammars. In Roche, E. and Schabes, Y., editors, *Finite–State Language Processing*, pages 329–354. MIT Press, Cambridge.

Gross, M. and Danlos, L. (1988). Building Electronic Dictionaries for Natural Language Processing. In *Programming of Future Generation Computers*, Amsterdam. North Holland, Elsevier Science Publishers.

Hajič, J., Hladká, B., and Pajas, P. (2001). The Prague Dependency Treebank: Annotation Structure and Support. In *Proceeding of the IRCS Workshop on Linguistic Databases*, pages 105–114, Philadelphia, USA.

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., and Trón, V. (2004). Creating Open Language Resources for Hungarian. In *Proceedings of the LREC 2004 Conference*.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 539–545, Nantes, France.

Héja, E., Kuti, J., and Sass, B. (2009). Jelentésegyértelmũsítés - egyértelmũ jelentés? [Word Sense Disambiguation - Ambiguous Sensation?]. In *Proceedings of the Hungarian Computational Linguistics Conference*, Szeged.

Hócza, A. (2006). Learning Tree Patterns for Syntactic Parsing. *Acta Cybernetica*, 17(3).

Jackendoff, R. (1990). *Semantic Structures*. The MIT Press, Cambridge, Massachusetts.

Joanis, E. and Stevenson, S. (2003). A General Feature Space for Automatic Verb Classification. In *Proceedings of the EACL Conference*, pages 163–170.

Joanis, E., Stevenson, S., and James, D. (2008). A General Feature Space for Automatic Verb Classification. *Natural Language Engineering*, 14(3):337–367.

Joshi, A., Vijay-Shanker, K., and Weir, D. J. (1991). The Convergence of Mildly Context–Sensitive Grammatical Formalisms. In Sells, P., Shieber, S., and Wasow, T., editors, *Foundational Issues in Natural Language Processing*, pages 31–81. MIT Press, Cambridge.

Joshi, A. K. and Vijay-Shanker, K. (1985). Some Computational Properties of Tree Adjoining Grammars. In *23rd Annual Meeting of the Association for Computational Linguistics*, pages 82–93, Chicago, Illinois, USA.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (International Edition)*. Prentice Hall.

Kálmán, L. (2006). Miért nem vonzanak a régensek? [Why don't Governors Attract?]. In Kálmán, L., editor, *KB 120 A titkos kötet. [KB 120: The. Secret Volume]*, pages 229–246. Research Institute for Linguistics, HAS.

Kálmán C., G., Kálmán, L., Nádasdy, A., and Prószéky, G. (1989). A magyar segédigék rendszere [The System of Hungarian Auxiliary Verbs]. *Általános Nyelvészeti Tanulmányok*, XVII:49–103.

Kaplan, R. and Bresnan, J. (1982). Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press, Cambridge, MA.

Karlsson, F. (2007). Constraints on Multiple Center–embedding of Clauses. *Journal of Linguistics*, 43(02):365.

Karttunen, L., Gaál, T., and Kempe, A. (1997). Xerox Finite-State Tool. Technical Report MSU-CSE-00-2, Xerox Research Centre Europe, Meylan, France.

Kay, P. and Fillmore, C. (1999). Grammatical Constructions and Linguistic Generalizations: the What's X doing Y? Construction. *Language*, 75(1):1–33.

Kiefer, F. (1992). Az aspektus és a mondat szerkezete. [Aspect and the Structure of the Clause]. In Kiefer, F., editor, *Strukturális magyar nyelvtan. I. Mondattan*, pages 797–886. Akadémiai Kiadó, Budapest.

Kipper, K., Dang, H. T., and Palmer, M. S. (2000). Class-Based Construction of a Verb Lexicon. In *Proceedings of the AAAI/IAAI Conference*, pages 691–696.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A Large-Scale Classification of English Verbs. *Journal of Language Resources and Evaluation*, 42(1):21–40.

Kipper-Schuler, K. (2005). *Verbnet: a Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA. AAI3179808.

Koeling, R. (2000). Chunking with Maximum Etropy Models. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7*, ConLL '00, pages 139–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koenig, J.-P. and Davis, A. (2003). Semantically Transparent Linking in HPSG. In *Proceedings of the HPSG03 Conference*, pages 222–235.

Koenig, J.-P. and Davis, A. (2006). The KEY to Lexical Semantic Representations. *Journal of Linguistics*, 42:71–108.

Komlósy, A. (1992). Régensek és vonzatok [Governors and Complements]. In Kiefer, F., editor, *Strukturális magyar nyelvtan. I. Mondattan*, pages 299–527. Akadémiai Kiadó, Budapest.

Korhonen, A. (2002). *Subcategorization Acquisition*. PhD thesis, University of Cambridge.

Korhonen, A. (2009). Automatic Lexical Classification - Balancing between Machine Learning and Linguistics. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*.

Korhonen, A. and Briscoe, T. (2004). Extended Lexical-Semantic Classification of English Verbs. In Moldovan, D. and Girju, R., editors, *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 38–45, Boston, Massachusetts, USA. Association for Computational Linguistics.

Korhonen, A., Gorrell, G., and McCarthy, D. (2000). Statistical Filtering and Subcategorization Frame Acquisition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

Korhonen, A., Krymolowski, Y., and Marx, Z. (2003). Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71.

Kornai, A. (2008). *Mathematical Linguistics*. Advanced Information and Knowledge Processing. Springer.

Koskenniemi, K. (1984). A General Computational Model for Word-Form Recognition and Production. In *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 178–181, Morristown, NJ, USA.

Koskenniemi, K. (1990). Finite-State Parsing and Disambiguation. In *Proceedings of the 13th conference on Computational linguistics - Volume 2*, COLING '90, pages 229–232, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koskenniemi, K. and Yli-Jyrä, A. (2009). CLARIN and Free Open Source Finite-State Tools. In *Proceedings of the 2009 conference on Finite-State Methods and Natural Language Processing: Post-proceedings of the 7th International Workshop FSMNLP 2008*, pages 3–13, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Kudo, T. and Matsumoto, Y. (2001). Chunking with Support Vector Machines. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL)*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Kupsc, A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. In *Actes de la confèrence Traitement Automatique des Langues Naturelles (TALN)*, Toulouse.

Kuti, J., Varasdi, K., Gyarmati, Á., and Vajda, P. (2007). Hungarian WordNet and Representation of Verbal Event Structure. *Acta Cybernetica*, 18(2):315–328.

Lafferty, J., Mccallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Lapata, M. (1999). Acquiring Lexical Generalizations from Corpora: A Case Study for

Diathesis Alternations. In *Proceedings of the 37st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 397–404, Maryland, USA.

Laporte, E. (2000). Mots et niveau lexical. In Pierrel, J.-M., editor, *Ingénierie des langues*, pages 25–49. Hermès.

Levin, B. (1993). *English Verb Classes and Alternations: a Preliminary Investigation*. University of Chicago Press, Chicago and London.

Levin, B. and Hovav, M. R. (2005). *Argument Realization*. Research Surveys in Linguistics. Cambridge University Press, Cambridge, UK.

Li, J. and Brew, C. (2008). Which are the Best Features for Automatic Verb Classification? In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL08-HLT)*, pages 434–442, Columbus, Ohio.

Lightfoot, D. (1979). *Principles of Diachronic Syntax*. Cambridge University Press, Cambridge, UK.

Liu, J. and László, J. (2007). A Narrative Theory of History and Identity : Social Identity, Social Representations, Society and the Individual. In Moloney, G. and Walker, I., editors, *Social Representations and History*, pages 85–107. Palgrave-Macmillan.

Lopatková, M. (2003). Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics*, pages 37–59.

Loper, E. and Bird, S. (2002). NLTK: the Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, pages 63–70, Morristown, NJ, USA. Association for Computational Linguistics.

Manguin, J.-L. (2003). Utilisation d'un corpus catégorisé pour l'étude et la représentation de la synonymie en contexte. In *Actes de 3emes journées de linguistique de corpus*.

Manguin, J.-L. (2004). Le dictionnaire électronique de synonymes de CRISCO. Un mode d'emploi à trois niveaux.

Manning, C. and Sag, I. (1999). Dissociations between Argument Structure and Grammatical Relations. In Kathol, A., Koenig, J.-P., and Webelhuth, G., editors, *Lexical and Constructional Aspects of Linguistic Explanation*. Stanford: CSLI Publications.

Manning, C. D. (1993). Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Meeting of the Association for Computational Linguistics*, pages 235–242.

Manning, C. D. and Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition.

Marquez, L. (2009). Tutorial on Semantic Role Labeling at ACL-IJCNLP 2009.

Mason, O. (2004). Automatic Processing of Local Grammar Patterns. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, pages 166 –171, University of Birmingham, UK.

McCallum, A., Freitag, D., and Pereira, F. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598, Stanford University, Stanford, CA, USA. Morgan Kaufmann.

McCarthy, D. (2001). *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations*. PhD Thesis, University of Sussex.

Merlo, P. and Esteve Ferrer, E. (2006). The Notion of Argument in PP Attachment. *Computational Linguistics*, 32(2).

Merlo, P. and Stevenson, S. (2001). Automatic Verb Classification based on Statistical Distribution of Argument Structure. *Computational Linguistics*, 27:3:373–408.

Merlo, P. and van der Plas, L. (2009). Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics Conference*, pages 288–296.

Messiant, C. (2010). *Acquisition automatique d'informations lexicales à partir de textes*. PhD thesis, Université de Paris-Nord.

Messiant, C., Nakamura, T., and Voyatzi, S. (2009). La complémentarité des approches manuelle et automatique en acquisition lexicale. In *Actes de la 16e Conférence sur le traitement automatique des langues naturelles (TALN)*, Senlis.

Molina, A. and Pla, F. (2002). Shallow Parsing Using Specialized HMMs. *Journal of Machine Learning Research*, 2:595–613.

Moreau, E. and Tellier, I. (2009). The Crotal SRL System: a Generic Tool Based on Tree–Structured CRF. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 91–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nagy, V., Vajda, P., and Oravecz, C. (2007). Bootstrapping Nooj Morphology by an External Morphological Analyzer for Heavily Inflected Languages. In *Proceedings of the 2007 NooJ Conference*, Barcelona, Spain.

O'Donovan, R., Burke, M., Cahill, A., van Genabith, J., and Way, A. (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-ii and Penn-iii Treebanks. *Computational Linguistics*, 31(3):329–366.

Oravecz, C. and Dienes, P. (2002). Efficient Stochastic Part-of-Speech Tagging for Hungarian. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 710–717, Las Palmas, Spain.

Oravecz, C., Nagy, V., and Varasdi, K. (2005). Morphological Idiosyncrasy in Hungarian Multiword Expressions. In *Proceedings of the Third International Seminar on the Computer Treatment of Slavic and East European Languages*.

Padó, S. and Lapata, M. (2009). Cross-lingual Annotation Projection for Semantic Roles. *Journal of Artificial Intelligence Research*, 36:307–340.

Pantel, P. and Lin, D. (2002). Discovering Word Senses from Text. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Alberta, Canada.

Partee, B., ter Meulen, A., and Wall., R. E. (1990). *Mathematical Methods in Linguistics*. Kluwer Academic Publishers.

Pereira, F. C. N. and Wright, R. N. (1997). Finite–State Approximation of Phrase Structure Grammars. In Roche, E. and Schabes, Y., editors, *Finite–State Language Processing*, pages 149–173. MIT Press, Cambridge.

Phan, X.-H. (2006). CRFChunker: CRF English Phrase Chunker. http://crfchunker.sourceforge.net/.

Pollard, C. (1984). *Generalized Context-Free Grammars, Head Grammars, and Natural Language*. PhD thesis, Stanford University.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

Pólya, T. and Gábor, K. (2010). Linguistic Structure, Narrative Structure and Emotional Intensity. In *Proceedings of the Workshop on Corpora for Research on Emotion and Affect at LREC 2010*, Valetta, Malta.

Preiss, J., Briscoe, T., and Korhonen, A. (2007). A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In *Meeting of the Association for Computational Linguistics*, pages 912–918, Prague.

Prószéky, G. (2003). NewsPro: automatikus információszerzés gazdasági rövidhírekbõl [NewsPro: Information Extraction from Business News]. In *Proceedings of the Hungarian Computational Linguistics Conference*, pages 161–166, Szeged.

Prószéky, G. and Tihanyi, L. (1992). A Fast Morphological Analyzer for Lemmatizing Agglutinative Languages. In *Papers in Computational Lexicography (COMPLEX)*, pages 265–278, Budapest, Hungary. Linguistics Institute of the HAS.

Prószéky, G. and Tihanyi, L. (2002). MetaMorpho: A Pattern-Based Machine Translation System. In *Proceedings of the 24th Translating and the Computer Conference*, pages 19–24, London, United Kingdom.

Pullum, G. K. and Gazdar, G. (1982). Natural Languages and Context–Free Languages. *Linguistics and Philosophy*, 4(4):471–504.

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press, Cambridge.

Pusztai, F. (2003). *Magyar Értelmező Kéziszótár [Concise Dictionary of Hungarian]*. Akadémiai Kiadó, Budapest.

Rákosi, G. (2006). On the Need for a more Refined Approach to the Argument-Adjunct Distinction: the Case of Dative Experiencers in Hungarian. In Butt, M. and King, T. H., editors, *The proceedings of the LFG06 Conference*, pages 416–436. Stanford: CSLI Publications.

Ramshaw, L. and Marcus, M. (1995). Text Chunking Using Transformation-Based Learning. In Yarovsky, D. and Church, K., editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey. Association for Computational Linguistics.

Recski, G., Rung, A., Zséder, A., and Kornai, A. (2010). NP-Alignment in Bilingual Corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Recski, G. and Varga, D. (2009). NP–Chunking in Hungarian. *The Odd Yearbook. SEAS Working Papers in Linguistics*.

Resnik, P. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

Roche, E. and Schabes, Y., editors (1997). *Finite–State Language Processing*. Bradford Book. MIT Press, Cambridge, Massachusetts, USA.

Sabine Schulte im Walde (2009). *The Induction of Verb Frames and Verb Classes from Corpora*. Mouton de Gruyter.

Sanfilippo, A. and Poznanski, V. (1992). The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources. In *Proceedings of the third conference on Applied natural language processing*, pages 80–87.

Sass, B. (2007). First Attempt to Automatically Generate Hungarian Semantic Verb Classes. In *Proceedings of the 4th Corpus Linguistics Conference, Birmingham*.

Sass, B. (2009). A Unified Method for Extracting Simple and Multiword Verbs with Valence Information. In *Proceedings of RANLP 2009*, pages 399–403.

Saul, L. K. and Pereira, F. (1997). Aggregate and Mixed-order Markov Models for Statistical Language Processing. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 81–89.

Schabes, Y. and Waters, R. C. (1993). Lexicalized Context-Free Grammars. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 121–129.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49.

Schmid, H. (2005). A Programming Language for Finite-State Transducers. In *5th International Workshop on Finite-State Methods and Natural Language Processing, FSMNLP*, pages 308–309, Helsinki, Finland.

Schulte im Walde, S. (2000). Clustering Verbs Semantically According to their Alternation Behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 747–753, Saarbrücken, Germany.

Schulte im Walde, S. and Brew, C. (2002). Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA.

Sha, F. and Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shieber, S. M. (1985). Evidence Against the Context-Freeness of Natural Language. *Linguistics and Philosophy*, 8:333–343.

Silberztein, M. (1987). The Lexical Analysis of French. In *Electronic Dictionaries and Automata in Computational Linguistics*, pages 93–109.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Masson, Paris.

Silberztein, M. (1997). INTEX: An Integrated FST Toolbox. In *Workshop on Implementing Automata*, Lecture Notes in Computer Science, pages 185–197. Springer.

Silberztein, M. (2003). *NooJ Manual. http://www.nooj4nlp.net*.

Silberztein, M. (2004). NooJ: an Object-Oriented Approach. In Muller, C., Royauté, J., and Silberztein, M., editors, *INTEX pour la Linguistique et le Traitement Automatique des Langues*, pages 359–369. Presses Universitaires de Franche–Comté, Besançon.

Silberztein, M. (2005). NooJ: a Linguistic Annotation System for Corpus Processing. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, British Columbia, Canada. The Association for Computational Linguistics.

Silberztein, M. (2007). An Alternative Approach to Tagging. In *12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007*, Lecture Notes in Computer Science, pages 1–11. Springer.

Simon, E., Serény, A., and Babarczy, A. (2010). Automatic Acquisition of Hungarian Subcategorization Frames. In *Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods. LREC 2010 Workshop*.

Simov, K. I., Simov, A., Kouylekov, M., Ivanova, K., Grigorov, I., and Ganev, H. (2003). Development of Corpora within the CLaRK System: The BulTreeBank Project Experience. In *Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 243–246.

Somers, H. (1984). On the Validity of the Complement–Adjunct Distinction in Valency Grammar. *Linguistics*, 22:507–530.

Steedman, M. (1996). *Surface Structure and Interpretation*. MIT Press.

Stefanowitsch, A. (2006). Negative Evidence and the Raw Frequency Fallacy. *Corpus Linguistics and Linguistic Theory*, 2(1):61–77.

Stevenson, S. and Joanis, E. (2003). Semi-Supervised Verb Class Discovery Using Noisy Features. In *Proceedings of the CONLL-03 Conference*.

Sun, L. and Korhonen, A. (2009). Improving Verb Clustering with Automatically Acquired Selectional Preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sun, L., Korhonen, A., and Krymolowski, Y. (2008). Verb Class Discovery from Rich Syntactic Data. In *Computational Linguistics and Intelligent Text Processing, CICLing*, pages 16–27.

Surányi, B. (2006). Scrambling in Hungarian. *Acta Linguistica Hungarica*, 53(4):393–432.

Swier, R. S. and Stevenson, S. (2004). Unsupervised Semantic Role Labelling. In *Proceedings of the EMNLP Conference*, pages 95–102.

Tamm, A., Abari, K., and Olaszy, G. (2008). Accent Assignment Algorithm in Hungarian Based on Syntactic Analysis. In *Proceedings of the INTERSPEECH-2007 Conference*, pages 466 – 469, Lancaster, UK.

Tihanyi, L. (2006). A MetaMorpho Projekt 2006-ban [The MetaMorpho Project in 2006]. In *Proceedings of the Hungarian Computational Linguistics Conference*. Szeged University Press.

Tjong Kim Sang, E. F. and Buchholz, S. (2000). Introduction to the CoNLL–2000 Shared Task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, pages 127–132, Morristown, NJ, USA. Association for Computational Linguistics.

Tolone, E. (2011). *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*. PhD thesis, LIGM, Université Paris-Est, France, Laboratoire d'Informatique Gaspard-Monge, Université Paris-Est Marne-la-Vallée, France.

Vajda, P., Nagy, V., and Dancsecs, E. (2004). A ragozási szótártól a nooj morfológiai moduljáig. [From the Dictionary of Inflections to the Morphological Module of NooJ]. In *Proceedings of the Hungarian Computational Linguistics Conference*, pages 183–190. Szeged University Press.

van den Eynde, K. and Blanche-Benveniste, C. (1978). Syntaxe et mécanismes descriptifs: présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32:3–27.

van den Eynde, K. and Mertens, P. (2006). *Le dictionnaire de valence Dicovalence : manuel d'utilisation*. Manuscript, Leuven.

van der Plas, L., Manguin, J.-L., and Tiedeman, J. (2008). Extraction de synonymes à partir d'un corpus multilingue aligné. In *Actes des journées de linguistique de corpus*, Lorient, France.

Van Valin, R. J. and Wilkins, D. (1996). The Case for Effector: Case Roles, Agents and Agency Revisited. In *Grammatical Constructions: Their Form and Meaning*, pages 289–322. Oxford University Press, Oxford.

Váradi, T. (2002). The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 385–389, Las Palmas, Spain.

Váradi, T. (2003). Shallow Parsing of Hungarian Business News. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 845 – 851, Lancaster, UK.

Váradi, T. and Gábor, K. (2004). A magyar Intex fejlesztésérõl. [On Developing the Hungarian Intex Module]. In *Proceedings of the Hungarian Computational Linguistics Conference*, pages 3–10. Szeged University Press.

Vendler, Z. (1957). Verbs and Times. *The Philosophical Review*, 66(2).

Véronis, J. (2003). Sense Tagging: does it Make Sense? In *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*. Frankfurt: Peter Lang.

Véronis, J. (2004). Hyperlex: Lexical Cartography for Information Retrieval. *Computer Speech and Language*, 18(3):223–252.

Vincze, O., Gábor, K., and László, J. (2010). Technológiai fejlesztések a NooJ pszichológiai alkalmazásában. [Improvements in the Psychological Module of NooJ]. In *Proceedings of the Hungarian Computational Linguistics Conference 2009*, Szeged, Hungary.

Voutilainen, A. (1997). Designing a (Finite–State) Parsing Grammar. In Roche, E. and Schabes, Y., editors, *Finite–State Language Processing*, pages 283–310. MIT Press, Cambridge.

Yli-Jyrä, A., Koskenniemi, K., and Lindén, K. (2006). Common Infrastructure for Finite-State Methods and Linguistics Descriptions. In *International Workshop Towards a Research Infrastructure for Language Resources. LREC 2006 Workshop*, Genova, Italy.

Zarcone, A. and Lenci, A. (2008). Computational Models for Event Type Classification in Context. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).