CZECH CLITICS IN HIGHER ORDER GRAMMAR

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By

Jiri Hana, Mgr., RNDr.

* * * * *

The Ohio State University

2007

Dissertation Committee:

Professor Carl Pollard, Adviser

Professor Brian Joseph

Professor Detmar Meurers

Professor Michael White

Approved by

____

Adviser

Graduate program in Linguistics

ABSTRACT

This thesis has three interrelated goals:

The main goal is an analysis of Czech clitics, units of grammar on the borderline between morphology and syntax with rather peculiar ordering properties both relative to the whole clause and to each other. We examine the actual set of clitics, their rather rigid ordering properties, and finally the properties of so-called clitic climbing. The analysis evaluates previous research, but it also provides new insights, especially in the position of the clitic cluster and in the constraints on clitic climbing. We show that many of the constraints regarding position of the clitic cluster suggested in previous research do not hold. We also argue that cases when clitics do not follow the first constituent are in fact not exceptions in clitic placement but instead unusual frontings.

The second goal is the development of a framework within Higher Order Grammar (HOG) supporting a transparent and modular treatment of word order. Unlike previous versions of HOG, we work with signs (containing phonological, syntactic and potentially other information) as actual objects of the grammar. Apart from that, we build on the simplicity and elegance of the pre-formal part of the linearization framework within Head-driven Phrase Structure Grammar.

Finally, the third objective is to test the result of the second goal by applying it on the results of the first goal.

To my grandfather František Andrlík

ACKNOWLEDGMENTS

The thesis did not emerge in vacuum and I would like to thank:

My advisor, Carl Pollard, for endless hours of discussions, for his willingness to explore my ideas even when he disagreed with the premisses, for those simple questions which showed in an instant that I did not have a clue what I was talking about.

Brian Joseph, Detmar Meurers and Mike White for serving on my committee, for their insights, comments, flexibility and suggestions. To Daniel Collins for serving as the Grad School representative, for reading the thesis in a very short time given to him by the Grad School and for his comments and suggestions.

Petr Sgall, for introducing me to linguistics, for feedback on early versions of the clitic analysis and for all the discussions about Czech syntax.

Alexander Rosen, for extensive feedback on the analysis Czech clitic, for all his examples, whether they supported or undermined my conclusions.

Andrea Sims and Richard Janda for comments on previous versions of the analysis of clitics.

Many other people who helped me with various linguistic and technical problems, most notably Mary Beckman, Mike Daniels, Markus Dickinson, Jakub Dotlačil, Eva Hajičová, Pavel Kosek, Denisa Lenertová, Marcela Michálková, Jarmila Panevová, Petr Savický, Shari Speer, Jana Šindlerová, Ludmila Uhlířová, Radoslav Večerka, Lída Veselovská, Šárka Zikánová.

Peter Culicover, who showed me that it is really possible to work effectively. I should have recalled more often his advice that it is easier to get lost in a long sentence than in a short one.

Chris Brew, for his *why* questions.

Anna for feedback, for being a friend and a colleague.

I would not have survived sane if it weren't for our pub evenings with George, Detmar, Luiz and Martin. I will really miss the heated discussions about free trade, wars, airlines or just about anything else. I would also like to thank Saša for similar support in Prague.

Finally, I would like to thank to Hanka, Filip and Matýsek for too many things to list them here, but especially for their patience . . .

VITA

1998	Mgr, Computer Science, Charles University, Prague
2001	RNDr, Computer Science, Charles University, Prague
1998–2001	Researcher, Charles University, Prague
2001–2006	Fulbright Fellow
2001–2002	University Fellow, The Ohio State University
2002–2006	Graduate Research and Teaching Associate, The Ohio State University
2006–2007	Presidential Fellow, The Ohio State University

PUBLICATIONS

Research Publications

Hana, J., A. Feldman, L. Amaral, and C. Brew (2006). Tagging Portuguese with a Spanish Tagger Using Cognates. In *Proceedings of the Workshop on Cross-language Knowledge Induction, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), Trento, Italy.*

Feldman, A., J. Hana, and C. Brew (2006). A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy.*

Feldman, A., J. Hana, and C. Brew (2006). Experiments in Morphological Annotation Transfer. In A. Gelbukh (Ed.), *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing)*, Lecture Notes in Computer Science. Springer-Verlag.

Hana, J. (2004). Czech clitics in Higher Order Grammar. In A. D. Sims and M. Whiting (Eds.), *Proceedings of the First Graduate Colloquium on Slavic Linguistics, November 8, 2003, at the Ohio State University; Working Papers in Slavic Studies*, Volume 3. Columbus, Ohio: Department of Slavic and East European Languages and Literatures. Significantly revised in 2005; backdated to 2004.

Hana, J. and A. Feldman (2004). Portable Language Technology: The case of Czech and Russian. In *Proceedings of the Midwest Computational Linguistics Colloquium, June 25-26, 2004*, Bloomington, Indiana.

Hana, J., A. Feldman, and C. Brew (2004). A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2004*, Barcelona, Spain.

Pollard, C. and J. Hana (2003). Ambiguity, neutrality, and coordination in Higher Order Grammar. In G. Jaeger, P. Monachesi, G. Penn, and S. Wintner (Eds.), *Proceedings of Formal Grammar*, Wien, pp. 125–136.

Kruijff, G.-J., E. Teich, J. Bateman, I. Kruijff-Korbayová, H. Skoumalová, S. Sharoff, L. Sokolova, T. Hartley, K. Staykova, and J. Hana (2000). A multilingual system for text generation in three Slavic languages. In *Proceedings of the 18th Conference on Computational Linguistics (COLING), July 31 - August 4 2000*, pp. 474–480. Universität des Saarlandes, Saarbrücken, Germany.

FIELDS OF STUDY

Major Field: Linguistics

TABLE OF CONTENTS

Chapters:

Appendices:

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

This thesis has three interrelated goals. The first one is the description of Czech clitics, units of grammar on the borderline between morphology and syntax with rather peculiar word-order properties. The second one is the development of a framework within Higher Order Grammar supporting a transparent and modular treatment of word order. Finally, the third goal is to test the framework by using it to formalize the analysis of clitics.

1.1 Word Order, Clitics

Although languages around the world share many common properties, they also significantly differ. One type of difference is in word order and its interpretation. For example, in the English sentence *Martin gave Paul a book,* Martin is giving and Paul is receiving the book. If one swaps these two names, the sentence means something very different. On the other hand, in Czech, any scrambling of the words in (1) is not only grammatical but also consistently means that Martin was giving and Paul was receiving. The sentences differ in their pragmatics but the basic meaning is the same. Czech uses morphological endings instead of word order to mark subject and object. Many languages are like English, many are like Czech, and many are somewhere in between.

(1) Martin dal Pavlovi knihu. (Czech)
Martin$_{nom}$ gave Pavel$_D$ book$_A$
'Martin gave Pavel a book.'

Another type of word-order differences pertains to whether or not a language prefers phrases to be continuous or whether there can be discontinuities. English prefers the former, while other languages are more complicated and allow phrases to have 'holes' where pieces of other phrases can be inserted. There are various factors causing these discontinuities. For example, in Dutch or German, the word order of verb complexes is highly restricted. Verbs occur in a particular order at the end of the sentence. On the other hand, their arguments occur relatively freely in the so-called Mittelfeld.

Among the more complex discontinuities are those caused by sentential clitics – units that are transitional between words and affixes, sometimes behaving like the former and sometimes like the latter. Consider, for example the Serbo-Croatian sentence in (2). The phrase *taj pisac* 'that author' is interrupted by clitics *bi* 'would' and *mi* 'to-me'.

(2) Taj *bi* *mi* pisac napisao knigu. (Serbo-Croatian, clitics in italics)
That$_{m.nom}$ would me$_D$ author$_{m.nom}$ write book$_A$
'That author would write me a book.'

The Czech sentence in (3) also contains discontinuities: the infinitival phrase phrase *opravit mu to* 'to repair it for him' is interrupted by the auxiliary verb *bych* 'would' and a reflexive particle *se* (belonging to *nesnažil* 'not-try').

(3) Opravit *bych* *se* *mu* *to* nesnažil. (Czech, clitics in italics)
repair$_{\mathsf{inf}}$ would$_{1sg}$ refl$_A$ him$_D$ it$_A$ not-tried
'As for repairing it for him, I would not try it.'

Clitics in general, and Slavic clitics in particular, present a great challenge to existing formalisms. Their ordering principles are complex and quite different from the properties of those that govern normal words and phrases. Also, they are subject to interacting constraints arising from various levels of grammar – syntactic, morphological, phonological, pragmatic and stylistic. While Czech word order is very free, the position of clitics is quite restricted. They accumulate in certain fixed positions within the sentence – see example (3) – and even their ordering within these positions is for the most part fixed.

Similar kinds of phenomena occur in many languages around the world belonging to very different language families, including French, Spanish, Albanian, Pashto and Tagalog. Although any six-year old native speaker of such a language can use sentences similar to those two above without any problems, linguists have struggled for decades to uncover the principles that determine which orderings are possible.

An even bigger problem is to express these rules in a precise and formal way. Expressing grammars of natural languages formally is important for at least two reasons: (i) it facilitates scientific progress: it is easier to test, falsify and compare hypotheses that are precisely formulated than those formulated in an unclear and vague way; (ii) a formal grammar can serve as a basis for automatic systems processing language. Any such grammar covering more than just a trivial language fragment is enormously complex and requires cooperative work of many specialists.

1.2 Higher Order Grammar

Higher Order Grammar (HOG; e.g., Pollard 2004*a*; Pollard and Hana 2003) builds on the positive attributes of Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag 1994), while avoiding its pitfalls. HOG is based on typed higher order logic/lambda calculus (Church 1940). This means it supports higher-order functions, i.e., functions that can receive other functions, possibly also higher-order, as parameters. Research in computer science (e.g., Hughes 1989; Thompson 1997) has identified higher order functions as crucial for achieving modularity and reusability. Both of these properties are important for linguistics. Modularity is essential for creation of any large scale grammars, necessary a team development. Reusability is important for both theoretical and practical reasons. On the one hand, reusability is central to linguistic work, usually referred to as "capturing generalizations". On the other hand, a high degree of reusability means common aspects of different grammar rules are written just once, thus making the grammar more transparent and less error-prone. Also, higher-order logic, the underlying logic of HOG, unlike RSRL (Richter 2000), the underlying logic of HPSG, is a standard mathematical theory; therefore, necessary cooperation with specialists in computer science and mathematics becomes much easier.

1.3 Roadmap

In **Chapter 2 Higher Order Grammar**, we discuss the basic setup of Higher Order Grammar. HOG is a grammar of signs. Signs capture semantic, tectogrammatical (abstract/deep syntactic) and phenogrammatic (concrete/surface syntactic) properties of language expressions. A grammar describes their individual components, their relations and their combinations. We also compare HOG to other frameworks.

In **Chapter 3 Basics of Czech word order**, we describe and analyze basic properties of Czech word order, including its relation to Information Structure, the integration of sentences into discourse. The data and conclusions from this chapter are meant both as a place of reference for the following chapter analyzing Czech clitics and as a case-study for the linearization framework developed in Chapter 5.

In **Chapter 4, Czech Special Clitics**, we provide an analysis of a certain class of Czech special clitics. We examine and characterize the set of Czech clitics; identify their position within the clause and then the order of clitics within this position. Finally we analyze so-called clitic climbing.

In **Chapter 5, Czech in HOG**, we gradually develop a simple grammar of Czech in HOG. Its tectogrammar and phenogrammar. This chapter also develops the bulk of a linearization framework for HOG. At the end, we focus on Czech clitics, formalizing most of the empirical results from Chapter 4.

Appendix A, Czech describes some basic properties of Czech grammar, focusing mainly on morphology and to some extent on syntax. The main purpose of this appendix is to provide some guidance for a non-native speaker when reading the Czech examples in the thesis.

Appendix B, Data, Examples, Glosses discusses the source of examples (mostly various corpora) and the format of glosses.

Appendix C, HOL for HOG summarizes the formal foundations of Higher-Order Grammar. It provides definitions and some discussion for all the terms used. Appendix C can be seen as a formal complement to the informal introduction to the logic of HOG in the beginning of Chapter 2.

Appendix D, General functions summarizes general purpose, language independent functions used in Chapter 5.

CHAPTER 2

HIGHER ORDER GRAMMAR

2.1 Basic properties

Higher Order Grammar (HOG; e.g., Pollard 2001*a*,*b*, 2004*a*,*b*,*c*; Pollard and Hana 2003; Hana 2004) is a logical framework for linguistic analysis that can be viewed simultaneously as generative-enumerative, like Type Logical Grammar (Morrill 1994) and Principles & Parameters (Chomsky 1981), or model theoretic, like Head-Driven Phrase Structure Grammar (Pollard and Sag 1994) or Lexical Functional Grammar (Bresnan 2001).

HOG consists of three subtheories – phenogrammar, tectogrammar and semantics. The distinction between tectogrammar (abstract syntax) and phenogrammar (concrete syntax + phonology) follows Curry's 1961 distinction between what he called phenogrammatics and tectogrammatics. Tectogrammar captures the abstract way in which linguistic signs combine (argument structure), while phenogrammar handles the concrete processes of string formation. Thus, tectogrammar handles traditional syntax except for word order, i.e., government, syntactic argument structure, morphosyntactic agreement, etc. Phenogrammar captures word order and phonology, including prosody. A more or less similar distinction between tectogrammar and phenogrammar has been made by many others (e.g., de Groote 2001; Dowty 1996; Kathol 1995; Muskens 2001*b*; Penn 1999*a*; Ranta 2004; Reape 1994; Sgall et al. 1969, see §2.10 for comparison). Semantics models Fregean senses but unlike in Montague's Semantics (Dowty et al. 1981; Montague 1970), intensions are not simulated via possible worlds. Such a multistratal setup has formal and practical advantages. Most importantly, the framework is modular, allowing each level and dimension to be studied to a great extent separately.

HOG is based on typed higher order logic/lambda calculus (Church 1940) and influenced by Lambek's (1988; 1999) categorical (i.e., expressed in Category Theory) grammar. The advantage of using HOL lies not only in its expressiveness, but also in the fact that it is a standard off-the-shelf formalism with extensive research already completed in both formal and computational areas.

In the following, we discuss the basic setup of HOG, first its individual components and then their cooperation. Finally, we compare HOG to other frameworks. Note that HOG is lstill under development, thus the cited papers differ not only in their approach to particular linguistic phenomena, but often also in details of the underlying framework. The final section, comparing HOG to other theories, discusses briefly various variants of HOG as well.

2.2 Formalism of HOG informally

In this section, we very briefly and informally introduce the major constructs of the formalism used in Higher Order Grammar. Appendix §C contains a much more thorough and formal presentation of the same topic. Figure 2.1 provides a brief overview of the basic notation used.

A grammar in HOG is a theory in Higher Order Logic (Church 1940), which is based on a typed lambda calculus. This means that functions (and relations) in HOG are first class citizens; for example, they can be passed to other functions as arguments. In this respect, the formalism resembles a typed functional programming language like Haskell (haskel.org) or ML (Milner et al. 1997). The logic can be thought of as Ty2 (Gallin 1975, a higher order logic equivalent to Montague's IL (Montague 1970, 1973)) with some additions. HOG has a larger set of basic types, including a type of natural numbers. It has also tuples, similar to records in some programming languages.[1] The type system is (schematically) polymorphic and allows the definition of supertypes and separation subtypes.

Types Every expression in HOG must have a type. This is written as $\mathit{term} : \mathsf{Type}$, for example:

(1)

$1 : \mathbb{N}$	a term of the type $\mathbb{N}$ (natural numbers)
$+ : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$	addition function
$\lambda x \,.\, x + 1 : \mathbb{N} \to \mathbb{N}$	an increment function
$\mathsf{dog} : \mathsf{N}$	a term of the type N
$\mathsf{the} : \mathsf{N} \to \mathsf{NP}$	a function taking a noun and returning an NP
$\alpha \,\&\, \neg\alpha : \mathsf{Bool}$	a boolean formula (assuming α is also of type Bool)

A type can be thought of as a set of terms of that type. In fact, the denotation of a type in some interpretation of the theory is a set that contains the denotations of the type's terms as elements. In formulas, typing is often omitted when it is possible to unambiguously infer it from context. The type system consists of a set of basic types like NP, S, etc. From these types, other types are constructed by a set of type constructors like functions and tuples (records). Lists and sets are definable.

[1] We use the word *tuple* to refer to both tuples indexed by natural numbers and by other indexes (e.g., SUBJ , TECTO). We also use the word *record* to refer to the tuples with non-numeric indexes.

1. Typeface:
    (a) types: $\mathsf{Bool}, \mathsf{NP}, \mathsf{Tecto}$
    (b) type constructors: $\mathsf{List}(\mathsf{A})$
    (c) terms: $\mathsf{concat}, \mathsf{kim}$
    (d) variables: A, B (types), a, b (terms)
    (e) tuple indexes (record attributes): $\textsc{subj}, \textsc{comps}, \textsc{tecto}$
    (f) phonology – usually replaced by spelling: *eat* used instead of /it/
2. Subtypes and supertypes:
    (a) subtype of A defined by an A-predicate φ: $\mathsf{A}_\varphi = [\, x : \mathsf{A} \,|\, \varphi(x) \,]$
    (b) supertype (coproduct) of A and B: $\mathsf{A} + \mathsf{B}$
    (c) predicate testing a type of a term: $a :: \mathsf{A}$ (note that $a : \mathsf{A}$ is not a predicate)
3. Functions:
    (a) functions from type A to type B: $\mathsf{A} \to \mathsf{B}$
    (b) lambda abstraction: $\lambda x : \mathsf{A} \,.\, b$
    (c) composition is written in the order of application: $f.g = \lambda x \,.\, g(f(x))$.
    (d) Application of a one argument function is written as $\mathsf{case}(x)$ or as $x.\mathsf{case}$.
4. Tuples (records, products):
    (a) type: $[\textsc{subj}\ \mathsf{NP},\ \textsc{comps}\ \mathsf{NP}]$ ($\textsc{subj}, \textsc{comps}$ are indexes)
    (b) term: $[\textsc{subj}\ \mathsf{john},\ \textsc{comps}\ \mathsf{mary}]$
    (c) indexes are used as projections: $[\textsc{subj}\ \mathsf{john}, \textsc{comps}\ \mathsf{mary}].\textsc{subj} = \mathsf{john}$
    (d) natural number indexes are omitted: $[\mathsf{john}, \mathsf{mary}] = [1\ \mathsf{john}, 2\ \mathsf{mary}]$.
    (e) tuple terms and types are often written as AVMs
5. Collections:
    (a) type: $\mathsf{Set}(\mathsf{A}) := \mathsf{A} \to \mathsf{Bool}$ (set); $\mathsf{List}(\mathsf{A})$ or A^* (list)
    (b) term: $\{\mathsf{a},\ \mathsf{b},\ \mathsf{c}\}$ (set); $\langle \mathsf{a},\ \mathsf{b},\ \mathsf{c} \rangle$ (list) (a, b, c have the same type)
    (c) singularizer: $\mathsf{sing} : \mathsf{Set}(\mathsf{A}) \to \mathsf{A}$; $\mathsf{sing}(\{\mathsf{a}\}) = \mathsf{a}$
6. Other:
    (a) logical connectives: $\Rightarrow$ (implication), & (conjunction), $\vee$ (disjunction), $\Leftrightarrow$ (equivalence), $\neg$ (negation), $\forall$ (universal quant.), $\exists$ (existential quant.)
    (b) statement (axiom or theorem); constraint on Sign: $\vdash\ \varphi$ (φ is a predicate)
    (c) typing judgement (axiom or theorem): $\vdash\ a : A$ (there is a term a and it is of type A)

Figure 2.1: Overview of HOG notation

Montague semantics also uses a typed lambda calculus, but it has no tuples (hence all the curried functions). HPSG is not higher-order – its type system has no functional types; moreover relations are not typed.

Subtypes and supertypes (see §C.4 for more details) There are two ways to express a type-subtype relationship in HOG.

1. For any countable set of types we can define a supertype of those types. There are two basic ways of defining a supertype:

   (a) By explicitly listing the set of subtypes.

   For example, $\mathsf{NominalP}$ is the type of all noun phrases and adjectival phrases:

   (2) $\mathsf{NominalP} = \mathsf{NP} + \mathsf{AP}$

   (b) By closing a set of types by a type constructor $\mathsf{ClosingSupertype}$. The type constructor takes a set of types, closes the set by the available type constructors and returns the (smallest) supertype of the closed set.

   For example, Tecto is the type of all tecto phrases, i.e., the supertype of all types obtained by closing the set of basic types by the available type constructors:

   (3) $\mathsf{Tecto} = \mathsf{ClosingSupertype}(\{\mathsf{NP}, \mathsf{N}, \mathsf{S}\})$

   Thus Tecto is a supertype of NP, N and S, but also of $\mathsf{NP} \to \mathsf{NP}$, $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$, $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$, etc.

2. For any type and a predicate on that type (i.e., a function from that type into Bool), we can define a subtype determined by the predicate.

   For example, the type of all accusative noun phrases is denoted as:

   (4) $\mathsf{NP}_{\lambda x:\mathsf{NP}.\mathsf{case}(x)=\mathsf{acc}}$ or $[\, x : \mathsf{NP} \,|\, \mathsf{case}(x) = \mathsf{acc} \,]$, usually written as $\mathsf{NP}_{\mathsf{acc}}$

The formalism is polymorphic (although only schematically), which means we can define type operators (functions receiving types as arguments and/or returning them as results). For example the type operator List for any type gives the type of the lists with that type; if $\mathsf{nps} : \mathsf{List}(\mathsf{NP})$, then nps is a list of NPs. Similarly, it is possible to define functions applicable to various types. For example, $\mathsf{reverse} : \mathsf{List}(A) \to \mathsf{List}(A)$ is a polymorphic function reversing list of any type.

Curry-Howard isomorphism (§C.7.1) The Curry-Howard isomorphism (Curry and Feys 1958; Howard 1980) states that the type system forms a full intuitionistic propositional logic. Type expressions, like NP, $\mathsf{NP} \times \mathsf{NP} \to \mathsf{S}$ are propositions in that logic. In (5), the notation usually used with expressions propositional logic is compared to the notation used for type expressions.

(5)

$\Rightarrow$	implication	$\to$	function space (exponential)
&	conjunction	$\times$	product (tuple type, record type)
$\vee$	disjunction	$+$	coproduct (disjoint union)
true	true	Unit	nullary product
false	false	Zero	nullary coproduct
$\neg$	negation		defined as $\neg \mathsf{A} = \mathsf{A} \to \mathsf{Zero}$
	atomic formulas		basic types

In such a view, (closed) terms are proofs. A type expression i.e., a formula in the logic of types, is a theorem if there is a term of that type. In other words, a type is "true" if it is inhabited, which means that in any model of the logic, there are objects which are members of the interpretation of the type. Constants (basic terms) are then nonlogical axioms.

Models, proofs (§C.7) A model of a grammar in HOG is an interpretation of the logic, in which all the theorems of the term logic have the same interpretation as the formula true, i.e., a Henkin model (Henkin 1950). In such a model, syntactic categories like NP correspond to sets. Words and phrases correspond to members of those sets. Subtypes are subsets. Thus HOG is a model theoretic framework, like Head-Driven Phrase Structure Grammar (Pollard and Sag 1994) or Lexical Functional Grammar (Bresnan 2001).

Because of the Curry-Howard isomorphism, the model-theoretic aspect of the grammar is automatically connected with the proof-theoretic aspect. The types are formulas in a propositional logic and the terms are proofs. In this respect, HOG is a generative-enumerative framework, like Type Logical Grammar (Morrill 1994), Combinatory Categorial Grammar (Steedman 2000*b*; Steedman and Baldridge 2003) and Principles & Parameters (Chomsky 1981; Chomsky and Lasnik 1993).

2.3 Signs

As mentioned above, HOG consists of three subtheories – tectogrammar, phenogrammar and semantics. All three theories are theories in a higher-order logic. While each of the theories describes particular dimension of a language, none of them describes it exhaustively. Only some objects licensed in each of these theories make sense from the point of view of the other theories. For example, while a sequences of phonological words /bɾuː tidi paluɾabə ə'bɔːrkə/ would be probably licensed by any English phenogrammar (it satisfies all the phonotactical rules), it would not correspond to any tectogrammatical nor semantic terms.

Words, phrases and sentences are modeled as signs. Signs are the combinations of tecto,[2] pheno and semantic objects that 'makes sense'. The pheno component of a sign is the pronunciation of the tecto component and the semantic component is the meaning of it. The type Sign is a subtype of the following tuple type:

(6) $$\begin{bmatrix} \textsc{tecto} & \mathsf{Tecto} \\ \textsc{pheno} & \mathsf{Pheno} \\ \textsc{sem} & \mathsf{Meaning} \end{bmatrix}$$

The grammar specifies the set of signs in a recursive manner. The lexicon determines what are the basic signs and then there are constraints determining how signs are combined into more complex signs.

For example, the lexicon specifies that there is a sign corresponding to the word *cat* which consists of the tecto object cat, the pheno object [/kæt/], and the semantic object cat'. This is stated as non-logical axiom in (7).

(7) $$\vdash \begin{bmatrix} \textsc{tecto} & \mathsf{cat} \\ \textsc{pheno} & \langle/\text{kæt}/\rangle \\ \textsc{sem} & \mathsf{cat'} \end{bmatrix} :: \mathsf{Sign}$$

In addition, there are also triples like the one on the left side of (8) that are not signs.

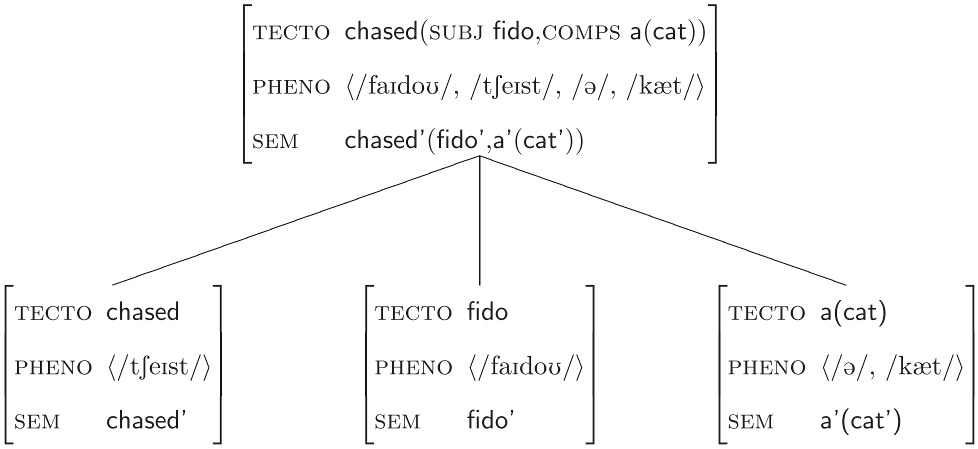
(8) $$\vdash \begin{bmatrix} \textsc{tecto} & \mathsf{snores}(\textsc{subj}\ \mathsf{kim}) \\ \textsc{pheno} & \langle/\text{kaʊ}/\rangle \\ \textsc{sem} & \mathsf{cat'} \end{bmatrix} : \begin{bmatrix} \textsc{tecto} & \mathsf{Tecto} \\ \textsc{pheno} & \mathsf{Pheno} \\ \textsc{sem} & \mathsf{Meaning} \end{bmatrix}$$

The grammar would also specify that the three signs at the bottom of (9), corresponding to *chased*, *Fido* and *a cat*, can combine into a sign at the top corresponding the sentence *Fido chased a cat.* A constraint might require that whenever a transitive verb (chased) combines with its subject (fido) and object (a(cat)) in tectogrammar, the pheno of the complex sign will be the concatenation of

[2] We use the term *tecto* for *tectogrammatical* and *pheno* for *phenogrammatical*.

pheno objects corresponding to the subject (⟨/faɪdoʊ/⟩), the verb (⟨/tʃeɪst/⟩) and the object (⟨/ə/, /kæt/⟩).

(9)



Therefore, a HOG grammar can be split into two parts:

1. constraints on individual components of signs. This means there are separate theories of tectogrammar, phenogrammar and semantics.
2. constraints on whole signs and their possible combinations. These can be further divided as
   (a) general constraints on individual signs.

   For example, it can be required that certain tecto terms (e.g. finite clauses) correspond to continuous pheno object.

   (b) lexical constraints determining the set of basic signs. The signs correspond to individual words and possibly idioms.

   (c) constraints specifying how to combine signs into more complex signs. For each possible combination in one component of the grammar, it must be specified what combination (if any) happens in the other parts.

The distinction between tectogrammar, phenogrammar and semantics can be seen as similar to the distinction between deep/surface syntax, phonetic form and logical forms, respectively, in Principles and Parameters (Chomsky and Lasnik 1993). However, there is an important difference. In HOG, the relation of the three components is compositional. P & P, on the other hand, relates the levels of representation for whole sentences.

Below, we first introduce the three HOG sub-theories: tectogrammar, phenogrammar and semantics. After that, we discuss the part of HOG dealing with whole signs, first the lexicon then the induction rules combining signs.

2.4 Tectogrammar

The main purpose of tectogrammar is to take care of the combinatorics of linguistic expressions – valency and modification. Linguistics, often figuratively, talks about verbs as functions. In HOG's tectogrammar, they *are* functions. For example, a transitive verb is of the type $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$, i.e., it is a function accepting a tuple of two NP's and returning a sentence. Tectogrammar handles only unordered hierarchical structure; the word order considerations are dealt with in phenogrammar.

2.4.1 Non-linear logic

This division of labor means that HOG can be a theory in an ordinary higher-order intuitionistic logic, with all structural rules:

- It is commutative. Therefore, it needs only one functional type constructor: ($A \to B$, implication in the type logic). Categorial Grammar has at least two function types: $B\backslash A$ and B/A depending whether the function takes its argument on the left or on the right.[3]
- It does not need to do book-keeping of resources (i.e., it does not need to ensure that every word token in a sentence is used exactly once) as linear logic does, since that is handled by phenogrammar.

Therefore, HOG's tectogrammar is similar to the abstract syntax of Grammatical Framework (Ranta 2004; see §2.10.5) which is also not linear, i.e., commutative with no resource book-keeping. However, in GF the underlying formalism is a lambda calculus without equality, not a logic. On the other hand, tectogrammar in Lambda Grammars (Muskens 2004; see §2.10.4) is expressed in linear logic.

2.4.2 Specification

To specify a tectogrammar, it is necessary to provide:

1. types:
   (a) basic types of expressions:[4] e.g., $\mathsf{N}, \mathsf{NP}, \mathsf{S}$.

[3] It is however possible to introduce multiple implications for other reasons than capturing word order; see §2.10.2.

[4] We use *tecto expressions* to refer to those tecto-objects that model tecto phrases and words. For example, objects of the type NP, S or $[\textsc{subj}\ \mathsf{NP}] \to \mathsf{S}$, but not of types such as Case, $\mathsf{Number} \to \mathsf{Bool}$, which model other linguistic properties.

The type Tecto is a supertype of all tecto types.

In HOG, syntactic categories are not thought to be labels of expressions but rather denote sets of expressions. Thus NP is a type of all tecto terms corresponding to noun phrases and its interpretation has the interpretation of those terms as its members.

For convenience, it is possible to give some of the more often used types a name. For example, $\mathsf{Det} = [\textsc{spec}\ \mathsf{N}] \rightarrow \mathsf{NP}$. Note, however, that once the set of types contains types N and NP and there is an index SPEC, the type $[\textsc{spec}\ \mathsf{N}] \rightarrow \mathsf{NP}$ exists, whether it is given the name Det or not.

(b) types of "feature values", e.g., $\mathsf{Case}, \mathsf{Number}$.

(c) product indexes (record attributes), e.g., SUBJ, COMPS, SPEC

2. constants:

(a) features, e.g., $\mathsf{case} : \mathsf{NP} \rightarrow \mathsf{Case}$[5]

(b) feature values, e.g., $\mathsf{nom}, \mathsf{acc} : \mathsf{Case}$

(c) tecto words, e.g., $\mathsf{we} : \mathsf{NP}_{\mathsf{nom}}$, $\mathsf{saw} : [\textsc{subj}\ \mathsf{NP}_{\mathsf{nom}}, \textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}] \rightarrow \mathsf{S}_{\mathsf{fin}}$

The lexicon then constrains what their pheno and semantics can be.

2.4.3 Sample tecto grammar

Now, we are ready to present a simple tectogrammar licensing tecto terms corresponding to the words and phrases in the sentence *Fido chased a cat.* First, we assume that there are types of the basic syntactic categories

(10) $\mathsf{NP}, \mathsf{N}, \mathsf{S}$

with the obvious motivation. We choose one term of the type NP to correspond to the expression *Fido*:

(11) $\vdash\ \mathsf{fido} : \mathsf{NP}$

Note that fido is just a label. Instead, we could also write 123-17-B or eats or charles-bridge for the same term, as long as we were consistent and made sure that it were pronounced by the phenogrammar as /faɪdoʊ/.

[5] Although there are different ways to handle case, see §5.1.3.

We also pick one term of the type N to correspond to the expression *cat*:

(12) $\vdash\ \mathsf{cat} : \mathsf{N}$

In addition to the three primitive types, there are all the types derived by available type constructors, including all tuple and function types. We are interested in a type of transitive verbs – i.e., a type of functions accepting a subject NP and an object NP and returning a sentence. Assuming there are tuple indexes SUBJ and COMPS, the type is:

(13) $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \rightarrow \mathsf{S}$

We again choose one term of that type to correspond to the transitive verb *chased*

(14) $\vdash\ \mathsf{chased} : [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \rightarrow \mathsf{S}$

Finally, determiners can be modeled as a functions accepting nouns as specifiers and returning NPs. Their type is then

(15) $[\textsc{spec}\ \mathsf{N}] \rightarrow \mathsf{NP}$

and one term of that type can be used to model the indefinite article *a*:

(16) $\vdash\ \mathsf{a} : [\textsc{spec}\ \mathsf{N}] \rightarrow \mathsf{NP}$

Sample derivation. We can now show how the tecto terms in the informal derivation in (9) could be actually licensed by the tectogrammar. The derivation/proof, which is schematically depicted in Figure 2.2, proceeds as follows:

1. The grammar states that

   (17) $\vdash\ \mathsf{cat} : \mathsf{N}$

2. The underlying logic guarantees that we can form tuples out of any set of terms (see §C.2). Therefore, we know the following statement holds:

   (18) $\vdash\ [\textsc{spec}\ \mathsf{cat}] : [\textsc{spec}\ \mathsf{N}]$

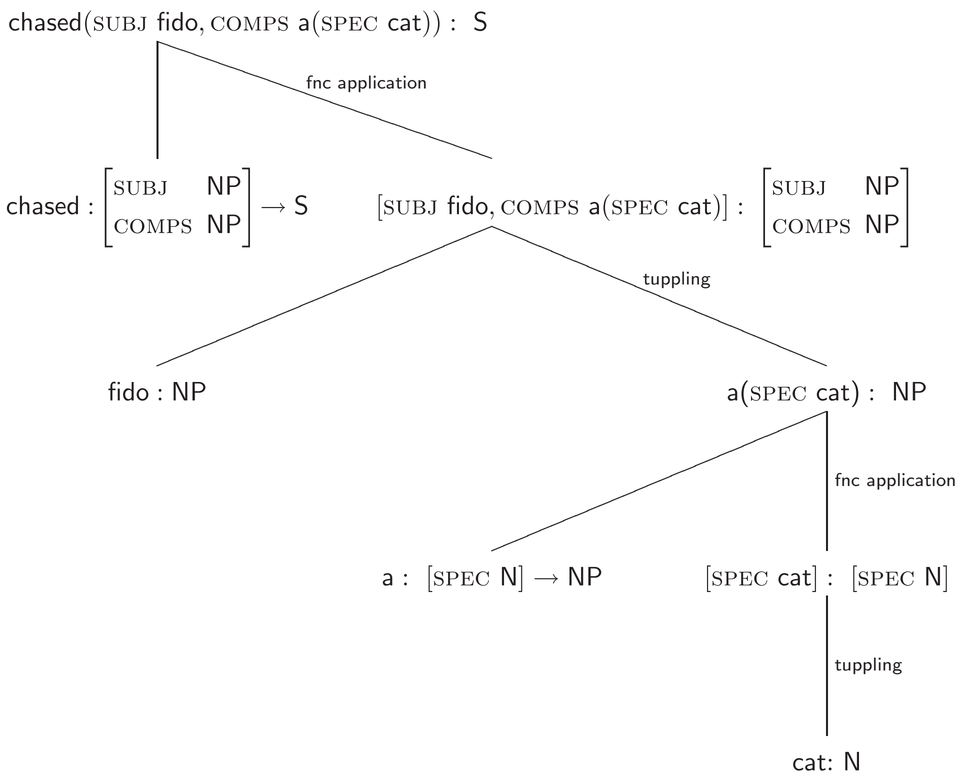

chased(SUBJ fido, COMPS a(SPEC cat)) : S
fnc application
chased : [SUBJ NP, COMPS NP] → S
[SUBJ fido, COMPS a(SPEC cat)] : [SUBJ NP, COMPS NP]
tuppling
fido : NP
a(SPEC cat) : NP
fnc application
a : [SPEC N] → NP
[SPEC cat] : [SPEC N]
tuppling
cat: N

Figure 2.2: Tecto derivation of *Fido chases a cat.*

3. By applying the determiner a

   (19) $\vdash\ \mathsf{a} : [\textsc{spec}\ \mathsf{N}] \rightarrow \mathsf{NP}$

   on the tuple in (18), we can show that there is a term of type NP that we can write as a(SPEC cat):

   (20) $\vdash\ \mathsf{a}(\textsc{spec}\ \mathsf{cat}) : \mathsf{NP}$

   Note that, while the format of the term a(SPEC cat) suggests the way in which it could have been derived, formally terms are indivisible and they do not record history of the way they were created. Instead of a(SPEC cat) we could have written term237114, because given a term, there is no way how to get the components it was created from. Similarly, given the number 4, there is no way to tell whether it is the result of $2 + 2$ or $3 + 1$; or given a list we do not know which lists, if any, it is a concatenation of.

4. Then we can create a tuple:

   (21) $\vdash$ $[\textsc{subj}\ \mathsf{fido}, \textsc{comps}\ \mathsf{a}(\textsc{spec}\ \mathsf{cat})] : [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$

5. Because the term chased

   (22) $\vdash$ $\mathsf{chased} : [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$

   accepts exactly that type of arguments, we know that there is a term of the type S, that we again write in a suggestive way as

   (23) $\vdash$ $\mathsf{chased}(\textsc{subj}\ \mathsf{fido}, \textsc{comps}\ \mathsf{a}(\mathsf{cat})) : \mathsf{S}$

   Which is what we wanted to show.

According to the Curry-Howard isomorphism (see §2.2, §C.7.1), $\mathsf{chased}(\textsc{subj}\ \mathsf{fido}, \textsc{comps}\ \mathsf{a}(\mathsf{cat}))$ is a proof of S. We have arrived at this proof from four non-logical axioms:

(24) $\vdash$ $\mathsf{fido} : \mathsf{NP}$
$\vdash$ $\mathsf{chased} : [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$
$\vdash$ $\mathsf{a} : [\textsc{spec}\ \mathsf{N}] \to \mathsf{NP}$
$\vdash$ $\mathsf{cat} : \mathsf{N}$

and the axioms of Higher Order Logic, allowing us to form tuples, and to apply functions on suitable arguments. In Curry-Howard isomorphism, tupling corresponds to conjunction introduction and function application corresponds to modus ponens (implication elimination).

It would be possible to prove a simple schematic lemma that would allow us to do tupling and function application in one step and thus hide the technical step of tupling, which is only needed to model/simulate functions of multiple arguments.[6] The proof trees would then look as usual (flat) syntactic structures. The proof of *Fido chased a cat* using such a lemma is in Figure 2.3.

2.5 Phenogrammar

In this chapter, we assume a very simple pheno-grammar: pheno objects are simply lists of phonological words, thus $\mathsf{Pheno} = \mathsf{PhonWord}^*$(recall that A^* is a list of elements of type A). Because

[6] Another possibility is currying (see §C.2.1). We could also provide functions of multiple arguments as primitives, similarly as, for example, in (Crole 1993).

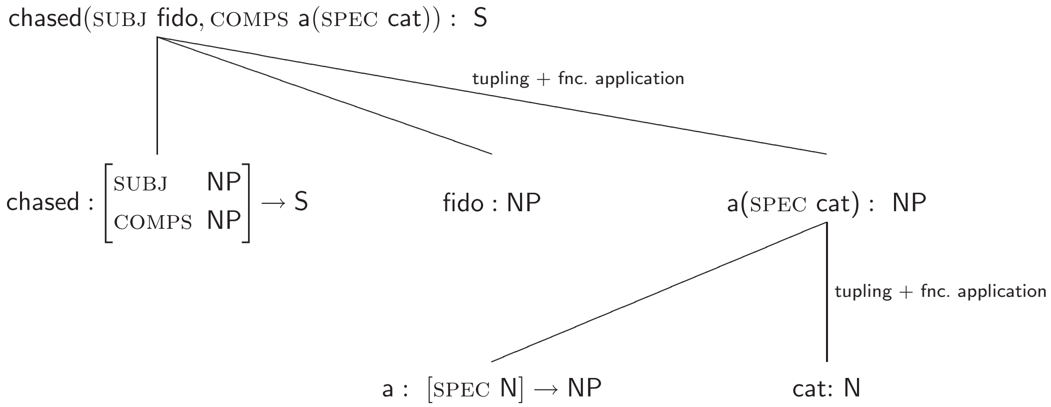

chased(SUBJ fido, COMPS a(SPEC cat)) : S
tupling + fnc. application
chased : [SUBJ NP, COMPS NP] → S
fido : NP
a(SPEC cat) : NP
tupling + fnc. application
a : [SPEC N] → NP
cat: N

Figure 2.3: Tecto derivation of *Fido chases a cat.* with "hidden" tupling

$\mathsf{PhonWord}^*$ is a list, the type forms a monoid with concatenation being the associative binary operation and empty list being the unit. A phonological word is a sequence of phonemes satisfying phonotactic constraints of the language. This means the type PhonWord is a subtype of the type Phoneme* determined by some predicate phonotactic-constraints formalizing constraints on possible lists of phonemes:

(25) $\mathsf{PhonWord} = [\, x : \mathsf{Phoneme}^* \mid \mathsf{phonotactic\text{-}constraints}(x)\,]$

Axioms like (26) assure existence of individual phonemes. Obviously, a realistic grammar would work with a less primitive notion of phonemes, introducing at least phonetic features along the lines of (Höhle 1999).

(26) $\vdash$/f/ : Phoneme

$\vdash$/a/ : Phoneme

$\vdash$/ɪ/ : Phoneme

$\vdash$/d/ : Phoneme

$\vdash$/o/ : Phoneme

$\vdash$/ʊ/ : Phoneme

and so on

For each phoneme in (26), we can form a singleton list containing that phoneme, thus we can prove:

(27) ⊢⟨/f/⟩ : $\mathsf{Phoneme}^*$
⊢⟨/a/⟩ : $\mathsf{Phoneme}^*$
and so on

We can concatenate those singleton lists, proving

(28) ⊢⟨/f/,/a/,/ɪ/,/d/,/o/,/ʊ/⟩ : $\mathsf{Phoneme}^*$

which we write simply as

(29) ⊢/faɪdoʊ/ : $\mathsf{Phoneme}^*$

Assuming that the sequence of phonemes satisfies phonotactic-constraints of the grammar, we can also prove that /faɪdoʊ/ is a phonological word:[7]

(30) ⊢/faɪdoʊ/ : $\mathsf{PhonWord}$

Again, we can form singleton lists containing phonological words and concatenate them, thus we can prove:

(31) ⊢⟨/faɪdoʊ/, /tʃeɪst/, /ə/, /kæt/⟩ : $\mathsf{PhonWord}^*$

Note, however that we can also prove

(32) ⊢⟨/faɪfaɪfaɪfaɪfaɪfaɪ/⟩ : $\mathsf{PhonWord}^*$

or

(33) ⊢⟨/faɪdoʊ/, /faɪdoʊ/, /faɪdoʊ/, /faɪdoʊ/, /ə/, /ə/, /ə/⟩ : $\mathsf{PhonWord}^*$

It is the cooperation of phenogrammar and tectogrammar (and semantics) that rules such sequences of phonemes and phonological words out.

[7] Formally, this is a little bit more complicated. As discussed in §C.4.4, the logic used in HOG requires that every term belongs to exactly one type (so-called *monotyping* property). Therefore formally, the two terms in (29) and (30) are two distinct terms. They are related by $\mathsf{ker}_{\mathsf{PhonWord},\mathsf{Phoneme}*}$, i.e., the function embedding $\mathsf{PhonWord}$ into $\mathsf{Phoneme}^*$. In this and similar cases, we abuse the notation and write both terms in the same way.

In Chapter 5, we introduce a more realistic phenogrammar suitable for capturing complex word-order constraints and discontinuities. In such phenogrammar, phonology is just one feature in a more structured pheno object.

2.6 Semantics

HOG can accommodate various types of semantics, including Montague Semantics (Dowty et al. 1981; Montague 1970). Below, we provide an overview of an integration of HOG and Hyperintensional Semantics (Pollard 2004*a*, 2005, to appear). Except for this section, the thesis does not address the semantic part of HOG.

Hyperintensional Semantics solves several problems of Montague Semantics, particulary the granularity problem (two intuitively distinct propositions are assigned the same meaning; for example, *One plus one equals two* is assigned the same meaning as *π is irrational*), including the problem of logical omniscience (one knows all the logical consequences of what they know; for example, according to Montague Semantics *Adam knows that one plus one equals two* is true if and only if *Adam knows that π is irrational* is true). The reason is that Montague Semantics is essentially extensional and intensionality is only simulated via sets of possible worlds. Sometimes, this simulation breaks down, for example two mathematical truths, have the same extension in all worlds and thus are assigned the same meaning.

Hyperintensional Semantics is intensional from the start and therefore it does not have this problem. The whole idea can be summarized as: (i) propositions are primitive notions while worlds are derived, and (ii) two propositions entailing each other can be distinct (equivalence does not imply equality). Note that the fact that propositions are primitive notions means we are agnostic only about their formal nature not about their properties. There are relations between propositions (e.g., entailment), they can be true in a particular world, etc. This situation is similar to other formal theories. For example, while we may have some basic intuitions about sets and memberships, in Set Theory, sets are primitive, further unspecified objects without any structure and so is the membership relation between them.[8]

[8] Sets are primitive in Zermelo-Fraenkel theory (see e.g., Jech 2003), the most commonly used variant. In the von Neuman-Bernays-Gödel system (see e.g., Mendelson 1997), classes are primitive instead and sets are certain classes.

Types. HOG hyper-intensional semantics has the following basic types:

- (Hyper)intensional types:
  - Ind – individuals
  - Prop – propositions
- Extensional types:
  - Ent – extensions of individuals
  - Bool – extensions of propositions; already provided by the logic

We can also define various kinds (sets of types):

- the kind of (hyper)intensional types: $\mathsf{HYPER} = \mathsf{closeKind}(\{\mathsf{Ind}, \mathsf{Prop}\})$
- the kind of extensional types: EXT. It contains for example $\mathsf{Ent}, \mathsf{Prop} \to \mathsf{Bool}$ (see the Ext type operator below)

Propositions. As said above, the propositions are primitive notions. This is different from Montague semantics where propositions are sets of worlds. Propositions are related by entailment relation:

(34) $\models : \mathsf{Prop} \times \mathsf{Prop} \to \mathsf{Bool}$

and the induced equivalence:

(35)
$$\begin{aligned} &\equiv\, : \mathsf{Prop} \times \mathsf{Prop} \to \mathsf{Bool} \\ &\equiv\, := \lambda p, q : \mathsf{Prop}\ .\ p \models q\ \&\ q \models p \end{aligned}$$

The entailment relation is constrained by nonlogical axioms to be a preorder (i.e., reflexive, transitive, but not antisymmetric). It is crucial that it is a preorder and not a partial order (i.e., preorder + antisymmetry; as in Montague semantics). The absence of antisymmetry allows two propositions to entail each other and still be distinct objects. This means equivalence ($\Leftrightarrow$) and equality on propositions are distinct relations. Equality implies equivalence but not vice versa. Adding the axiom in (36) would turn hyperintensional semantics into Montague semantics.

(36) $\vdash \forall p, q\, .\, (p \Leftrightarrow q) \Rightarrow (p = q)$

As in Montague semantics, the set of propositions with entailment $(\mathsf{Prop}, \models)$ forms a boolean pre-order or pre-algebra.[9] Therefore, the theory introduces the usual binary operators on propositions $(\mathsf{and'}, \mathsf{or'}, \ldots)$ corresponding to the obvious natural language expressions. But again, we do not want to collapse equivalent propositions into a single object, therefore unlike in Montague semantics, the nonlogical axioms constraining these operators are formulated so that the boolean structure is a prealgebra, not an algebra (roughly, the axioms use equivalence instead of equality). For example, the two propositions in (37) (for some p) are equivalent but not equal.

(37) a. p $\mathsf{or'}$ $\mathsf{not'}(p)$

b. $\mathsf{not'}(p)$ $\mathsf{or'}$ p

Worlds and references. The type of possible worlds is not a primitive type in HOG semantics, but can be defined. Intuitively, a Montagovian possible world is a maximal consistent set (i.e., an ultrafilter) of propositions. The propositions characterize the world. It is possible to define this in HOG using subtyping:

(38) $\mathsf{World} = [x \in \mathsf{Set}(\mathsf{Prop}) \mid \mathsf{maximally\text{-}consistent}(x)]$

The predicate maximally-consistent is lambda-definable and the definition can be found in (Pollard 2005, p. 42) or (Pollard to appear).

It is also possible to define a polymorphic function mapping hyperintensions to extensions in a particular world. To assign a type to this function, we need a type operator Ext that given a hyperintensional type returns the corresponding extensional type: $\mathsf{Ext} : \mathsf{HYPER} \to \mathsf{EXT}$.[10] Then the function is defined as:

(39)
$$\begin{array}{ll}
\mathsf{ext} & : \forall H : \mathsf{HYPER} \;.\; \mathsf{World} \times H \to \mathsf{Ext}(H) \\
\mathsf{ext}(w : \mathsf{World}, * : \mathsf{Unit}) & = * : \mathsf{Unit} \\
\mathsf{ext}(w : \mathsf{World}, p : \mathsf{Prop}) & = p \in w : \mathsf{Bool} \\
\mathsf{ext}(w : \mathsf{World}, p : A \times B) & = [\mathsf{ext}(w, \pi^0(p)), \mathsf{ext}(w, \pi^1(p))] : \mathsf{Ext}(A) \times \mathsf{Ext}(B) \\
\mathsf{ext}(w : \mathsf{World}, f : A \to B) & = \lambda x : A \;.\; \mathsf{ext}(w, f(x)) : A \to \mathsf{Ext}(B)
\end{array}$$

[9] The prefix *pre* means that equivalence does not imply equality.

[10] For example, $\mathsf{Ext}(\mathsf{Prop}) = \mathsf{Bool}$, $\mathsf{Ext}(\mathsf{Ind}) = \mathsf{Ent}$, $\mathsf{Ext}(\mathsf{Prop} \to \mathsf{Prop}) = \mathsf{Prop} \to \mathsf{Bool}$; see (Pollard 2005, p. 39) or (Pollard to appear) for more details.

Montague. Finally, (Pollard 2004*b*, to appear) defines a polymorphic function modeling Montague semantics. The function takes a (hyper)intension and maps it to Montagovian intension, i.e., a function from worlds to extensions.

$$(40)\quad\begin{array}{ll}\mathsf{montague} & : \forall H : \mathsf{HYPER}\ .\ H \rightarrow (\mathsf{World} \rightarrow \mathsf{Ext}(H)) \\ \mathsf{montague} & = \lambda i : H\ \lambda w : \mathsf{World}\ .\ \mathsf{ext}(w, i) \\ \text{or equivalently} & \\ \mathsf{montague}(i : H) & = \lambda w : \mathsf{World}\ .\ \mathsf{ext}(w, i)\end{array}$$

Of course, this function simulates Montague semantics with all its problems, for example, the omniscience problem. In other words, the propositions corresponding to the sentences *Adam knows that one plus one equals two* and *Adam knows that π is irrational* are distinct in Hyperintensional Semantics, but they are equal modulo the montague function.

2.7 Lexicon

In HOG, the lexicon is a set of primitive primitive signs, i.e., words or idioms.

$$(41)\quad \vdash \mathsf{lex} : \mathsf{Set}(\mathsf{Sign})$$

To specify that there is a sign corresponding to *Fido*, it is necessary to state several things. First, we state that there is a tecto term fido of the type NP and semantic term fido' of the type Ind (existence of the term /faɪdoʊ/ follows from the constraints on phenogrammar):

$$(42)\quad\begin{array}{l}\vdash \mathsf{fido}\ \ : \mathsf{NP} \\ \vdash \mathsf{fido'} : \mathsf{Ind}\end{array}$$

And then, we need to state that the triple with such a tecto, pheno and semantics is a lexical sign:

$$(43)\quad \vdash \begin{bmatrix}\textsc{tecto} & \mathsf{fido} \\ \textsc{pheno} & /\text{faɪdoʊ}/ \\ \textsc{sem} & \mathsf{fido'}\end{bmatrix} \in \mathsf{lex}$$

Usually, we abbreviate this by simply listing the individual components of the signs, moreover writing only the phonological word for pheno. Thus a lexicon covering the sample *Fido chased a cat* would state the following:

$$(44)\quad \begin{array}{lll}
\mathsf{fido} & : \mathsf{NP}, & \text{/faɪdoʊ/}, & \mathsf{fido'} & : \mathsf{Ind} \\
\mathsf{chased} & : [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}, & \text{/tʃeɪst/}, & \mathsf{chased'} & : \mathsf{Ind} \times \mathsf{Ind} \to \mathsf{Prop} \\
\mathsf{a} & : [\textsc{spec}\ \mathsf{N}] \to \mathsf{NP}, & \text{/ə/}, & \mathsf{a'} & : \mathsf{Set(Ind)} \to \mathsf{Ind} \\
\mathsf{cat} & : \mathsf{N}, & \text{/kæt/}, & \mathsf{cat'} & : \mathsf{Set(Ind)}
\end{array}$$

A real grammar would not provide lexicon in a form of a list. Instead, we would probably introduce functions modeling morphology. Using higher-order formalism to express morphology has been studied, for example, by Forsberg and Ranta (2004). While any realistic grammar of an inflective language will require addressing morphology in some way, for our purposes it is enough to assume that there is a list of primitive signs available, whether it is simply listed or generated by some function.

2.8 Combining signs

As stated above, the set of possible signs, i.e., the set of possible combinations of pheno and tecto (and semantic) terms, is specified recursively. The lexicon introduced in the previous section specifies the basic signs. In this section, we introduce a way for the grammar to specify how signs combine into other signs.

For each term construction in any of the components of the sign, the grammar must specify how (if at all) the terms in the other components change. In some cases, the combination in one grammar component corresponds to the same combination in another component. For example, below we assume that tuples in tecto correspond to equivalent tuples in pheno. Sometimes there is no associated change, e.g., subtype embedding in tecto corresponds to no change in pheno, or the change may be more complex, as in the case of tecto function application.

In the following constraints, we make use of the following parametric type:

$$(45)\quad \mathsf{Sign}(T, P) = [\, s : \mathsf{Sign} \mid s.\textsc{tecto} :: T \,\&\, s.\textsc{pheno} :: P \,]$$

The type operator defines subtypes of the type sign, where the tecto term is of type T and the phenoterm is of type P. Often, we write the type also in the AVM notation:

(46) $\begin{bmatrix} \textsf{Sign} & \\ \textsc{tecto} & \mathsf{T} \\ \textsc{pheno} & \mathsf{P} \end{bmatrix}$

2.8.1 Tuples

We assume that tuples in tecto correspond to tuples in pheno. For example, for a binary product indexed by natural numbers, we assume the following schema holds:

(47) $\vdash \forall a : \mathsf{Sign}(A, \mathsf{Pheno})\ \forall b : \mathsf{Sign}(B, \mathsf{Pheno})\ \exists m : \mathsf{Sign}(A \times B, \mathsf{Pheno} \times \mathsf{Pheno})\,.$

$m.\textsc{tecto} = [a.\textsc{tecto}, b.\textsc{tecto}]\ \&$

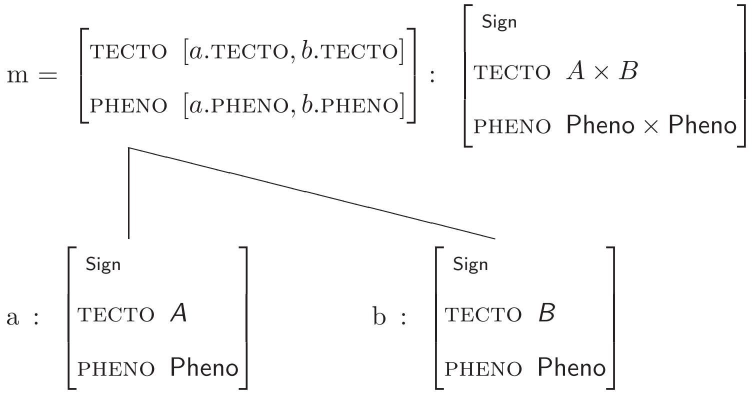
$m.\textsc{pheno} = [a.\textsc{pheno}, b.\textsc{pheno}]$

For every two signs, a and b there is a sign m such that its tecto is the pair of a's and b's tectos and its pheno is the pair of a's and b's phenos:

(48) $m = \begin{bmatrix} \textsc{tecto} & [a.\textsc{tecto}, b.\textsc{tecto}] \\ \textsc{pheno} & [a.\textsc{pheno}, b.\textsc{pheno}] \end{bmatrix}$

Informally, we can depict the axiom as a tree where the two signs a and b combine to form the sign m:

(49)


m = [TECTO [a.TECTO, b.TECTO], PHENO [a.PHENO, b.PHENO]] : [Sign, TECTO A × B, PHENO Pheno × Pheno]
a : [Sign, TECTO A, PHENO Pheno]
b : [Sign, TECTO B, PHENO Pheno]

We assume that analogous schemata exists for any set of indexes. In case of the tuples used in (9), this means:

(50) $\vdash \forall a : \mathsf{Sign}(\mathsf{NP}, \mathsf{Pheno})\ \forall b : \mathsf{Sign}(\mathsf{NP}, \mathsf{Pheno})$

$\exists m : \mathsf{Sign}([\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}], [\textsc{subj}\ \mathsf{Pheno}, \textsc{comps}\ \mathsf{Pheno}])\,.$

$m.\textsc{tecto} = [\textsc{subj}\ a.\textsc{tecto}, \textsc{comps}\ b.\textsc{tecto}]\ \&$

$m.\textsc{pheno} = [\textsc{subj}\ a.\textsc{pheno}, \textsc{comps}\ b.\textsc{pheno}]$

(51) $\vdash \forall a : \mathsf{Sign}(\mathsf{N}, \mathsf{Pheno})\ \exists m : \mathsf{Sign}([\textsc{spec}\ \mathsf{N}], [\textsc{spec}\ \mathsf{Pheno}])\,.$
$m.\textsc{tecto} = [\textsc{spec}\ a.\textsc{tecto}]\ \&$
$m.\textsc{pheno} = [\textsc{spec}\ a.\textsc{pheno}]$

In this grammar, tuples are used purely as a technical device needed by functions of multiple parameters. Linguistically, the actual combination of signs is done when there is function application in tecto. Using tuples in pheno can be thus seen as simply a way of holding the phenos of individual arguments together so that they are accessible for the "real" combination, i.e., function application.

Note, that we use the same indexes for the corresponding products in both tectogrammar and phenogrammar. One might argue that indexes like SUBJ, COMPS are motivated by syntactic functions and should not be used for indexing tuples in pheno. It would be possible to use different set of indexes for pheno, say numbers. However, formally there is not much difference. Because in our setup, there is one-to-one correspondence between tuples in tectogrammar and in phenogrammar, using the same indexes makes other constraints easier to follow.

2.8.2 Function application

The simple tecto-grammar in §2.4.3 covering the sentence *Fido chased a cat* contains two functions: the transitive verb chased and the determiner a. The combination of these functions with their arguments corresponds to certain combination of the corresponding pheno terms.

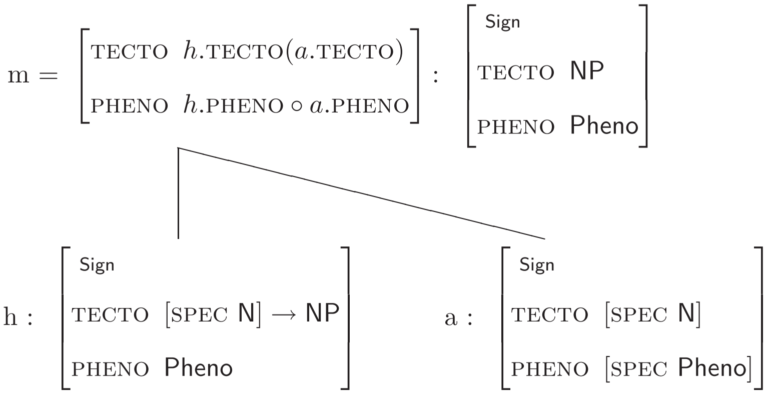
For example, applying a to cat corresponds to concatenation of the corresponding pheno terms ⟨/ə/⟩ and ⟨/kæt/⟩. Therefore a(cat) corresponds to ⟨/ə/,/kæt/⟩. In general the combination of determiners with nouns can be constrained by:

(52) $\vdash \forall h : \mathsf{Sign}([\textsc{spec}\ \mathsf{N}] \to \mathsf{NP}, \mathsf{Pheno})\ \forall a : \mathsf{Sign}([\textsc{spec}\ \mathsf{N}], [\textsc{spec}\ \mathsf{Pheno}])\ \exists m : \mathsf{Sign}(\mathsf{NP}, \mathsf{Pheno})\,.$
$m.\textsc{tecto} = (h.\textsc{tecto})(a.\textsc{tecto})\ \&$
$m.\textsc{pheno} = h.\textsc{pheno} \circ a.\textsc{pheno}.\textsc{spec}$

In prose: for every two signs h and a where tecto-grammatically, h is a determiner and a is a singleton tuple containing a noun, there is another sign m that has as its tecto the result of applying the determiner ($h.\textsc{tecto}$) on the noun ($a.\textsc{tecto}$) and as its pheno the concatenation of the pheno of the determiner and the pheno of the noun.

Displayed as a tree, this looks as follows:

(53)



m = [TECTO h.TECTO(a.TECTO) PHENO h.PHENO ∘ a.PHENO] : [Sign TECTO NP PHENO Pheno]
h : [Sign TECTO [SPEC N] → NP PHENO Pheno]
a : [Sign TECTO [SPEC N] PHENO [SPEC Pheno]]

Similarly, the combination of a transitive verb with its argument can be constrained by:

(54)
$$
\begin{array}{l}
\vdash \forall h : \mathsf{Sign}([\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}, \mathsf{Pheno}) \\
\qquad\qquad \forall a : \mathsf{Sign}([\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}], [\textsc{subj}\ \mathsf{Pheno}, \textsc{comps}\ \mathsf{Pheno}]) \\
\qquad\qquad \exists m : \mathsf{Sign}(\mathsf{S}, \mathsf{Pheno})\,. \\
\quad m.\textsc{tecto} = (h.\textsc{tecto})(a.\textsc{tecto})\ \& \\
\quad m.\textsc{pheno} = a.\textsc{pheno.subj} \circ h.\textsc{pheno} \circ a.\textsc{pheno.comps}
\end{array}
$$

The pheno component of the complex sign is again specified in the last line: it is a concatenation of the pheno of the subject, the pheno of the verb and the pheno of the object. A similar constraint would exist for intransitive verbs.

While it is possible to list individual constraints for every functor (intransitive verbs, transitive verbs, ditransitive verbs, determiners, prepositions, complementizers, . . .), in any realistic grammar, this would be rather unwieldy. Moreover, we would loose many generalizations. Therefore, in §5.2, we introduce a more convenient way of specifying possible combinations of signs corresponding to function application in tecto. Similarly, as in HPSG schemata, the constraints are then stated over triples of the three signs participating in the rule: [m,h,a] – the mother m, the head daughter h and the tuple of non-head daughters a.

2.8.3 Subtypes

A term construction in tecto need not be accompanied by any change in pheno (and vice versa). This is the case in embedding a subtype into a supertype – the embedding is transparent to the other parts of the grammar. This is what we would expect, because the type embedding functions are just technical devices required by the formal machinery of HOL with no direct linguistic motivation.[11]

2.9 A complete toy grammar of English

The toy grammar fragment of English used as illustration throughout this chapter can be summarized as:

1. Tectogrammar:

   (a) Tuple indexes. $\textsc{subj}, \textsc{comps}, \textsc{spec}$

   (b) Basic syntactic types. $\mathsf{NP}, \mathsf{N}, \mathsf{S}$

   The type of all tecto objects: $\mathsf{Tecto} := \mathsf{ClosingSupertype}(\mathsf{NP}, \mathsf{N}, \mathsf{S})$

   (c) Basic syntactic terms:

   $\vdash \mathsf{fido} : \mathsf{NP}$

   $\vdash \mathsf{chased} : [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \rightarrow \mathsf{S}$

   $\vdash \mathsf{a} : [\textsc{spec}\ \mathsf{N}] \rightarrow \mathsf{NP}$

   $\vdash \mathsf{cat} : \mathsf{N}$

2. Phenogrammar.

   (a) Type of phonemes as a primitive type: $\mathsf{Phoneme}$

   Type of phonological words $\mathsf{PhonWord} = [\, x : \mathsf{Phoneme}^* \mid \mathsf{phonotactic\text{-}constraints}(x)\,]$ (we leave $\mathsf{phonotactic\text{-}constraints}$ predicate unspecified).

   Type of pheno objects: $\mathsf{Pheno} := \mathsf{PhonWord}^*$

   (b) Individual phonemes as primitive terms:

   $\vdash$ /f/ : $\mathsf{Phoneme}$

   $\vdash$ /k/ : $\mathsf{Phoneme}$

   $\vdash$ /æ/ : $\mathsf{Phoneme}$

   $\vdash$ /ʃ/ : $\mathsf{Phoneme}$

   etc.

[11] Recall, that subtype-supertype relationship can be defined in two ways in HOG: (1) via predicate subtyping ($\mathsf{NP}_{\mathsf{acc}}$ is a subtype of NP), (2) via co-products ($\mathsf{NP} + \mathsf{PP}$ is a supertype of NP and of PP). In each case, the function mapping objects of the subtype to object of the supertype is different – see §2.2 and especially §C.4 for more details.

3. Lexicon. The lexicon (see §2.7), specifies that the following four pheno-tecto tuples are primitive signs:

$$\vdash \begin{bmatrix}\textsc{tecto} & \mathsf{fido}\\ \textsc{pheno} & \langle/\text{faɪdoʊ}/\rangle\end{bmatrix} \in \mathsf{lex}$$

$$\vdash \begin{bmatrix}\textsc{tecto} & \mathsf{chased}\\ \textsc{pheno} & \langle/\text{tʃeɪst}/\rangle\end{bmatrix} \in \mathsf{lex}$$

$$\vdash \begin{bmatrix}\textsc{tecto} & \mathsf{a}\\ \textsc{pheno} & \langle/\text{ə}/\rangle\end{bmatrix} \in \mathsf{lex}$$

$$\vdash \begin{bmatrix}\textsc{tecto} & \mathsf{cat}\\ \textsc{pheno} & \langle/\text{kæt}/\rangle\end{bmatrix} \in \mathsf{lex}$$

4. Sign combination constraints:

   (a) A determiner accepts a noun and forms a noun phrase:

$$\begin{aligned}\vdash\ \forall h &: \mathsf{Sign}([\textsc{spec}\ \mathsf{N}] \to \mathsf{NP}, \mathsf{Pheno})\\ \forall a &: \mathsf{Sign}([\textsc{spec}\ \mathsf{N}], [\textsc{spec}\ \mathsf{Pheno}])\\ \exists m &: \mathsf{Sign}(\mathsf{NP}, \mathsf{Pheno})\,.\\ m.\textsc{tecto} &= (h.\textsc{tecto})(a.\textsc{tecto})\ \&\\ m.\textsc{pheno} &= h.\textsc{pheno} \circ a.\textsc{pheno}.\textsc{spec}\end{aligned}$$

   (b) A transitive verb accepts a subject and complement NPs and forms a sentence:

$$\begin{aligned}\vdash\ \forall h &: \mathsf{Sign}([\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}, \mathsf{Pheno})\\ \forall a &: \mathsf{Sign}([\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}], [\textsc{subj}\ \mathsf{Pheno}, \textsc{comps}\ \mathsf{Pheno}])\\ \exists m &: \mathsf{Sign}(\mathsf{S}, \mathsf{Pheno})\,.\\ m.\textsc{tecto} &= (h.\textsc{tecto})(a.\textsc{tecto})\ \&\\ m.\textsc{pheno} &= a.\textsc{pheno}.\textsc{subj} \circ h.\textsc{pheno} \circ a.\textsc{pheno}.\textsc{comps}\end{aligned}$$

   (c) tuples correspond to equivalent tuples (§2.8.1)

In §2.4.3, we have show that we can prove:

(55) $\vdash \mathsf{chased}(\textsc{subj}\ \mathsf{fido}, \textsc{comps}\ \mathsf{a}(\textsc{spec}\ \mathsf{cat}))\ :\ \mathsf{S}$

which means tectogrammar licences a term $\mathsf{chased}(\textsc{subj}\ \mathsf{fido}, \textsc{comps}\ \mathsf{a}(\textsc{spec}\ \mathsf{cat}))$ of the type S. The term denotes the tectogrammatical entity corresponding to the sentence *Fido chased a cat*. The phenogrammar allows us to prove that

(56) $\vdash \langle/\text{faɪdoʊ}/,\ /\text{tʃeɪst}/,\ /\text{ə}/,\ /\text{kæt}/\rangle\ :\ \mathsf{PhonWord}^*$

The full grammar licences the sign corresponding to the above sentence:

(57) $$\vdash \begin{bmatrix} \textsc{tecto} & \mathsf{chased}(\textsc{subj}\ \mathsf{fido}, \textsc{comps}\ \mathsf{a}(\textsc{spec}\ \mathsf{cat})) \\ \textsc{pheno} & \langle /\text{faɪdoʊ}/, /\text{tʃeɪst}/, /\text{ə}/, /\text{kæt}/ \rangle \end{bmatrix} : \begin{bmatrix} \mathsf{Sign} & \\ \textsc{tecto} & \mathsf{S} \\ \textsc{pheno} & \mathsf{Pheno} \end{bmatrix}$$

The most straightforward proof of this statement is shown in Figure 2.4.

2.10 Comparison with other approaches

Below, we first discuss difference between various versions of HOG. A comparison with three type logical approaches based on tecto-pheno distinction (Lambda Grammar, Abstract Categorial Grammars, Grammatical Framework) follows.

2.10.1 "Old" HOG (Pollard 2001*b*, 2004*a*; Pollard and Hana 2003)

The main difference with the version of HOG presented in this dissertation and HOG as presented in (Pollard 2001*b*, 2004*a*; Pollard and Hana 2003), is that the older version assumes phenogrammar and semantics are related to tectogrammar by homomorphic functors. The functor is determined by its value on primitive signs as specified in the lexicon. The condition on homomorphism means that type and term constructors in tecto correspond to the same type and term constructors in semantics and pheno (e.g., function application to function application, tuple to tuple). This is simply too restrictive – it would, for example, mean that a single tecto term can correspond only to single pheno term, in other words neither lexical nor structural synonymy would be possible. Currently the relation between the three structures is much looser – it is a general relation and it is constrained by the tree types of constraints on whole signs and their possible combinations as mentioned in §2.3. In this respect, HOG in this thesis is closer to HPSG's treatment of the relation between phonology, syntax, and semantics. They are mutually constrained, not one function of the other.

2.10.2 HOG in (Pollard 2006)

The grammar of English in HOG as presented in (Pollard 2006) differs from the framework in this thesis in several ways (apart from minor notational differences). The most important one is that Pollard uses several different functional type constructors, each for a different type of complements ($\multimap_{\textsc{subj}}$, $\multimap_{\textsc{comp}}$, $\multimap_{\textsc{spec}}$). This means the logic of types has several implications. Each of the functional types has its own application and abstraction. We handle the same differences with one

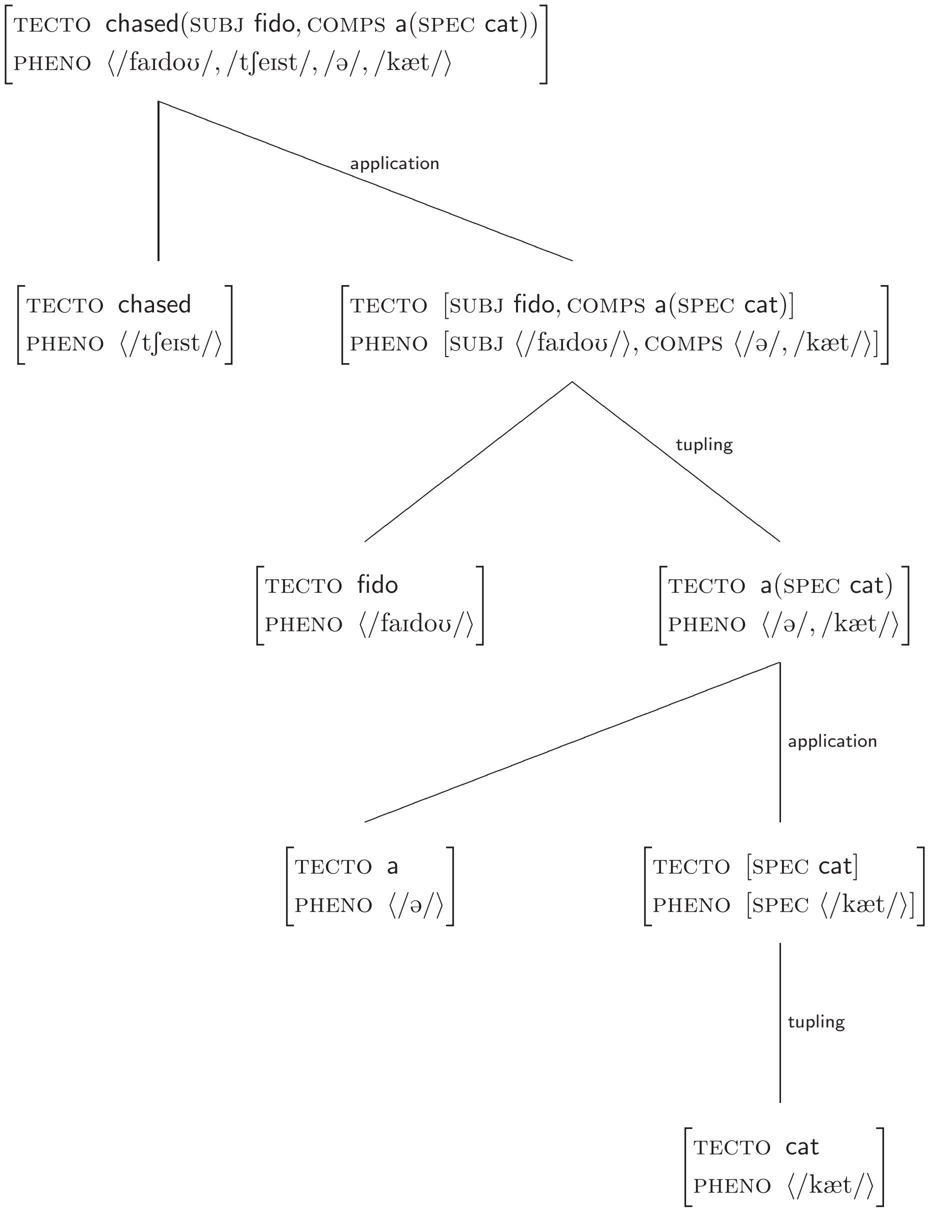


Figure 2.4: Parallel tecto and pheno derivation corresponding to *Fido chased a cat*

functional type and indexed products. Thus the tecto type of an transitive verb chase in (Pollard 2006) would be

(58) $\textsf{chase} : \textsf{NP} \multimap_{\textsc{comps}} (\textsf{NP} \multimap_{\textsc{subj}} \textsf{S})$

instead of the term used in this thesis:

(59) $\textsf{chase} : [\textsc{subj}\ \textsf{NP}, \textsc{comps}\ \textsf{NP}] \rightarrow \textsf{S}$

Note, that the different implications of (Pollard 2006) are not intended to correspond to different word orders (at least not directly), as different implications do in categorial grammar (e.g., Morrill 1994; Steedman 2000*b*), but rather to different grammatical functions.

2.10.3 HPSG

HPSG (Pollard and Sag 1994) is one of the most commonly used syntactic formalisms, and probably the most frequently used grammar formalism in computational linguistics. Formally, the main difference between HOG and HPSG is that functions (and relations) in HOG are first class citizens. For example, they can be passed to other functions as arguments. Another difference is that in HOG, syntax (tectogrammar) and semantics are clearly separated, while in HPSG they share the same structures.

HPSG has been successfully applied to many linguistic problems. However, formalization of grammars of languages with free-phrase order and relatively common discontinuous phrases, although theoretically possible, is complicated, nonmodular, and nonintuitive as Penn (1999*b*) or Rosen (2001) show. Penn formulates his analysis of Serbo-Croatian clitics both in a precise natural-language and in HPSG. The contrast between the two analyses is striking – the former is elegant and simple, while the latter is rather complicated and non-modular. Rosen (p.c.) draws a similar conclusion from his formalization of Czech word order.[12]

HOG and HPSG have the same theoretical expressive power: both lambda calculus (Szudzik 2005) and HPSG (Kepser 2004; Søgaard 2007,Keselj 2002, p. 118) are Turing equivalent. However, HOG is based on higher order logic, while HPSG is based on RSRL (Relational Speciate Re-entrant Language; Richter 2000). The former is a standard, well researched formalism, widely used in mathematics, computer science and other areas. The latter is an idiosyncratic formalism, unknown

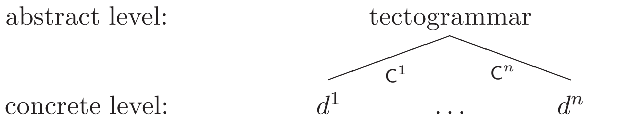
[12] Rosen does not use standard HPSG, but he uses RSRL (Richter 2000), the underlying logic of HPSG to formalize Functional Generative Description (FGD; Sgall et al. 1986)

to anybody except a small group of formal linguists. For a more extensive discussion of HPSG problems, both formal and practical, see Pollard (2001*a*). Here, we limit ourselves to mentioning a very surprising fact that even finite models are undecidable in RSRL (Kepser 2004)!

2.10.4 Lambda grammars

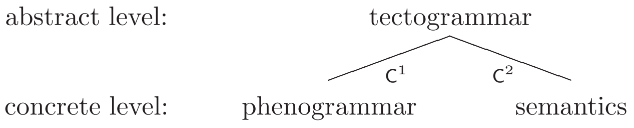
General setup. Lambda-grammars (Muskens 2004; also Muskens 2001*a*,*b*, 2003) have two levels – an abstract one and a concrete one – see the graph in (60). The abstract level contains a single structure, called tectogrammar, responsible for combinatorics of expressions. Tectogrammar is realized in n concrete structures (the dimensions of the concrete level).[13]

(60)



Usually, just two dimensions are assumed: phenogrammar (called syntax, dimension 1) and semantics (dimension 2):

(61)



The objects of the grammar are signs, n-dimensional tuples $[a_1, \ldots, a_n]$. The whole tuple is a tectogrammatical entity assigned a tectogrammatical type. Each of the terms a_i is an entity of the concrete grammar of the i-th dimension. Thus usually, a sign is a pair of the corresponding phenogrammatical and semantic terms. For example, [/faɪdoʊ/, fido'] : NP. This means that $\lambda-$grammars are very much like mainstream Categorial Grammar. The difference is that unlike in CG, tecto terms are non-directional.

Essentially, $\lambda-$grammars can be viewed as a certain collection of Abstract Categorial Grammars (ACG; de Groote 2001, 2002): each abstract-concrete pair corresponds to one ACG. Unlike $\lambda-$grammars, ACG's can also be chained when the concrete grammar of one level is the abstract grammar of the next level. However, formally this difference is not important, because for every such a chain of grammars, there is a single equivalent ACG grammar.

[13] We modify the notation slightly to make the same concepts written in the same way as in HOG. For example, a functional type is written as $A \rightarrow B$ instead of (AB), Bool is used instead of t.

Logics. An n-dimensional λ-grammar consists of n + 1 logical theories. Tectogrammar uses the implicative ($\multimap$) fragment of linear logic.[14] The concrete grammars can use various logics, but Muskens uses multimodal logic. In HOG (as presented in this thesis), all theories, including tectogrammar, are theories in non-linear Higher Order logic with products and coproducts.

This means λ-grammar splits Multimodal Categorial Grammar (MCG; Moortgat 1997) into two parts: the combinatorial part in tectogrammar and the multimodal part in phenogrammar.

Mappings. The tectogrammar is realized into each dimension via concretization functions, marked as C^d. These functions are homomorphisms, therefore, it is enough to specify them for lexical entries only. A tuple of concretized tectogrammatic expressions is called a generated sign. For example, if cat is a tectogrammatic expression, $\langle \mathsf{C}^1(\mathsf{cat}), \mathsf{C}^2(\mathsf{cat}) \rangle$ is a generated sign.

This setup is very similar to older versions of HOG, C^1 corresponds to the interpretation homomorphism phen and C^2 to sem. However, as mentioned in §2.10.1, in the current HOG, the relation between tectogrammar and phenogrammar on one side and semantics on the other is much looser. They are still derived compositionally, but phen (and sem) are relations, not function and even less homomorphisms.

Advantages. Such a multistratal setup has formal and practical advantages: (1) the framework is modular, allowing each level and dimension to be studied to a great extent separately; (2) the actual grammars are much simpler than corresponding grammars in MCG.

The setup also has linguistic advantages. As an example, Muskens (2004) presents a treatment of medial gaps (*the book that Sue gave _ to Bill*) and certain quantifier expressions. In MCG, the technique for analyzing medial gaps is much more complicated than that for peripheral gaps. The complications are probably just an artifactual consequence of the formal toolkit, not of some real property of natural language. In λ-grammars the treatments of medial and peripheral gaps are equally complex.

[14] Roughly: the implicative fragment means that there are no type constructors other than → (e.g., no products/tuples or coproducts); the linear logic means that every lambda binds exactly one variable, thus $\lambda x \,.\, x + x$ is impossible, because λ binds two instances of x.

Phenogrammar. Abstract tecto terms correspond to *sets* of phrase structure trees. Although the PS trees are objects of the phenogrammar, their combinatorial admissibility is handled in tectogrammar. Within phenogrammar, the trees are (possibly) transformed and translated to strings using the usual apparatus of multimodal grammars Moortgat (1997).[15]

The set of modal operators is quite standard:

1. $\bullet$ – constituency; separates sisters in a binary branching tree;
2. $\circ$ – string concatenation (as HOG $\circ$ concatenation);
3. $\Diamond_y$ – tree-to-string conversion (yield), compacts the part of tree in the Reape (1994)/Kathol (1995) sense; $\Diamond_y(a \bullet b) = \Diamond_y a \circ \Diamond_y b$ and $\Diamond_y a = a$ if a is lexical.
4. $\Box_m$, $\Diamond_m$ – tools for the usual key-lock gymnastics of a multimodal-grammar.

For example, the following pair would correspond to the transitive verb *kiss* (terms and types are written in HOG notation):

(62) $[\lambda t_1, t_2 \,.\, (t_2 \bullet (\mathit{kisses} \bullet t_1)), \lambda x, y : \mathsf{Ent} \,.\, \mathsf{kiss'}yx] : \mathsf{NP} \rightarrow (\mathsf{NP} \rightarrow \mathsf{S})$

and one of the two signs corresponding to the sentence *Every boy kisses a girl* would be:

(63) $[((\mathit{every}\bullet\mathit{boy})\bullet(\mathit{kisses}\bullet(a\bullet\mathit{girl}))), \lambda i : \mathsf{World}\ \forall x : \mathsf{Ent} \,.\, (\mathsf{boy'}xi\Rightarrow\exists y : \mathsf{Ent}(\mathsf{girl'}yi\&\mathsf{kiss'}xyi))] : \mathsf{S}$

the first part of the tuple represents a binary branching tree. Applying the yield operator ($\Diamond_y$) recursively on it produces the expected result

(64) $[\mathit{every} \circ \mathit{boy} \circ \mathit{kisses} \circ a \circ \mathit{girl}, \lambda i : \mathsf{World}\ \forall x : \mathsf{Ent} \,.\, (\mathsf{boy'}xi \Rightarrow \exists y : \mathsf{Ent}(\mathsf{girl'}yi \,\&\, \mathsf{kiss'}xyi))] : \mathsf{S}$

Comparison with HOG In many aspects, Lambda Grammar is very similar to HOG. However, there are also many differences. In Lambda Grammar, there are no tecto terms per se, instead they are the same entities as signs. The underlying logic is also different. While HOG uses the full (intuitionistic) HOL for tectogrammar, lambda-grammars use only the implicative fragment of linear logic. Recently, Muskens (p.c.) suggested dropping the requirement of linearity and require only non-vacuousness of lambda bindings (for parsability and learnability reasons).

[15] For this it is important that the pheno objects are *sets* of trees and not just trees. The sets have the obvious boolean structure ($\subseteq$, $\cup$, $\cap$, full set, $\emptyset$). For example, the constituency $\bullet$ operator does not combine two subtrees into a tree, but two sets of subtrees into a set of trees.

Moreover, Lambda Grammar follows the tradition of Categorial Grammar by being extremely lexicalist – there are no nonlogical axioms except those expressing the lexicon. Therefore, unlike HOG, Lambda Grammar cannot express HPSG-like constraints.

2.10.5 Grammatical Framework (GF)

Although Grammatical Framework (Ranta 2004) is in many respects similar to Lambda-grammars, ACG and HOG, its development has been motivated and driven by different goals. It was developed as part of various practically oriented projects (multilingual authoring, translation of proofs into NL, etc). As a consequence, it is formally less general than Lambda-grammars, but very elaborated in particular aspects (modularity and reuse support, morphology). It is also fully implemented (parser, generator, multilingual authoring, etc.)

In GF, the tectogrammar is in a typed lambda calculus with only meta equality, thus it does not form a logic. This means it is not possible to constrain the possible tecto expressions. Similarly as in HOG, and unlike in Lambda-grammars, similarly as in HOG, the tecto lambda calculus is not linear.

CHAPTER 3

BASICS OF CZECH WORD ORDER

In this section, we describe and analyze the basic properties of Czech word order. First, we discuss word order in Czech in general. After that, we summarize the relation of word order and Information Structure. Then we briefly mention some elements with syntactically determined word order: prepositions and complementizers. Finally, we provide a slightly less elementary analysis of topicalization or fronting. In Chapter §5 these properties are analyzed within HOG.

A complete analysis of Czech word order phenomena is well beyond the scope of this thesis; below we present only the basic properties. We also leave out many other phenomena relevant to Czech word order, notably the so-called wh-movement, comparatives and parentheticals. Clitics are discussed to a significantly greater depth in Chapter 4.

First a short note about examples. Appendix B discusses the presentation of data and their sources in more detail. I have tried to avoid constructing my own examples; instead I have used as many real utterances as possible – usually drawing them from various subcorpora of the Czech National Corpus (CNC) or the Prague Dependency Treebank (PDT). The Czech National Corpus includes two synchronic spoken corpora containing fiction, non-fiction and news (syn2000 abbreviated as syn0, syn2005/syn5), two spoken corpora (Oral2006, PMK), a corpus of private correspondence (KSK), news corpora (syn2006pub/syn6) and a few others. Any example that does not have a source listed is based on my own Czech native competence. Searching the corpora for evidence for a particular phenomenon is often far from trivial (c.f. e.g. Meurers 2005). While most of the corpora used are annotated with morphological and PDT also with syntactic information, the morphological annotation was mostly automatic and obviously is not perfect. The nature of current tagging technology means that errors are more common in less frequent constructions and especially in constructions involving discontinuities, both of concern in this thesis.

In the examples, information structure is marked in the English translation: Rheme is marked by the use of capitals and subscript R and contrast in theme by sans-serif and subscript C.

Finally, it is necessary to mention that there are two variants of Czech (see §A for more details): Official (Literary, Standard) Czech and Common (Colloquial) Czech. The two variants differ mainly in morphology and lexicon. One might argue that there are no native speakers of Official Czech. However, in the area of clitics, the grammatical differences are quite limited, and we discuss them where they arise. Simplifying somewhat, the spoken corpora can be seen as capturing Common Czech, and the written corpora, especially the news texts, as capturing Official Czech. The KSK corpus of private correspondence mixes features of both, sometimes even within the same sentence.

3.1 Free word order

Czech has exceptionally free word order in comparison with many other languages in general, and with English in particular. Unlike English, where word order is mostly fixed and is mainly used to express grammatical functions, word order in Czech is used to express Information Structure (see the next subsection).[16] Thus for example, the four words in sentence (1) can be rearranged in all

[16] And probably also definiteness, as in Russian, another Slavic language (Brun 2000, 2001).

24 (=4!) possible ways. Each of the sentences has a different information structure, but all of them are grammatically correct.

(1) Včera Petr viděl Marii.
yesterday Peter$_N$ saw Mary$_A$
'Yesterday, Peter saw Mary.'

More precisely, Czech word order is very free with respect to the possibility of moving entire phrases; virtually any scrambling is possible. However, scrambling resulting in discontinuous phrases is much less common, although it is far more common than in English.[17] It is mostly limited to discontinuities due to certain constructions (e.g., comparison), to clitics (see §4) and to sentences involving so-called *split fronting* (see §3.4). One of the first more systematic survey of discontinuous constructions in Czech can be found in (Uhlířová 1972). (Holan et al. 1998, 2000; Plátek et al. 2001) have suggested several measures expressing complexity of discontinuities and their reflection in the complexity of parsers.

Discontinuities in the Prague Dependency Treebank. Discontinuities in the surface syntax layer of the Prague Dependency Treebank (PDT; see §B) have been analyzed by (Hajičová et al. 2004; Zeman 2004). They report that about 23% of the 73,000 sentences contain some kind of nonprojectivity (roughly, discontinuity).[18] However, many of the discontinuities are of a rather technical nature (many involve punctuation that is included as part of the syntactic structure) or are theory-dependent (e.g., they involve structures that could be analyzed as coordination of elliptical clauses, non-constituent coordination, gapping, etc., where only some of such analyses involve discontinuities). Finally note that PDT is a news corpus; the number and distribution of discontinuities in spoken and/or informal language are likely to be significantly different.

[17] According to (Holan et al. 2000), English allows a maximum of three discontinuity gaps in a phrase, while Czech does not impose any limit on the number of gaps. Of course, this is the competence point of view; the performance point of view is quite different – in a way parallel to, for example, relative-clause embedding which is also unlimited in competence but rather restricted in performance.

[18] Projectivity is defined on dependency trees. A dependency tree is a rooted ordered tree where the nodes are the words (tokens) of the sentence. In a dependency tree, the head word dominates its dependents (i.e., there is no distinction between a mother and its head daughter).

A dependency edge between a daughter d and mother m is projective iff all nodes that are between d and m in the word-order relation, are transitively dominated by m. A dependency tree is projective if all edges are projective, otherwise it is nonprojective. Various measures of degrees of non-projectivity have been explored, for example in (Havelka 2007).

3.2 Elements with restricted word order

While Information Structure (together with phrases embedding, see below) is the main factor determining word order in Czech, there are elements with fixed or highly restricted word order. In this section, we address prepositions and complementizers. Clitics, other set of elements with a restricted placement, are discussed in Chapter 4.

Prepositions. Prepositions immediately precede their NPs, as shown by *o* 'for' in (2a,c). There is no preposition stranding in Czech, as (2c) illustrates.

(2) a. Požádali jsme je [o krátký rozhovor].
asked aux$_{1pl}$ them$_A$ for short interview
'We asked them for a short interview.' [syn6]

b. O co jste je požádali?
for what aux$_{1pl}$ them$_A$ asked
'For what did you ask them?'

c. * Co jste je požádali o?
what aux$_{1pl}$ them$_A$ asked for
Intended: 'What did you ask them for?'

Complementizers. Complementizers precede the clause as illustrated by *že* 'that' in (3):

(3) Doufám, že [ses tam nenudila].
hope$_{1sg}$ that aux$_{2sg}$+refl$_A$ there not-bored
'I hope you weren't bored there.' [ksk]

3.3 Information structure and Information Packaging

There is general agreement that different parts of an utterance make different informational contributions to the discourse. An utterance can be divided into two parts according to the informational contribution it makes. The new information communicated by the utterance is expressed by the part usually called *rheme* (e.g., in Firbas 1957; Steedman 2000*a*) or *focus* (e.g., in Sgall et al. 1986). On the other hand, the part usually called *theme* or *topic* connects *rheme* to the information already present in the common ground.[19] Informally, one might say rheme is what the utterance says about

[19] Note that these terms are in some theories used differently. For example Steedman (2000*a*) uses *focus* to refer to contrast (both in theme and rheme). The term *topic* is sometimes used as synonymous to theme (e.g., Sgall et al.

the theme. Although there is some agreement about these basic properties of theme and rheme, anything beyond the intuitive characterization is controversial, including the exact nature of those items, their manifestation, existence of transitional items, etc. In the words of Enric Vallduví:

> A number of proposals for the informational articulation of the sentence – sometimes incompatible – are found in the literature. The differences among them are significant [..]. What all the approaches have in common is the recognition that in the sentence there is some sort of informational split between a more informative part and a less informative part. Where that split is and what kind of split it is – a continuum or a dichotomy – is a a matter of disagreement, but the split is nevertheless present. In our terms, it could be said that information is concentrated on a subpart of the sentence, while the remainder is licensed only as an anchoring vehicular frame for that informative part to guarantee an optimal entry into the hearer's knowledge-store. (Vallduví 1993, p. 35).

The distinction was perhaps first suggested by Weil (1844).[20] Gabelentz (e.g., Gabelentz 1891) distinguished *psychological subject* (roughly theme) and *psychological predicate* (roughly rheme). In the Prague school, Information Structure has been studied extensively by Mathesius (1915, 1929, 1939), Firbas (1957, 1992, etc.; using the term Functional Sentence Perspective), Daneš (1974) and Sgall & Hajičová (Sgall et al. 1986, etc.; Topic-Focus Articulation). The Prague School's main concern has been relation of the Information Structure to word order. The work by (Halliday 1967) is probably responsible for bringing the ideas to Generative Syntax (Jackendoff 1972; Selkirk 1984, and many others).

In this thesis, we treat Information Structure along the lines of Functional Generative Description (hence FGD; e.g., Sgall et al. 1986). The decision is primarily a pragmatic one; most of the empirical work on the Information Structure in Czech has been done in FGD or theories closely related. No other theory has been tested so extensively on Czech data. For example, in the Prague Dependency Treebank (see §B.1), about 50,000 have been manually annotated for Information Structure.[21] The

1986), sometimes only as its contrastive part. Finally, *comment* is complimentary to *topic* in either of these meaning, so sometimes it is synonymous with rheme and sometimes refer to the part of the sentence that is not contrastive theme. See (Vallduví 1993, §3.1) for a comparison of terminology.

[20] He calls *initial notion* or *point of departure* what we would call theme and *information being imparted* or *goal of discourse* what we would call rheme (Weil 1887 [1844], p. 30); he even suggests that what Latin expresses by word order, English expresses with emphasis (p.49 Note 7).

[21] To be precise, lexemes in the tectogrammatical layer of PDT are annotated for contextual boundness (see below). Information about theme and rheme can be derived from such annotation. On the portions that were processed by

theory differs from other theories in many important aspects; however at the level of detail needed here, it is largely compatible with many other treatments of Information Structure.

For some researchers, the terms theme and rheme refer to pragmatic or cognitive categories. While we do not dispute that such categories exist, we use the terms to refer to their syntactic counterparts (similarly as tense is related to time, aspect to Aktionsart, etc.).

3.3.1 Theme – Rheme

Following FGD, but also the general treatment of Information Structure in Czech syntax (e.g., Daneš et al. 1987), we partition the words in the tecto-structure of an utterance into *theme* and *rheme*.[22] Theme and rheme are syntactic (tectogrammatical) categories that have cognitive/pragmatic counterparts and are expressed by various means, primarily by intonation and word order (see below). As examples, consider the four sentences from (Vallduví and Vilkuna 1998) in (4) (rhemes are marked by capitals).

(4) a. What about pipes? In what condition are they?
The pipes are RUSTY.$_R$

b. What about pipes? What's wrong with them?
The pipes ARE RUSTY.$_R$

c. Why does the water from the tap come brown?
THE PIPES ARE RUSTY.$_R$

d. I have some rust remover. You have any rusty things?
THE PIPES$_R$ are rusty.

Theme is the syntactic counterpart of being *given by the Question Under Discussion* (Roberts 1996), and rheme is the syntactic counterpart of *Information Focus* (Roberts 1998), which provides a (partial) answer to the question under discussion. In fact, FGD uses the so-called Question Test to identify focus (e.g., Sgall et al. 1986, §3.31). The difference is that in case of FGD, the questions are

several annotators, the agreement on tokens is about 76% (3 annotators, about 10,000 sentences) or about 68% (6 annotators, about 900 sentences) (Zikánová et al. 2007).

[22] FGD usually uses the term *topic* for *theme* and *rheme* for *focus*. We chose theme and rheme because they seem to be less ambiguous across theories.

Also, including only words in the Information Structure is a simplification. FGD distinguishes topic/focus also for grammatical morphemes. For example, a past tense morpheme can belong to focus, while the verb itself belongs to topic, event though they are realized as a single word (at least in 3rd person).

just tests, while, in Roberts' theory, the questions under discussion are abstract entities modeling the discourse. Similarly FGD's themes and rhemes are very similar to themes and rhemes of Vallduví (e.g., Vallduví and Vilkuna 1998) or Steedman (e.g., Steedman 2000*a*).

Although theme and rheme are related to old (familiar) and new information, they are still syntactic notions – they express how the speaker *decides* to modulate the information. Theme does not necessarily need to be old information. As Roberts (1996, p. 19) shows, theme can be used to communicate new information via presuppositions it triggers. On the other hand, rheme does not necessarily need to present new information – consider, for example, the dialog in (5) between a student and a professor. In an ideal situation, both know the answers, thus the rheme of the student's answer does not add to the common ground any information about Kepler discovering how planets work, but rather that the student knows the answer, is able to present it in an appropriate way, etc.

(5) Professor: What did Johannes Kepler discover while in Prague?

Student: He discovered TWO OF HIS PLANETARY MOTION LAWS$_R$.

3.3.2 Contrast

In addition to the theme-rheme distinction, it is common to distinguish between contrastive and noncontrastive elements. Consider the following dialog from (Jackendoff 1972):

(6) a. Well, what about FRED? What did HE eat?

b.	FRED	ate the	BEANS.
accent:	fall-rise (B)		fall (A)
focus:	independent		dependent

(7) a. Well, what about the BEANS? Who ate THEM?

b.	FRED	ate the	BEANS.
accent:	fall (A)		fall-rise (B)
focus:	dependent		independent

The fall accent (Jackendoff's *A-accent*) marks what Jackendoff calls *dependent focus*, and fall-rise accent (*B-accent*) marks *independent focus.* The difference is that (6b) cannot occur in the context of (7a) and (7b) cannot occur in the context of (6a). In Czech, the same distinction would be usually expressed by word order, with an optional fall-rise accent on the independent focus and fall accent on the dependent focus:

(8) a. A co Fred? Co (ten) jedl?
And what Fred? What (that) ate?
'And about what Fred? What did he eat?'

b. Fred jedl fazole.
Fred ate beans
'Fred ate the beans'

(9) a. A co fazole? Ty jedl kdo?
And what beans? Those ate who?
'And what about the beans? Who ate them?'

b. Fazole jedl Fred.
beans ate Fred
'Fred ate the beans'

For some researchers, the presence of contrasts implies the unit is rhematic. For example, according to Roberts (p.c.), both independent focus and dependent focus are rhematic. On the other hand, for some researchers contrast is orthogonal to the theme-rheme distinction, so parts of both theme and rheme can be contrasted. In such a view, Jackendoff's independent focus is usually considered thematic and dependent focus rhematic. This is true, for example, for Steedman (1991, contrast is called *focus*), Vallduví and Vilkuna (1998, *kontrast*), and probably also Kadmon (2001, *TOPIC-focus* = contrastive theme, *FOCUS-focus* = contrastive rheme).

FGD falls roughly into the latter group. Contrast is independent of the theme-rheme distinction, so there is a contrastive and noncontrastive theme (usually called contrastive and noncontrastive topic). The distinction between contrastive and noncontrastive rheme is not made for Czech. According to Sgall (p.c.), the reason is that while the distinction is cognitively relevant, it has no linguistic manifestation in Czech.

3.3.3 Theme Proper, Rheme Proper

According to FGD, in addition to the simple distinction of theme and rheme, there is a more fine-grained distinction of so called deep word order, a linear order expressing increasing communicative load (so-called *communicative dynamism*) of items in the utterance. Items in the theme come before items in the rheme in such ordering. Within the theme, the order of items reflects the items' decreasing salience (see Hajičová and Vrbová 1982; Hajičová et al. 1990). The minimal item in such ordering, the most salient item, i.e., the most "thematic"-theme, is called *Theme Proper* (*Topic Proper*) and the most "rhematic"-rheme is called *Rheme Proper* (*Focus Proper*).

Theme proper and rheme proper usually correspond to individual clausal constituents, but there are exceptions. It is well known that they may correspond to a partial constituent, see for example (10) (the English translation corresponds to one of several possible interpretations, see §3.3.4.2.)

(10) [Sportovec] je Pavel [dobrý].
sportsman$_{m.sg.nom}$ is Pavel good$_{m.sg.nom}$
'Pavel is a GOOD$_R$ sportsman$_C$.' (As a sportsman, ...)

Although not discussed in the literature, in the light of multiple constituents involved in long-fronting (§3.4.3) and multiple constituents preceding clitics (§4.4.4), it seems reasonable to suggest that under certain circumstances a theme proper may consist of several constituents, or at least things that are traditionally regarded as multiple constituents.

In addition to the utterance level theme-rheme (topic-focus) dichotomy, the FGD theory of Information Structure distinguishes so-called contextually bound and contextually unbound elements (e.g., Sgall et al. 1986); they are primitive notions, but in a prototypical case, *context bound* corresponds to a contextually *given/familiar* and *context unbound* to a *new* expression. Neither of these notions is used in this thesis.

3.3.4 Information packaging

Different languages mark Information Structure in different ways. Distinct intonation and word order are the most common means in most languages, including Czech. In Czech, as a free word order language, the function of word order in expressing information structure is far more important than in languages like English.

3.3.4.1 Intonation

Until recently, relatively little attention was devoted to Czech prosody. Most of the statements about prosody are rather vague, with little or no grounding in exact phonetic experiments. The prosodic marking of rheme proper is usually called *intonation center* while contrastive theme is simply marked by *contrastive stress*, corresponding to Jackendoff's B-accent.

According to Nino Peterek (p.c.), preliminary results suggest that contrastive theme is marked by a rising tune, but it is unclear whether it corresponds to something like L+H* or H*, or even L+H* L of the ToBI system developed for English (Silverman et al. 1992). Rheme has a falling tune; when positioned sentence finally, it is marked simply by L%. For discussion of various realizations of contrastive themes, see for example (Veselá et al. 2003).

3.3.4.2 Word order

Objective ordering Usually, sentences follow so-called *objective ordering* (Mathesius 1939, 1975). In that case, according to FGD:

1. The Intonation Center (the tune marking rheme proper) is at the end.
2. Thematic expressions precede rhematic expressions; contrastive theme tend to come before non-contrastive theme:

   Theme Proper < other Theme < other Rheme < Rheme Proper

3. The order within the theme is constrained by salience, more salient items coming first (see, e.g., Hajičová and Vrbová 1982; Hajičová et al. 1990).
4. Rhematic expressions are usually ordered by a default word order, the so-called *systemic ordering* (Sgall et al. 1995).

The traditional and most straightforward way to interpret this is to see word order in Czech as the means of expressing theme and rheme. Thus Weil's statement that the "syntactic march is not the march of ideas" (Weil 1887 [1844], p. 21) is more true of English than of Czech.[23]

There are many exceptions to this general pattern; see (Rosen 2001) for a summary. For example, word order in certain syntactic constructions is usually fixed regardless of IS (e.g., there is a strong preference for adjectives to precede their nouns); the finite verb occurs in the second position also more frequently than would be predicted by its IS function (this is probably an influence of German); as in many other languages, heaviness of constituents influences their placement; etc. Also, constituents with heterogenous IS (e.g., adjective belongs to rheme, noun belongs to theme) tend to stay continuous. However, the constituent might be split, especially if one part belongs to Theme Proper and the other to Rheme Proper. This is discussed in more detail in the following section.

[23] One might say that in English, word order is relatively fixed and prosody is relatively free, while in Czech it is just the opposite. However, it is also possible to see the situation from a different perspective, along the lines suggested by (Roberts 1998, p. 146). In that view, word order in Czech would not express Information-Structure per se, but instead is only responsible for placing the rheme into the position where the Intonation Center can be realized. In our view, the problem with such a view is that (1) the IC can be under certain circumstances placed sentence non-finally (see below) and (2) the ordering within the theme by item salience would need to be considered a different phenomenon.

Subjective ordering In addition to the general objective ordering principle, there is a so-called *subjective ordering* (Mathesius 1939, 1975).[24] In this ordering, the Rheme Proper is placed at the beginning:

(11) Rheme Proper < Theme Proper < other Theme < other Rheme

Subjective ordering is usually used in excited speech; it is also quite common in newspapers, especially in titles (it probably adds some flavor of speed, urgency, etc.).

In addition to this simple case, there are also intermediate orders where a bigger portion or even the whole rheme is placed sentence initially. According to L. Uhlířová (p.c.) there is no systematic study on subjective ordering. We are therefore forced to leave this for future research and assume only the simplest possibility when only Rheme Proper occurs clause initially.

3.3.4.3 Analysis of Information Packaging

In the following sections and chapters we will the following reflection of Information Structure in word order.

Sentences having two parts:

1. The first part contains the theme proper (if there is any) in objective ordering and the rheme proper in subjective ordering. We will call such expression a *fronted* expression and analyze it in more detail in the following section.

2. Following the fronted expression is the rest of the sentence and it is ordered according to the increasing communicative dynamism:

   Theme Proper < other Theme < other Rheme < Rheme Proper

   Note that not all items must be present in this part of the sentence. A particular element may not present at all (only Rheme Proper is obligatory) or it could have been fronted.

[24] In Weil (1887 [1844], pp. 43–47) the term *the pathetic order* refers to a similar phenomenon in Greek:

> When the imagination is vividly impressed, or when the sensibilities of the soul are deeply stirred, the speakers enters into the matter of the discourse at the goal, and we do not become aware, till afterward, of the successive steps by which he could have entered had his mind been in a more tranquil state. (Weil 1887 [1844], p. 45)

This gives us the following two orders:

	fronted	rest of the sentence
objective:	Theme Proper	other Theme < other Rheme < Rheme Proper
subjective:	Rheme Proper	Theme Proper < other Theme < other Rheme

We assume that there are sentences without a fronted element. For example, the response in (12) is a rheme-only sentence in objective ordering. The first constituent is neither theme proper neither rheme proper, and we assume that it was not fronted.

(12) *Context: Proč máš takovou radost? – Why are you so happy?*

Martin odjel do Francie.
Martin went to France.
'Martin went to France.'

3.3.5 Summary of the adopted Information Structure for Czech

In the following, we assume the following basic view of Information Structure and Information Packaging in Czech. It is clear that more research is needed in this area.

1. Nature:
   (a) Every sentence is partitioned into theme and rheme. The rheme must be nonempty.
   (b) The most thematic/salient part of the theme is theme proper, the most rhematic part of the rheme is rheme proper.
   (c) Every item in the theme is either contrastive or noncontrastive.
   (d) Contrast is not linguistically distinguished for rheme (rheme proper might but need not express contrast).
2. Realization:
   (a) The word order reflects the IS of an utterance, either by objective ordering or subjective ordering. If there is a contrast in the theme, it tends to be on the theme proper.
   (b) The objective and subjective ordering differ in the nature of their initial (fronted, see next section) element: in the objective ordering it is the theme proper (if there is any), while in the subjective it is the rheme proper.

(c) The rest of the sentence is ordered according to the following order:

Theme Proper < other Theme < other Rheme < Rheme Proper

(d) Constituents with heterogenous IS tend to stay continuous; however there are exceptions. For example, as discussed in the next section, even partial constituent can under certain circumstances undergo fronting.

(e) Prosodically, the rheme proper is marked by the so-called Intonation Center. The contrastive theme is marked by a falling-rising tone, which is optional if the contrastive theme is sentence initial.

(f) Some expressions (e.g., complementizers or clitics) are not ordered by IS

3.4 Fronting

In this section, we will explore the basic properties of a phenomenon usually called fronting or topicalization. We avoid the term topicalization, because this suggests the construction marks an expressions as a topic (whether that means theme or only contrastive theme); which is true only in objective ordering. In subjective ordering, the fronted expression is rhematic.

In comparison with English or German, many aspects of Czech fronting are rather understudied. This applies mostly to so-called long fronting (where the expression occurs in a higher clause) and split fronting (where only part of a clausal constituent is fronted). Given the complexity and diversity of constraints on split and long fronting in other languages, it is unlikely that Czech would be significantly simpler in this area, yet these phenomena have been little discussed for Czech.

3.4.1 Short Constituent fronting – scrambling

As discussed in §3.3 above, theme proper (contrastive or not) and, in subjective ordering, rheme proper tend to occur sentence initially. For clausal constituents, this tendency is close to a strict rule. We analyze their presence in initial position, e.g., *housky* ‘rolls’ in (13) as simply a result of ordering the clausal constituents.

(13) a. *Objective ordering:*
Context: Kdo koupil housky? – Who bought the rolls?

Housky koupil Martin.
rolls bought Martin.
‘Martin$_R$ bought the rolls.’

b. *Subjective ordering:*

Context: Co koupil Martin? – What did Martin buy?

Housky	koupil	Martin.
rolls	bought	Martin.

‘Martin bought THE ROLLS$_R$.’

3.4.2 Split fronting

The situation when the theme or rheme proper correspond to only a part of a clausal constituent is more complex. We can distinguish two cases:

1. The whole constituent occurs in the position appropriate for the IS function of its head and the distinct IS of the subexpression is marked only by intonation. This possibility seems to be always available and it is not analyzed here.

(14) *Question: A co teda koupil makovýho a co kmínovýho? – And what did he buy with poppy-seeds and what with caraway?*

Martin	koupil	[makový	HOUSKY$_R$]	a	kmínový	ROHLÍKY$_R$.
Martin	bought	poppy-seed$_{adj.pl.acc}$	rolls$_{pl.acc}$	and	caraway$_{adj.pl.acc}$	bread-sticks$_{pl.acc}$

‘Martin bought poppy-seed ROLLS$_R$ and caraway BREAD-STICKS$_R$.’

(As for poppy-seed things, Martin bought rolls and as for caraway things, he bought bread-sticks.)

2. The part of the constituent belonging to the theme proper or rheme proper is fronted, resulting in a discontinuity. This possibility is available only in certain circumstances, which are the topic of this section.

(15) *Question: A co teda koupil makovýho a co kmínovýho? – And what did he buy with poppy-seeds and what with caraway? (the same as in (14))*

[Makový]	Martin	koupil	[_ housky]	a	kmínový	rohlíky.
poppy-seed$_{adj.pl.acc}$	Martin	bought	rolls$_{pl.acc}$	and	caraway$_{adj.pl.acc}$	bread-sticks$_{pl.acc}$

‘Martin bought poppy-seed ROLLS$_R$ and caraway BREAD-STICKS$_R$.’

(As for poppy-seed things, Martin bought rolls and as for caraway things, he bought bread-sticks.)

The examples below show fronted partial expressions of various categories:[25]

(16) Split NPs

a. AP from NP

[Makový] koupil [_ housky].
Poppy-seed$_{adj.pl.acc}$ bought rolls$_{pl.acc}$

'He bought poppy-seed rolls.' (As for poppy-seed things, he bought rolls.)

b. N from NP

[Housky] koupil [makový _].
rolls$_{pl.acc}$ bought poppy-seed$_{adj.pl.acc}$

'He bought poppy-seed rolls.' (As for rolls, he bought poppy-seed ones)

c. PP from NP

[O syntaxi] jsem si půjčil [knihu _].
about$_{loc}$ syntax$_{f.sg.loc}$ aux$_{1sg}$ refl$_{D}$ borrowed book

'I have borrowed a book about syntax.' [after De Kuthy 2002 (1)]

d. Possessive Adj from NP

[Dvořákovu] snesu [_ operu], ale symfonii ani náhodou.
Dvořák's$_{f.sg.acc}$ can-bear$_{1sg}$ opera$_{f.sg.acc}$, but symphony not-even by-accident

'I can bear Dvořák's opera, but never his symphony.'

(17) Split predicative NPs

a. N from predicative NP

[Práce] to byla [galejnická _].
job$_{pl.acc}$ it was galley-like

'It was a very hard job.' [Uhlířová 1972 p. 174]

b. A from predicative NP

[Dobrý] je Pavel [_ sportovec].
good$_{m.sg.nom}$ is Pavel sportsman$_{m.sg.nom}$

'Pavel is a good sportsman.' (As for good ...)

[25] To support the orientation of a nonnative reader, the examples contain the symbol _ in place where the fronted expression would be if it weren't fronted (i.e., if it had the same IS function as the non-fronted part of the constituent). This is for expository reasons only; it is not meant to suggest that the analysis of the data should include the notion of a trace. Also, it shows only the phrase the fronted expression syntactically belongs to, not the exact position it would occur in if it weren't fronted, which because of scrambling is not clear. The _ is placed in an unmarked position.

(18) Split AP

a. Adj out of AP

[Hrdý] je [_ na své děti.]
proud is [on self children]

'He is proud of his children.' [after De Kuthy and Meurers 2001 (1c)]

b. PP out of AP

[Na své děti] je [hrdý _ .]
on self children is [proud]

'He is proud of his children.'

Verbal attribute. In the traditional Czech syntax, sentences involving (seemingly) split phrases are sometimes analyzed by means of a so-called complement[26] or verbal attribute. Informally, in this view, split NPs are analyzed as two sister phrases – an NP and a verbal attribute. The adjective agrees with the noun in the NP in the usual way. According to this analysis, the attribute relates semantically both to the verb and to the NP at the same time. Supposedly, it relates less to the NP than a normal adjective and less to the verb than a normal adjunct.

This view is roughly analogous to the reanalysis approach to similar phenomena in English or German (see De Kuthy and Meurers 2001 and references cited there). However, for Czech, this analysis has never been formally spelled out, especially its relation to semantics. Even informal analyses are rather limited (Svoboda 1969; Úličný 1969, 1970). There is little agreement in this area: some authors (Karlík et al. 1996) analyze all discontinuities with adjectives as verbal attributes, some (e.g., Daneš et al. 1987, p. 168) reject the notion completely, while others (Hajič et al. 1999; Uhlířová 1972) differ in the place of putting the boundary between the two cases. Unfortunately, the argument for or against never exceeds a few paragraphs.

Examples like (19), where the noun *hrušku* 'pear$_{f.sg.acc}$' might be replaced by a pronoun, suggest that analysis involving verbal attributes might be a better option than assuming discontinuous constituents. Because most analyses would assume that in (19b) *velkou* 'big$_{f.sg.acc}$' is not an attribute of the pronoun *ji* 'her$_{acc}$', it seems natural to assume that analogously, in (19b), it is not an attribute of the noun *hrušku* 'pear$_{f.sg.acc}$'.[27]

[26] This term is not directly related to complements in phrase structure grammars. In this sense, a complement complements the verb in addition to its subject, objects and adjuncts. In addition to split fronting, complements are used to analyze control verbs and predicatives.

[27] Although Jarmila Panevová (p.c.) suggests analyzing (19b) as replacement of the thematic noun *hruška* by a pronoun in the surface syntax layer of (a variant of) Functional Generative Description (Sgall et al. 1986).

(19) a. Hrušku dal Martin Petrovi velkou.
pear$_{f.sg.acc}$ gave Martin Petr$_{dat}$ big$_{f.sg.acc}$
'Martin gave a BIG$_R$ pear$_C$ to Petr.'

b. Martin ji dal Petrovi velkou.
Martin her$_{acc}$ gave Petr$_{dat}$ big$_{f.sg.acc}$
'Martin gave a BIG$_R$ one to Petr.'

However, certain other cases suggest that an analysis involving discontinuous constituents is more plausible. For example, it seem more natural to analyze *o irským* 'about Irish$_{n.sg.loc}$' and *pivu* 'beer$_{n.sg.loc}$' in in (20) as two parts of a split PP. Locative is strictly prepositional, thus analysis involving two continuous clausal constituents would require to treat the preposition-less *pivu* 'beer$_{n.sg.loc}$' as an exception.

(20) *Context: Australský víno je dobrý. A co říkáš irskýmu$_C$?*
'Australian wine is good, and what do you think about Irish$_C$ wine?'

[O irským] jsem slyšel jen [_ pivu].
about Irish$_{n.sg.loc}$ aux$_{1sg}$ heard only beer$_{n.sg.loc}$
'I have heard only about Irish$_C$ beer.'

Note, however, that some speakers allow repeating the preposition, which would be an argument for a reanalysis view:

(21) [O irským] jsem slyšel jen [o pivu].
about Irish$_{n.sg.loc}$ aux$_{1sg}$ heard only about beer$_{n.sg.loc}$
'I have heard only about Irish$_C$ beer.'

Such constructions are however clearly impossible in my idiolect and my informants are split. For example, Jarmila Panevová (p.c.) judges them as better than those without the second preposition.

In the following, we assume the phrases are indeed discontinuous. The actual choice is not important for our purpose – we need *some* analysis of split-fronting so that we can analyze placement of clitics in the next chapter. Whether clitics follow the first part of a split constituent or a full reanalyzed constituent has the same consequences.

3.4.3 Unbounded Dependencies

As in English, the dependency between the fronted expression and its head (or the trace) can cross clausal boundaries. Unlike in the case of English (see e.g., Levine and Hukari 2006), this is a rather understudied area of Czech, and we are not aware of any in-depth study of the phenomenon. Brief analyses of the phenomenon can be found in (Štícha 1996) and (Petkevič 1998).

(22) a. [Makový] říkal Martin, že koupil [_ housky].
Poppy-seed$_{adj}$ said Martin that bought rolls.
'Martin said he bought poppy-seeed rolls.'

b. [Makový] říkal Martin, že si myslí, že Petr koupil [_ housky].
Poppy-seed$_{adj}$ said Martin that refl$_D$ thinks that Petr bought rolls.
'Martin said he thinks that Petr bought poppy-seeed rolls.'

c. [Pivo] jsem přece hlásil, že podávají jenom [lahvové _].
beer aux$_{1sg}$ emph announced that serve$_{spl}$ only bottled
'I did announce that they serve beer only in bottles.' [Rosen 1994 (37b)]

Such unbounded dependencies are also for non-split constituents:

(23) a. [Housky] jsem si myslel, že říkal Petr, že koupil _ Martin.
rolls aux$_{1sg}$ refl$_D$ thought that said Peter that bought Martin
'The rolls, I thought Peter said Martin had bought.'

b. [Toho kluka] si myslím, že jsem včera viděl _.
That boy refl$_D$ think that aux$_{1sg}$ yesterday saw
'That boy, I think I saw yesterday.' [Petkevič (176)]

c. [Zítra] předpokládáme, že _ tlaková výše postoupí k jihu.
tomorrow suppose$_{1pl}$ that pressure height moves to south
'Tomorrow, we suppose the pressure height will move to the south'[(Štícha 1996, p. 30) & Uhlířová]

3.4.4 Multiple Fronted Expressions

The theory of Information Structure in FGD implies that the theme and the rheme proper consist of a single (possibly partial) constituent.[28] However, examples of long fronting in (24) show that fronting of multiple constituents is possible. We are not aware of any analysis of multiple fronting in Czech, but Avgustinova and Oliva (1995) discuss a special case of this phenomenon: a clitic clusters preceded by multiple constituents. Generalizing and extending their data, we can conclude that multiple constituents can be fronted when all are contrasted, express a path (from – through – to), or are spatio-temporal stage adverbials.

[28] As a dependency theory, FGD does not use the notion of constituents directly; here we mean a subtree of a node in a dependency tree.

(24) a. All contrasted:

[Petra do Francie] říkal Pavel, že si myslí, že Martin pošle hned.
Petr$_A$ to France said Pavel that refl$_D$ thinks that Martin will-send immediately
'Pavel said he thinks Martin would send Petr to France$_C$ immediately.'

b. Path:

[Z Paříže na Remeš] si myslím, že říkal, že se stopuje blbě.
from Paris direction Reims refl$_D$ think$_{1sg}$ that said$_{3sg}$ that hitch-hike badly
'I think he said that hitching from Paris in the direction of Reims does not go well.'

c. Period:

[Od pátku do neděle] očekáváme, že bude pršet.
from Friday till Sunday await$_{1pl}$ that will rain.
'We expect that it will be raining from Friday till Sunday.'

d. Stage:

[Zítra ve vyších polohách] očekáváme, že bude pršet.
tomorrow in higher altitudes await$_{1pl}$ that will rain.
'We expect that it will be raining tomorrow in higher altitudes.'

3.4.4.1 Constituents?

The expressions participating in multiple fronting are traditionally analyzed as consisting of several constituents in Czech syntax. In fact, it is not clear how they could be analyzed differently, because Czech is traditionally analyzed in a dependency theory, which is radically endocentric (every constituent has a head) and lexicalist (there are no null heads).

The expressions, however, share some properties with single constituents. As discussed in §4.4.4, they can occur before the main clitic cluster, a place usually occupied by a single constituent. Another similarity is that they can be coordinated:

(25) a. Coordinated path, short fronting:

[[Z Varů] [do Chebu] a [z Paříže na Remeš]] *se*$_1$ *mi*$_1$ vždycky
from (Carls)bad to Cheb and from Paris in-direction Reims refl$_A$ me$_D$ alway
stopovalo blbě.
hitchhiked badly
'I always had a hard time hitching from Carlsbad to Cheb and from Paris the direction of Reims.'

b. Coordinated path, long fronting:

[[Z Varů] [do Chebu] a [z Paříže na Remeš]] říkal Martin, že
from (Carls)bad to Cheb and from Paris in-direction Reims said Martin that
se$_1$ stopuje blbě.
refl$_A$ hitchike badly

‘Martin said that it is hard to hitchhike from Carlsbad to Cheb and from Paris direction Reims.’

c. Coordinated complements, short fronting:

[[Petra] [do Francie]] a [[Marii] [do Německa]] *bych* ještě poslal, ale Martina do
Petr$_A$ to France and Marie to Germany would$_{1sg}$ still send but Martin$_A$ to
Maďarska ani náhodou.
Hungary not-even by-accident

‘I could possibly send Petr$_C$ to France$_C$ and Marie$_C$ to Germany$_C$, but never Martin$_C$ to Hungary$_C$.’

d. Coordinated complements, long fronting

[[Petra] [do Francie]] a [[Marii] [do Německa]] si myslím, že by šéf ještě
Petr$_A$ to France and Marie to Germany refl$_D$ think$_{1sg}$ that would$_3$ boss still
poslal, ale ...
sent, but ...

‘I think that the boss could possibly send Petr$_C$ to France$_C$ and Marie$_C$ to Germany$_C$, but ...’

However, the expressions participating in multiple fronting also differ from constituents in many respects; for example it is hard to use a pronoun to refer to them.

3.4.4.2 “Internal” coordination

An interesting fact that we are not ready to provide analysis of is that not only the contrasted expressions can be coordinated as group with other contrasted expressions, but that the conjunction *a* ‘and’ can be inserted between them, as in (26). This adds a certain gradation of the contrast and is easier to accept when in a negative sentence or at least in a sentence contrasted with a negative one. For example (26a) suggests that Martin is a bad choice and together with Hungary it is even worse. Without the conjunction, the statement refers only to the whole combination (*Martin visiting Hungary*) as a bad choice, and the individual conjuncts might be possible, just not together (*Martin can go to Italy and Hungary can be visited by Eva*). A similar effect has the insertion of a pause instead of the conjunction.

(26) a. [[Petra] a [do Francie]] *bych* ještě poslal, ale Martina a do Maďarska ani
Petr$_A$ and to France would$_{1sg}$ still send but Martin$_A$ a to Hungary not-even
náhodou.
by-accident
Roughly: 'I would send Petr$_C$ to France$_C$, but never Martin$_C$ to Hungary$_C$.'

b. [[Všechny sny] a [najednou]] *se mu* určitě nesplní.
All dreams and at-once refl$_A$ him$_D$ definitely not-fulfil.
Roughly: 'There is no way all his dreams will come true at the same time.'

The conjunction *a* or a prosodic boundary have similar consequences when inserted in a middle of a constituent. Consider (27a). The implication of the sentence is simply that I do not dare to babysit for the Nováks. However, when a pause is inserted in (27b) or the conjunction *a* in (27c), the implication is roughly along these lines: *I have a hard time with babysitting in general, and babysitting for Nováks is just something I do not dare at all.* These data suggest that even expressions that are traditionally analyzed as constituents with a single head can undergo multiple fronting.

(27) a. [Hlídat$_2$ děti Novákům] *si*$_1$ teda netroufnu$_1$.
watch$_{\mathsf{inf}}$ children Nováks$_D$ refl$_D$ so not-dare
'I DO NOT DARE$_R$ to babysit for the Nováks$_C$.'

b. [Hlídat děti | Novákům] *si*$_1$ teda netroufnu$_1$.
watch$_{\mathsf{inf}}$ children Nováks$_D$ refl$_D$ so not-dare
'I DO NOT DARE$_R$ to babysit$_C$ for the Nováks$_C$.'

c. [[Hlídat děti] a [Novákům]] *si*$_1$ teda netroufnu$_1$.
watch$_{\mathsf{inf}}$ children and Nováks$_D$ refl$_D$ so not-dare
'I DO NOT DARE$_R$ to babysit$_C$ for the Nováks$_C$.'

3.4.4.3 Constraints?

It is not clear whether any two (or more) expressions that can be fronted independently can be also fronted together. As we show in §4.4.4.2, the constraint suggested by (Avgustinova and Oliva 1995, pp. 36/37) in connection with clitics is too restrictive even for clitic placement. It is therefore, even more incorrect for fronting in general. In our opinion, the restrictions are more of a pragmatic than of a syntactic nature. Certain sentences with multiple frontings seem impossible simply because it is hard to imagine a context for them, especially if presented by themselves without sufficient context. We leave this issue for further study.

A similar phenomenon occurs in German, where the so-called Vorfeld has been argued to sometimes contain expressions that have been traditionally categorized as several constituents. Müller (2002, 2003, 2005) argues for analyzing them as a single constituent with an empty verbal head, which successfully constraints the Vorfeld to being interpreted as dependents of the same verbal head. However the meaning of such constructions is different in German than in Czech.

3.4.5 Some restrictions on split fronting

Czech is a free-constituent language, and therefore any clausal constituent can be fronted (with the exception of clitics; there are other constituents with restricted placement, such as determiners, but they are not clausal). However, as in other languages, there are limitations on split fronting. Below, we explore the more obvious ones.

3.4.5.1 Category Limitations.

Not every syntactic category can be fronted in a split fronting, and similarly not every category can be left behind. For example, while both relative clauses and prepositions can occur clause initially, they cannot be fronted as a result of the split fronting alone.

(28) Embedded Relative Clause

a. Napsal jsem [knížku, která půjde dobře na odbyt].
wrote aux$_{1sg}$ book which will-go well prep sale
'I wrote a book that will sell well.'

b. * [Která půjde dobře na odbyt,] napsal jsem [knížku _].
which will-go well prep sale wrote aux$_{1sg}$ book

c. * [Která půjde dobře na odbyt,] si myslím, napsal jsem [knížku _].
which will-go well prep sale refl$_D$ think wrote aux$_{1sg}$ book

(29) Clausal Relative Clause

a. [Která půjde dobře na odbyt,] jsem poznal hned.
which will-go well prep sale aux$_{1sg}$ recognized right-away
'I recognized right away which one will sell well.'

b. [Která půjde dobře na odbyt,] si myslím, že jsem poznal hned.
which will-go well prep sale refl$_D$ think that aux$_{1sg}$ recognized right-away
'I think I recognized right away which one will sell well.'

(30) Preposition (fronting NP from PP)

a. * [Na] polož tu knihu _ stůl, ne pod (stůl).
on put the book table$_A$, not under table
Intended: 'Put the book on$_C$ the table, not under$_C$ it.'

b. Polož tu knihu na$_C$ stůl, ne pod$_C$ stůl.
put the book on table$_A$, not under table
'Put the book on$_C$ the table, not under$_C$ the table.'

Fronting the demonstrative *ten* 'the/this/that' also does not seem possible.

(31) a. Včera četl [ten básník] ze své knihy.
yesterday read the poet from his book
'Yesterday, the poet was reading from his book.'

b. * [Ten] vcera četl [_ básník] ze své knihy.
the yesterday read poet from his book
'Yesterday, the poet was reading from his book.'

c. [Ten básník] včera četl ze své knihy.
the yesterday read poet from his book
'Yesterday, the poet was reading from his book.'

3.4.5.2 Embedding Limitations.

Hajičová et al. (2004) claim that a contrastive expression has a strong tendency to stand in the initial position in the surface word order, no matter how deeply it is embedded in the underlying structure of the sentence. However, this does not seem to be correct. Generally only a clausal constituent can be split (in this respect, dependents of complex predicates and of prepositions act as clausal constituents). The existence of such limitations on embedding should not be really surprising; they exist in many other languages. See for example (De Kuthy 2002, p. 11) for constraints on split NPs in German.

In (32), only the whole complement of the verb or an dependent of that complement can be fronted (although stylistically this is not the best choice). Fronting of more embedded constituents as in (32d) is clearly out. It is also impossible to front the adjective *magisterských* 'Master's', and or any other possible modifier of *diplomů* (e.g., *všech univerzit* 'of all universities'.)

(32) a. Vláda předepisuje [velikost písmen [na deskách [magisterských diplomů.]]]
government$_N$ regulates size letters$_G$ on covers Master's$_G$ diplomas$_G$
'The government regulates the character size on the covers of Master's diplomas.'

b. [Velikost písmen [na deskách [magisterských diplomů]]] vláda předepisuje.
size letters$_G$ on covers Master's$_G$ diplomas$_G$ government$_N$ regulates
'The government regulates the character size on the covers of Master's diplomas.'

c. ? [Na deskách [magisterských diplomů]] vláda předepisuje [velikost písmen _]]
on covers Master's$_G$ diplomas$_G$ government$_N$ regulates size letters$_G$
'The government regulates the character size on the covers of Master's diplomas.'

d. * [Magisterských diplomů] vláda předepisuje [velikost písmen [na deskách _]]
Master's$_G$ diplomas$_G$ government$_N$ regulates size letters$_G$ on covers

Similarly, the PP in (33) is "too embedded" to be fronted. The intended meaning can be expressed by fronting the whole PP and using the intonation to put contrast on the adjective *kategoriální* 'categorial'.

(33) a. * [O kategoriální] jsem si půjčil [knihu [_ gramatice]].
about$_{loc}$ categorial$_{f.sg.loc}$ aux$_{1sg}$ refl$_D$ borrowed book grammar$_{f.sg.loc}$
Intended: 'I have borrowed a book about categorial$_C$ grammar.'

b. * [Kategoriální] jsem si půjčil [knihu [o _ gramatice]].
categorial$_{f.sg.loc}$ aux$_{1sg}$ refl$_D$ borrowed book about$_{loc}$ syntax$_{f.sg.loc}$
Intended: 'I have borrowed a book about categorial$_C$ grammar.'

c. [O kategoriální$_C$ gramatice] jsem si půjčil [knihu _] (o závislostní článek).
about$_{loc}$ categorial$_{f.sg.loc}$ grammar$_{f.sg.loc}$ aux$_{1sg}$ refl$_D$ wrote book

'I have borrowed a book about categorial$_C$ grammar (about dependency grammar, I borrowed an article).'

3.4.5.3 Prepositions.

Rosen (2001, p. 195) shows that in a split PP, the preposition and attribute must be fronted together, as in (34). This is similar to the situation in Polish (Kupść 2000, §2.4.2) and Serbo-Croatian (e.g., Penn 1999*a*, p. 179).[29]

(34) AP from PP

a. [O [jak dotovanou]] se jedná [_ soutěž].
about how financed$_{f.sg.loc}$ refl$_A$ is talked competition$_{f.sg.loc}$
'How financed a competition is it?' [Rosen 2001 (150b)]

[29] Penn discusses split PP in connection with so-called 2W placement of clitics. In such placement, the clitics follow the first prosodic word of a sentence and can thus split the initial constituent. It is claimed (Halpern 1998, p. 111) that at least some 2W placement cannot be explained by independently split constituents, e.g., due to fronting. In Czech, clitics can follow a partial constituent only in cases when the constituent is split for other reasons.

Penn's concern is thus opposite to ours. In his analysis, it is natural to ask why anything else is required to stand initially with the preposition. In our case, it is natural to ask why the preposition is required to stand initially when anything else is fronted.

b. *Context: Australský víno je dobrý. A co říkáš irskýmu$_C$?*
'Australian wine is good, and what do you think about Irish$_C$ wine?'

[O irským] jsem slyšel jen [_ pivu].
about Irish$_{n.sg.loc}$ aux$_{1sg}$ heard only beer$_{n.sg.loc}$
'I have heard only about Irish$_C$ beer.'

c. * [Irským] jsem slyšel jen [o _ pivu].
about Irish$_{n.sg.loc}$ aux$_{1sg}$ heard only beer$_{n.sg.loc}$
'I have heard only about Irish$_C$ beer.'

However, the situation applies to any PP-split; the preposition must precede even a fronted noun. (Recall that _ denotes the unmarked position of the fronted expressions, not traces.)

(35) P+N from PP

a. [O pivu] jsem slyšel jen [_ irským _].
about beer$_{n.sg.loc}$ aux$_{1sg}$ heard only Irish$_{n.sg.loc}$
'I have heard only about Irish beer$_C$.'

b. * [Pivu] jsem slyšel jen [o irským _].
beer$_{n.sg.loc}$ aux$_{1sg}$ heard only about Irish$_{n.sg.loc}$

It does not seem that the constraints behind the examples above could be prosodic. While certain Czech prepositions are indeed proclitics, the situation applies even to multisyllabic non-clitic preposition like *kolem* 'around'. On the other hand, it is possible that the constraint is a generalization of an (originally?) prosodic constraint.

3.4.6 Summary of §3.4

In a simple case, theme proper and rheme proper correspond to clausal constituents; in objective ordering theme proper and in subjective ordering rheme proper are fronted – i.e., they occur clause initially, and can climb to higher clauses.

Split fronting, i.e., fronting of expressions that are not clausal constituents is also possible. When a theme/rheme proper does not correspond to a clausal constituent, the expression can be topicalized if the minimal constituent containing it is a clausal constituent. In this respect NPs of clausal PPs and dependents of complex predicates act as clausal constituents. The topicalized expression may, but need not, include a head of the clausal constituent it is part of. There are certain additional syntactic restrictions, for example, prepositions and non-clausal relative clauses cannot be topicalized.

The topicalized expression may consists of several expressions if they are all contrasted, if they are so-called stage adverbials, or if they express path or period.

CHAPTER 4

CZECH SPECIAL CLITICS

In this chapter, we provide an informal analysis of a certain class of Czech clitics. Many of the aspects presented here are then analyzed in Higher Order Grammar in the next chapter.

Clitics are units that are transitional between words and affixes, having some properties of words and some properties of affixes. Czech clitics (e.g. Avgustinova and Oliva 1995; Fried 1994; Hana 2004; Rosen 2001; Toman 1980, 1986, 1996, 2000), Slavic clitics (e.g. Franks and King 2000; Penn 1999*a*) and clitics in general (e.g. Anderson 1993; Zwicky 1977), present a great challenge to existing formalisms. Their ordering properties are often complex and quite different from the properties of both normal words and affixes. Also, they are subject to constraints coming from various levels of grammar – syntactic, morphological, phonological, pragmatic and stylistic.

This chapter is organized as follows: first we provide a brief discussion of clitics in general across languages, then we introduce the basic properties of Czech clitics; then we characterize the set of Czech clitics; identify their position within the clause and then the order of clitics within this cluster; and finally we analyze so-called clitic climbing. This chapter is by no means meant to be an exhaustive study of Czech clitics. Instead it focuses on core problems and especially ordering problems that are known to be hard to handle in other frameworks.

In the examples, all relevant clitics are given in italics for easier orientation. Often, numerical subscripts show the relation between clitics and the word governing them; the subscripts increase with the degree of embedding of the governors. Clitic auxiliaries have subscript zero. Otherwise the examples and their sources are presented in the same way as in the previous chapter, see Appendix B for more details.

4.1 Clitics in General

Clitics have attracted attention for a long time. They are units that are transitional between words and affixes, having some properties of the former and some of the latter. The exact mix of these properties varies considerably across languages. This means there is a whole spectrum of units between clear affixes and clear words. Delimitation of the set of clitics, and if they are treated as a separate category at all, is to a great extent an arbitrary or theory-internal decision. In the next chapter, we treat clitics as special words, with some affix-like properties, but nevertheless words.

Wackernagel (1892) was one of the first to study clitic placement. He observed that, in Greek, enclitics follow the first word of the sentence and suggested that this was a rule in Proto-Indo-European. In recent decades, there has been a been significant amount of work on clitics in general (esp. Anderson 1992; Halpern 1995; Klavans 1985; Zwicky 1977) – see (Nevis et al. 1994) for a comprehensive list.

A clitic must attach to an adjacent word (possibly through another clitic), its *host*. Typical clitics are prosodically dependent on their host. A clitic following its host is called an *enclitic*; a clitic preceding it is called a *proclitic*. In addition, there are also *mesoclitics* occurring between the host and its affixes and *endoclitics*, analogous to infixes, occurring in the middle of their hosts. However, neither mesoclitics nor endoclitics are discussed in this dissertation.[30]

Zwicky (1977) divides clitics into two classes: *simple* clitics and *special* clitics.[31] A simple clitic is a clitic whose position within the sentence is the same as position of non-clitic words of the same class. Syntactically, simple clitics behave as other non-clitic words; the only difference is phonological. For example, English *has* and *'s* have the same word order properties. The position of special clitics, on the other hand, is determined by special constraints, different from the constraints determining the position of non-clitic words. The purpose of this chapter is to describe and analyze such special behavior of Czech special clitics, we leave simple clitics aside.

[30] The status of mesoclitics and endoclitics is rather controversial. Klavans (1995) claims they are impossible. On the other hand, Harris (2002) argues that endoclitics do exists, providing evidence from Udi.

[31] He also uses the term *bound words* for phrasal clitics, for example English possesive *'s*. However as Klavans (1982, p. 33) and others pointed out, the distinction between simple clitics and bound words is not clear.

4.1.1 Placement and other basic properties of clitics

Anderson (1992) identifies six places relative to some domain where special clitics can occur:

- Initial clitics.
- Final clitics. For example, English possessive *-s* within NP.
- Second-position clitics – the clitics follow some initial element. For example, Warlpiri auxiliaries within certain S (Donohue and Sag 1999), Slavic clitics within S.
- Penultimate-position clitics – the clitics precede some final element. For example, Nganhcara pronominals within S (Anderson 1994).
- Pre-head clitics. For example, Romance pronominal clitics.
- Post-head clitics. For example, Romance clitics in certain constructions, e.g., imperatives.

Clitics can also be characterized in terms of the the following three parameters:

- *Anchor*. The clitic is placed by reference to the *first*, *head*, or *last* element;
- *Orientation*. It *precedes* or *follows* the anchor.
- *Domain* (or *scope*). It is placed within a certain domain, e.g., S, VP, NP.

Table 4.1 shows how the combination of the anchor and orientation parameters corresponds to the 6 categories of (Anderson 1992).

Type of Clitics (Anderson 1992)	Anchor	Orientation	Schematically
initial	first	precedes	[clitic anchor ...]
final	last	follows	[... anchor clitic]
second-position	first	follows	[anchor clitic ...]
penultimate-position	last	precedes	[... clitic anchor]
pre-head	head	precedes	[... clitic head ...]
post-head	head	follows	[... head clitic ...]

Table 4.1: Characterization of clitic position

The value of the orientation parameter usually determines the phonological attachment (proclitics precede their anchor, and enclitics follow the anchor). However, as Klavans (1985) shows, this is not always the case. Thus she introduces an additional parameter expressing the direction of phonological *attachment* (*left* or *right*). For example, Kwakwala determiners are NP initial clitics (domain=NP, anchor=first, orientation=precedes) but attach to the left, i.e., to the word preceding the NP (Klavans 1985, p. 106). Consider the sentence in (1). Syntactically, *x̣a* ‘OBJECT’ and *sa* ‘OBLIQUE’ mark the following words, but phonologically they attached to the preceding words (this is marked by =). This means the Kwakala determiners are syntactically proclitics, but phonologically enclitics. In Klavans’ words, they are clitics with *dual citizenship.*

(1) nəp’idi-da gənanəm =x̣a guk^w =sa t’isəm
throw-DEIC child OBJ house OBL rock
‘The child hit the house with a rock by throwing.’ [Klavans 1985 (32)]

We would also add that in the case of the second and penultimate position clitics, it is necessary to specify the nature of the *element* – for example a word, a constituent, or a fronted expression.

4.2 Basic Characteristics of Czech special clitics

Czech special clitics (henceforth just clitics[32]), like most other Slavic clitics, fall into the category of second-position clitics. They are another case of clitics with dual citizenship. Syntactically they are enclitics, following their anchor, a certain clause-initial unit, usually the first constituent. However, phonologically, they can be both enclitics and proclitics, depending on circumstances (see §4.2.2). This means the above parameters do not have to be constant for a given language or even for a given clitic.

In this section, we introduce some basic properties of Czech clitics. We show that they indeed behave differently in respect to the rest of the grammar than normal words or affixes do. We briefly talk about their phonological properties, position within the sentence, their position to each other, so-called clitic climbing and finally we will briefly discuss them from a historic perspective. The rest of the chapter then discusses most of these problems in more detail.

[32] Czech also has clitics that are not special, i.e., they are ordered as other expressions of the same category (see §4.1 for more discussion of various types of clitics). For example, clitic prepositions immediately precede their NP, as non-clitic prepositions do. The negative marker *ne-* can be considered a clitic because unlike affixes it attaches to stems of various categories, but otherwise acts as a prefix. They are not discussed in this dissertation exactly for the reason that their word-order properties are straightforward.

4.2.1 Clitics and word order

Clitics differ from the rest of Czech grammar in two important dimensions:

- Word-order freedom: Czech word order is very free as regards the possibility of moving entire phrases – virtually any scrambling is possible. By contrast, the position of clitics is very restricted – they occur most frequently in so-called Wackernagel or second position (Wackernagel 1892) and even their ordering within this position is for the most part fixed.
- Constituent discontinuity:[33] While the order of constituents is mostly free, scrambling resulting in discontinuous phrases is rather rare.[34] As we mention in (Hana 2004), clitics, however, are frequently associated with the presence of discontinuous phrases. This stems from the fact that, while their position is restricted, the positions of their governors, if any, are not. There are various factors that make a sentence with clitics more or less acceptable, but, perhaps surprisingly, the number of discontinuities caused by the clitics is not among them.

The rigidity of clitic placement can be illustrated by comparing clitics to full NPs. The indirect object (*Petrovi* 'Peter$_D$') in sentence (2a) can also occur in any other place in that sentence (except within the PP) – for example in the theme position at the beginning of the sentence, as in (2b):

(2) a. Dal Petrovi psa k vánocům.
gave Peter$_D$ dog$_A$ for Christmas
'He gave Peter a dog for Christmas.'

b. Petrovi dal psa k vánocům.
Peter$_D$ gave dog$_A$ for Christmas
'He gave Peter$_C$ a dog for Christmas.'

However, when the noun phrases here are replaced by the corresponding weak pronouns (one type of clitic), the above word-order freedom is lost – compare (2b) with the ungrammatical (3b):

[33] For dependency grammar, the most prominent linguistic tradition in the analysis of Czech (Šmilauer 1947, more formally, e.g., Sgall et al. 1986), discontinuous constituents correspond to non-projective dependency trees (Hays 1964, p. 519, allegedly already in Hays 1960.)

[34] (Hajičová et al. 2004, ftn. 1) report statistics for the training part of the layer of surface-syntax (so-called analytical layer) of PDT. According to them, about 1.9% of word dependencies in the analytical layer are non-projective and about 23% of sentences contain one or more non-projectivities. Note, however, that existence of many of these non-projectivities is dependent on the chosen linguistic theory or annotation scheme.

(3) a. Dal *mu* *ho* k vánocům.
gave him$_D$ him$_A$ for Christmas
'He gave it to him for Christmas.'

b. * *Mu* dal *ho* k vánocům.
him$_D$ gave him$_A$ for Christmas

The clitics themselves have a fixed position within a clitic cluster. So, while the order of the direct object (*psa* 'dog') and the indirect object (*Petrovi* 'Peter$_D$') in sentence (2a) can be switched and still have the resulting sentence (4a) be fully grammatical, the corresponding change of word order in sentence (3a), with its clitics, results in the ungrammatical sentence (4b).

(4) a. Dal psa Petrovi k vánocům.
gave dog$_A$ Peter$_D$ for Christmas
'He gave Peter a dog for Christmas.'

b. * Dal *ho* *mu* k vánocům.
gave him$_A$ him$_D$ for Christmas

The occurrence of multiple discontinuous phrases associated with clitics is also interesting. Sentence (5) is a normal sentence that can occur in everyday conversation. Yet the clitics *jsem*, *se*, *mu*, *to* here participate in several discontinuities, as the phrase structure in Figure 4.1 shows.

In (6), an analogous sentence without clitics (though contentwise a little bit odd), pronominal clitics are replaced by full NPs (*auto* 'car' for *to* 'it', *Petrovi* 'Petr$_D$' for *mu* him$_D$), the past tense formed with clitic auxiliary *jsem* is replaced by the future nonclitic auxiliary *budu*, and the reflexive clitic *se* is eliminated by replacing the reflexive verb *snažil se* 'try' by non-reflexive *zkoušet* 'try'. The sentence still contains the contrasted VP headed by *opravit* 'repair$_{\textsf{inf}}$', but as can be seen in Figure 4.2, the structure is much simpler.

(5) Opravit *jsem* *se* *mu* *to* včera snažil marně.
to-repair aux$_{1sg}$ refl$_A$ him$_D$ it$_A$ yesterday tried fruitlessly
'I tried to repair$_C$ it for him yesterday WITHOUT SUCCESS$_R$.'

(6) Opravit Petrovi auto budu zítra zkoušet marně.
to-repair Petr$_D$ car$_A$ will$_{1sg}$ tomorrow try$_{\textsf{inf}}$ fruitlessly
'I will be trying to repair the car for Peter$_C$ tomorrow WITHOUT SUCCESS$_R$.'

4.2.2 Phonology – Enclitics? Proclitics? Either? Neither?

Typically, Czech (2P) clitics are phonological enclitics. However there are systematic exceptions to this. Already Trávníček (1951, §103 2b) said that, after a pause, clitics procliticize to the following

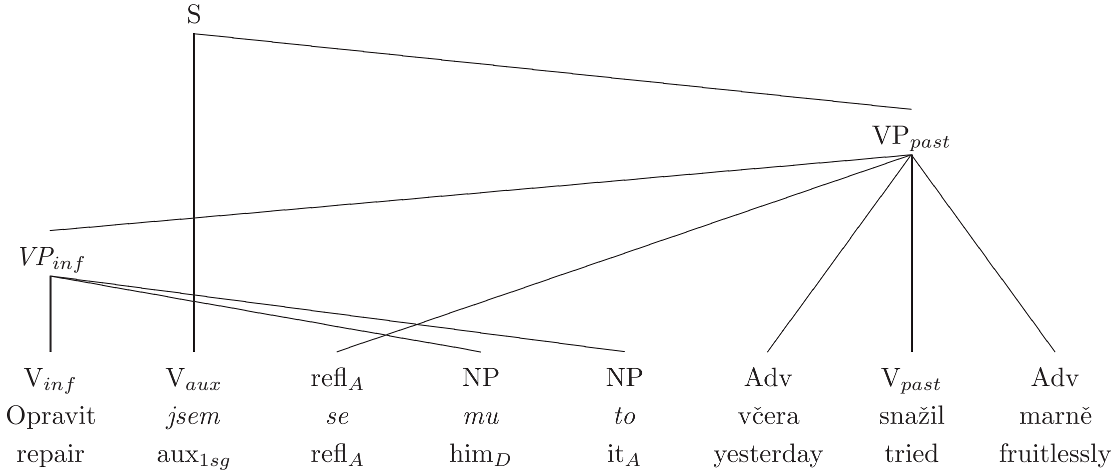


Figure 4.1: The syntactic structure of (5)

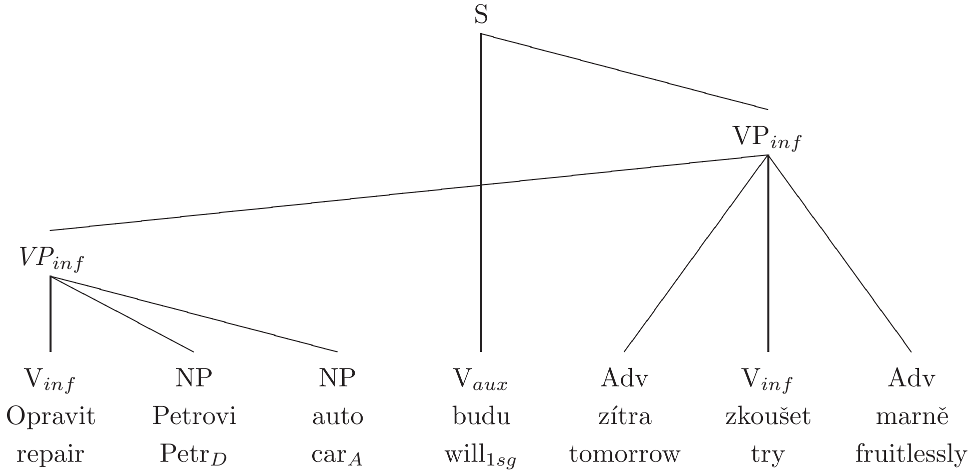


Figure 4.2: The syntactic structure of (6)

word. He claimed this was rare and unusual, which is not true in current Czech. A pause follows a heavy constituent (7), parenthetical (8), a contrastive theme (at least in some cases), or an initial constituent containing a clitic cluster (11). For example, in (7a), the clitic *se* forms a prosodic word with the material on its right, i.e., it procliticizes. It cannot encliticize, as (7b) shows (| marks a prosodic boundary).

(7) a. Knihy, které tady vidíte, | *se* dnes platí zlatem.
books which here see$_{2pl}$ refl$_A$ today pay with-gold$_I$
'The books you can see here are paid for with gold today.' [Toman 1996]

b. * Knihy, které tady vidíte, *se* | dnes platí zlatem. [Toman 1996]

(8) Ve středu, | teď se podržte kolegyně, | *jsem* navštívila hypermarket Globus.
on Wednesday, now refl$_A$ hold$_{2pl}$ colleagues$_{fem}$, aux$_{1sg}$ visited hypermarket Globus
'On Wednesday, and now hold on colleagues, I visited the supermarket Globus.' [ksk]

It is worth noting that, in Common Czech, clitics can occur even sentence-initially. The clitic *se* in (9a) and *jsme* in (9b) are obviously not enclitics. In Common Czech, sentence-initial clitics are not frequent but are possible, although they have a distinct "feel" and usually express (ostentatious) familiarity. They are are not approved in Literary Czech (if that's of any linguistic significance). Note however that (9b) was used by a governmental official on TV news.

(9) a. *Se* ví.
refl$_A$ knows$_{3sg}$
Of course.

b. (.. objevují [se] nějaké dokumenty, o kterých my *jsme* nevěděli.)
(... documents that we did not know of are surfacing.)

Jsme *se* domnívali, že *je* kompletní.
aux$_{1pl}$ refl$_A$ thought that is complete
'We thought, it [=the file] was complete.' [www.ceskenoviny.cz, 2006-05-22]

On the other hand, Czech clitics also cannot always be proclitics, as is clear from (10).

(10) Směju *se.*
Laugh$_{1sg}$ refl$_A$
I am laughing.

Toman (1996) shows that whether a clitic procliticizes or encliticizes is not a lexical property of the clitic. The sentence in (11) contains the same clitic *ji* 'her$_A$' twice in two different clitic clusters (see §4.6 for more information on multiple clitic clusters). As the object of the verb *nudilo*, it occurs in the main cluster *by ji*. In the other case, it is a part of the phrase *poslouchat ji* – the subject of the sentence. The prosodic boundary is identical with the syntactical boundary of the subject phrase, following the first *ji*. Therefore, the first *ji* encliticizes, while the second procliticizes.[35]

(11) a. Poslouchat$_2$ *ji*$_2$, | *by*$_0$ *ji*$_1$ asi nudilo.
to-listen her$_A$ would$_3$ her$_A$ probably bore.
It would perhaps bore her (e.g., Ann) to listen to her (e.g., Mary).

b. * Poslouchat$_2$ | *ji*$_2$, *by*$_0$ *ji*$_1$ asi nudilo.

c. * Poslouchat$_2$ *ji*$_2$, *by*$_0$ | *ji*$_1$ asi nudilo.

d. * Poslouchat$_2$ *ji*$_2$, *by*$_0$ *ji*$_1$ | asi nudilo. [Toman 1996]

Oliva (1998) even argues that clitics do not have to be a part of a larger prosodic unit at all and can be phonologically independent. According to him, in the most natural pronunciation of (12), the prosodic boundaries both precede and follow the clitic *bychom* 'would$_{1pl}$'.

However, we do not think their example can be generalized. First, many consulted speakers found having the prosodic boundary on both sides of *bychom* only marginally acceptable and instead preferred to procliticize it with *jak*.[36] Second, it seems that even such marginal acceptability is limited only to conditional clitics; it does not seem to be possible for, say, *se* as (13) shows. This may be related to their special status within the set of clitics. As discussed in §4.3.4.2, they can be contrasted or rhematic. Moreover, up to about century or so ago they were also used as nonclitic conjunctions to express purpose (Trávníček 1951, §103 2c). Although this usage is now archaic and has been replaced by the conjunction *aby*, it is probably still part of our passive competence and can thus influence phonological properties of the clitic in rare constructions like the one in (12). In sum, it does not seem that (12) is an example of some general possibility of Czech clitics to be phonologically independent.

(12) My všichni, co spolu chodíme, | *bychom*, | jak říká Zilvar z chudobince, měli
we all, that together walk, would$_{1pl}$, as says Zilvar from poorhouse, should$_{pl}$
držet za jeden provaz.
to-hold by one rope
'As Zilvar from the poorhouse says, all of us friends should stick together.' [Oliva 1998]

[35] *Asi* can but need not be a clitic in this example, see §4.3.6.

[36] However, some speakers, including A. Rosen, consider the variant with both boundaries fully acceptable.

(13) a. ??

My	všichni,	co	spolu	chodíme, |	*se*, |	jak	říká	Zilvar	z	chudobince,
we	all,	that	together	walk,	refl$_A$,	as	says	Zilvar	from	poorhouse,
nemáme	čeho	bát.								
not-have$_{1pl}$	of-what	to-be-scared$_{\mathsf{inf}}$								

‘As Zilvar from the poorhouse says, all of us friends have nothing to be scared of.’

b. My všichni, co spolu chodíme, | *se*, jak říká Zilvar z chudobince, nemáme čeho bát.

4.2.3 Position

We refer to the word-order position of sentential clitics within the clause as 2P. While formally, this is just a label, it is motivated by the fact that in most of the cases, this position is really the second position within the clause, in the sense of immediately following the first clausal constituent as in (14) or the head of the clause as in (3a). However, as we discuss in §4.4, there are many deviations. 2P can be preceded by (i) a complementizer + another constituent, (ii) a multi-constituent contrastive theme, and (iii) a complex adjunct (e.g., *from – to* expressions), sometimes considered to be individual constituents on the clausal level. These cases are not necessary disjoint. We refer to the material preceding clausal clitics as 1P (in the case of the embedded clauses, it is slightly more complicated; see §4.4.6).

(14)

1P		2P				
Příští	sobotu	*bych*	*mu*	*to*	mohl	dát.
next	Saturday	would$_{1sg}$	him$_D$	it$_A$	could	give$_{\mathsf{inf}}$

‘Next Saturday, I could give it to him.’

4.2.4 Multiple clitic clusters and climbing

Above, we talked about the position of clitics relative to the finite clause domain. We call this sequence of clitics the *main* or *clausal* clitic cluster. However a clause can contain additional *embedded* clusters in the domain of embedded infinitive VPs, NPs or APs, etc. In this case the clitics in general do not occur in second position; Toman (2000) uses the term *clitics in non-canonical positions.* In (15a), *se* is in the clausal cluster, *mu* in the cluster of the VP *pomoct mu ho najít* and *ho* in the cluster of the VP *najít ho.* Recall that a verb and clitics it governs are labeled with the same numerical subscripts increasing with the depth of verb embedding. Clitic auxiliary verbs get the zero subscript.

Clitics with more embedded governors can, under certain circumstances, occur in the clitic clusters of the larger domains, possibly in the clausal one – see (15b). This is traditionally referred to as *clitic climbing*. We analyze clitic climbing in more detail in §4.6; for now it is enough to say that clitic climbing is subject to several constraints and various preferences. For the following discussion it is also important to note that two clitic clusters can be adjacent, as in (16). The clitic *mu* is in the cluster of the VP *pomoct mu*, which in turn serves as the host for the clausal clitic cluster containing *se*. Phonologically *mu* is an enclitic while *se* is a proclitic, ’ and there is a *potential* prosodic boundary between them.

(15) a. Všichni *se*$_1$ snažili$_1$ [*mu*$_2$ pomoct$_2$ [*ho*$_3$ najít$_3$.]]
all refl$_A$ tried him$_D$ help$_{\mathsf{inf}}$ him$_A$ find$_{\mathsf{inf}}$
‘Everybody tried to help him to find it.’

b. Všichni *se*$_1$ *mu*$_2$ *ho*$_3$ snažili$_1$ [pomoct$_2$ [najít$_3$]].

c. [Pomoct$_2$ *mu*$_2$ *ho*$_3$ [najít$_3$]] *se*$_1$ snažili$_1$ všichni.

(16) [Pomoct$_2$ =*mu*$_2$] | *se*$_1$= snažili$_1$ všichni.
help$_{\mathsf{inf}}$ him$_D$ refl$_A$ tried all
‘Everybody tried to help him.’

4.2.5 Diachronic aspects

The constraints on the placement of Czech clitics have changed over time. According to Pavel Kosek (p.c.), the placement of Czech clitics after the first constituent is a rather new development; clitics probably did not occur in this position even in the early 1300’s. In Old Czech and in Old Slavonic, clitics usually encliticized to the first phonological word, as in (17a) (see also Trávníček 1962, p. 149). Non-functional clitics also often accompanied the finite verb, usually following it as in (17b), sometimes preceding it, as in (17c). According to Večerka (1989) the Wackernagel position after the first word is the primary position, while according to P. Kosek (p.c) the verb adjacent position was more common. Moreover, the modern accusative pronominal clitics and the conditional auxiliary were probably not constant clitics in the early stages of Czech.

(17) a. ten sě pes počě radovati
that refl$_A$ dog started to-be-happy
‘that dog started to be happy’ [Trávníček 1962, p. 149/passionl (1300’s)]

b. Gdyž přiblížieše sě Ježúš k Jeruzalému . . .
When approached refl$_A$ Jesus to Jerusalem
‘When Jesus approached Jerusalem . . . ’ [P. Kosek p.c./Mt 21,1-9]

c. Předmluva Mistra Vavřincova v Kniehy snového vykládanie tuto sě počíná . . .
foreword master Vavřinec to Books of-dream interpretation here refl$_A$ starts
'Here starts the foreword of Master Vavřinec to the Interpretation of Dreams ...'
[P. Kosek p.c./Vavřinec z Březové: Foreword to Snář . . . (early 1400's)]

While placement of clitics after the first prosodic word is still possible in modern Serbo-Croatian (Halpern 1995), this is in general not true in modern Czech. Czech clitics do follow a certain clause initial unit. However, what this unit is is determined mainly by syntax – by constituent structure and to certain extent by information structure – and only marginally by phonology. A similar development happened in other Slavic languages, including Slovak or Slovenian. So it is possible to say that historically, Slavic clitics could be roughly characterized by the following parameter configuration: domain=S, anchor=first, orientation=follows, element=phon-word and attachment=left. In Modern Czech, two parameters have different values: element=constituent and attachment=left/right. However the value of the element parameter is a simplification; there are many exceptions, as we briefly mentioned above and discuss in more detail below.

4.3 The set of Czech clitics

The set of Czech clitics is similar to that in many other Slavic languages: so-called weak pronouns, certain auxiliaries and some particles or adverbs. Clitics can be categorized as either *constant* or *inconstant* (see e.g., Karlík et al. (1996), already in Trávníček (1951, §103, §104)).[37] Constant clitics always behave as clitics; inconstant clitics can function as clitics but can also function as normal words (that is they can occur outside of a clitic cluster).[38]

4.3.1 Testing clitic-hood

Enumerating the exact set of clitics is far from trivial and probably impossible. The set is often different for different authors,[39] but the core stays the same – weak personal pronouns (including

[37] Avgustinova and Oliva (1995) use the terms *pure clitics* and *semi-clitics*.

[38] An inconstant clitic can be seen as a single word functioning two different ways or as two distinct words. The former view is implicit in most analyses of clitics; the latter view is adopted by for example Avgustinova and Oliva (1995) or Esvan (2000). We do not see any benefit in resolving this problem. As is seen in the following chapter, we choose the former possibility, but nothing hinges on that choice.

[39] For example, *to* is considered to be a clitic by (Karlík et al. 1996, p. 649), but not by (Rosen 2001, p. 212). All traditional sources list *li* 'whether' alongside the other 2P clitics, but this is disputed by (Fried 1994) and (Avgustinova and Oliva 1995). (Rezac 2005) leaves out the copula and most of the fringe clitics.

reflexives), past and conditional auxiliary. Inclusion of other clitics depends on the author: *li* 'if', *to* 'it', other auxiliaries and various short particles and adverbs, etc. are all sometimes included.

To identify that a particular unit is a clitic and not a regular affix or word, one has to obviously show it has properties different from those of normal affixes and properties different from those of normal words. Various criteria for clitic-hood have been suggested (e.g. Carstairs 1981; Klavans 1995); we use tests based on a subset of properties suggested by (Zwicky 1977, 1985; Zwicky and Pullum 1983).

It is relatively easy to distinguish all the clitic candidates from affixes. With the exception of -*li* 'if' and -*s* 'aux$_{2sg}$' in Official Czech, all candidates for clitic-hood discussed below can be hosted by any syntactic category. Affixes are selective of the stems they attach to. Pronominal clitics, in addition, often climb from embedded clauses (§4.6); such freedom of movement is also not found for affixes.

It is far more challenging to decide whether a particular candidate is a clitic or a normal word. Many authors use as the main or only criterion of clitic-hood the inability of clitics to carry accent on their own. However, as Zwicky (1985) remarks, this is the most unreliable test. First, there are many words that are not clitics and usually occur without accent. Second, (Klavans 1982, §2) shows that some clitics can bear accent under certain circumstances. In Czech, this is the case for proclitic prepositions. The conditional auxiliary can even bear contrastive accent – see §4.3.4.2. Moreover, unlike in many other languages, prosody plays only a secondary role in the grammar of Czech clitics – their direction of prosodic dependence is unspecified (§4.2.2), and prosody is nearly irrelevant in their placement. Obviously, the test is also hard to apply to inconstant clitics. For these reasons, we decided to exclude the test of prosodic deficiency. We consider a word to be a clitic when at least one of the following tests holds. The first two tests are useful only for identifying constant clitics, the third test can be used to identify (some) inconstant clitics. Note that while the features of clitics motivating these tests are rather universal, the tests themselves are dependent on the interplay of those features with the rest of the Czech grammar, and are thus suited only for identification of Czech clitics and not clitics in general.

1. **[*Alone]** Clitics cannot occur in isolation, e.g., as an answer to a question.

   In this respect clitics are similar to bound morphemes. The test is an instantiation of a more general *binding* principle formulated by (Zwicky 1977, p. 2): "Bound morphemes are affixes". The strength of the binding principle is language and clitic dependent. For example, in Czech the negative proclitic *ne*- or the enclitic -*li* (see this section below) cannot be separated from their host by a parenthetical. On the other hand, Czech 2P clitics can be preceded by a parenthetical. (However, in that case they attach phonologically to the following word, see §4.4.)

2. [***Final**] Clitics cannot occur sentence-finally.

   Clitics cannot stand sentence finally, unless the final position is 2P at the same time (the example must be constructed in such a way that such interpretation is impossible). This is a consequence of a more general property of clitics: clitics have more restricted distribution than normal words (although not as much as affixes). As mentioned in §4.1, in Czech, they occur in so-called 2P in the sentence. Because it is not easy to exactly identify that position, we use the slightly weaker test above.

   It is also true that, apart from a very colloquial register (§4.2.2), clitics cannot be sentence-initial. However, it is sometimes hard to separate this and other registers when making grammaticality judgments in less common cases. Note that this restriction does not follow from the prosodic deficiency of clitics. As mentioned above, Czech clitics do not need to lean phonologically on the expression preceding them; they can procliticize when preceded by a prosodic boundary.

3. A member of a clitic cluster is a clitic.

   This property can be instantiated in two specific tests:

   (a) **[1P-Cl]** A word between 1P and a clitic is a clitic.

   When true, the candidate is in 2P – (i) because it follows 1P, it is either in 2P or follows an empty 2P; (ii) since the candidate is followed by a clitic, 2P cannot be empty. One must make sure the candidate actually follows 1P and is not part of it. Using an uncontrasted proper name for 1P is a safe bet; the candidate cannot form a constituent with it, and none of the multiconstituent cases for 1P discussed in §4.4.4 are possible. This test was used by Rosen (2001, p. 208). This test is not able to identify clitics that are either required to be on the end right of the cluster, or that are separated from the end by such clitics. Unlike the previous two tests, this test can identify inconstant clitics.

   (b) **[Cl-Cl]** A word between two clitics without possibility of any prosodic boundaries between the three, is a clitic.

   This means all three words belong to the same clitic cluster and thus obviously all are clitics. It must be clear that the two surrounding clitics belong to the same cluster, see §4.6 for discussion of multiple clusters.

	dative			genitive/accusative		
	weak	either	strong	weak	either	strong
1sg	mi	mně [mɲɛ]			mě [mɲɛ]	mne*
2sg	ti [cɪ]		tobě [tobjɛ]	tě [cɛ]		tebe
3sg m	mu		jemu	ho, jej*		jeho
3sg n	mu		jemu	ho, jej*, je$^*_{acc}$		jeho
3sg f		jí [jiː]			ji [jɪ]	
1pl		nám			nás	
2pl		vám			vás	
3pl		jim			jich$_G$/je$_A$	

(* – rare; *je$_A$* – only in accusative, *jich$_G$* – only in genitive)

Table 4.2: Personal pronouns in genitive, dative and accusative

However, as is evident from the rest of this section, the boundary between clitics and non-clitics is often fuzzy. There are some obvious cases of clitics such as the weak personal pronouns but then there are less clear cases, especially among inconstant clitics. In one view, any short word without much lexical content can be considered an inconstant clitic – under certain conditions, when deaccented in theme, it can appear at the boundary of the clitic cluster. We discuss some of these borderline cases in §4.3.6. However, we are more interested in the complex word-order properties of clitics than in exactly enumerating them. For this purpose it is enough to limit the set of clitics to the more obvious cases.

4.3.2 Personal Pronouns

The Czech personal pronouns are summarized in Table 4.2. It is traditional to distinguish weak and strong forms of pronouns. Weak forms, e.g., *ti* 'you$_{sgD}$', are prototypical constant clitics, strong forms, e.g., *tobě* 'you$_{sgD}$', are never clitics.[40] Forms that can be either weak or strong, e.g., *nám* 'us$_D$', are inconstant clitics. Initial *j-* changes to *ň-* [ɲ] after a preposition,[41] e.g., *jej* 'him$_{G/A}$' vs. *bez něj* 'without him$_G$'.

Originally, *mně* 'me$_D$' (pronounced [mɲɛ], the same way as *mě* 'me$_{G/A}$') was only a strong pronoun, but now is frequently used as a weak one, too, as (18) shows.

[40] According to Veselovská (p.c.), in Moravia, the eastern region of Czechia, *mu* 'him/it$_D$' and *ho* 'he/it$_{GA}$' (and in some regions also *mi* 'me$_D$' and *ti* 'you$_D$') are used as strong pronouns, Bohemian Czech strong pronouns being rarely used.

[41] In spelling, *ň* + *i* → *ni*: *ji* → *ni* 'her$_{GA}$', *jí* → *ní* 'her$_D$', *jich* → *nich* 'them$_G$', *jim* → *nim* 'them$_D$'; and *ň* + *e* → *ně*: *jej* → *něj* 'him/it$_{GA}$', *jeho* → *něho* 'him/it$_{GA}$', *jemu* → *němu* 'him/it$_D$', *je* → *ně* 'it$_A$/them$_A$'.

(18) Dej *mi/mně* to!
Give me$_D$ it
‘Give it to me!’

In Common Czech, dative and accusative forms in the first and second person singular are sometimes used interchangeably – for example *mi* ‘me$_D$’ is sometimes used as an accusative clitic (20).[42]

(20) Vidíš *mě/mi*?
See me$_A$
‘Do you see me?’

In the 3rd person feminine, this neutralizations is complete – the pronoun can be pronounced with short vowel [jɪ] and long vowel [jiː] in both cases, although the long form is more common. The pronunciation and spelling of Official Czech must be learnt at school. Still many speakers, including myself, have to pause and think when they are required to use the “correct” form. On the other hand, *mne* ‘me$_{G/A}$’, *jej* ‘he/it$_{G/A}$’ and *je* ‘it$_A$’ are formal and are rarely used; *mě*, *ho* and *ho*, respectively are used instead. However, the preposition forms *něj* and *ně* are common. Note also that in Czech the demonstrative pronoun *to*, an inconstant clitic, is often used where English would use a 3rd person personal pronoun.

Examples (21 – 23) show the difference between the three types of personal pronouns. From (21), it is obvious that strong pronouns *tobě* ‘you$_{sgD}$’ and inconstant *jí* ‘her$_D$’ can be rhematic and stand sentence-finally, similarly to full NPs, while weak pronouns cannot. Instead, weak pronouns must occur in 2P, roughly following the first constituent, as in (21b) or (22). The sentence in (22) also shows that *jí* can be a clitic. Similarly to *ti* ‘you$_{sgD}$’, a constant clitic, it occurs in the middle of a clitic cluster, surrounded by constant clitics *bych* ‘would$_{1sg}$ and *ho* ‘him$_A$’. This is not possible for *tobě* ‘him$_D$’, a strong pronoun, or for a full NP. Similarly, (23) shows that while NPs and strong pronouns can occur in isolation, weak pronouns cannot.

[42] Some speakers judge this as ungrammatical in such sentences, but most accept it in more expressive utterances like:

(19) Kurva, Jituš, neser mi, co je na dluhách výhodnýho?
expletive Jituš not-piss-off me$_A$ what is prep debts advantageous
approx: ‘Jituš, do not piss me off, what is it that’s advantageous about debts?’
[syn5/M. Viewegh: Účastníci zájezdu; fiction 1996]

(21) *[*Final] a clitic cannot occur sentence finally:*

a. Marie dala sešit Petrovi / tobě / **ti* / jí.
Marie gave notebook Petr$_D$ / you$_{sgD}$ / you$_{sgD}$ / her$_D$.
'Marie gave a notebook to Petr / you / *you / her.'

b. Marie *ti* / *jí* dala sešit.
Marie you$_{sgD}$ / dather gave notebook
'Marie gave you / her a notebook.'

(22) *[Cl-Cl]*

Nedal *bych* *ti* / *jí* / **tobě* / *Petrovi *ho* ani za nic.
not-gave would$_{1sg}$ you$_{sgD}$ / her$_D$ / you$_{sgD}$ / Petr$_D$ him$_A$ not-even for nothing.
'I would not give it to you / her / *you / Petr for anything.'

(23) *[*Final] a clitic cannot occur sentence finally:*

A: Komu dala Marie sešit?
'Who did Marie gave a notebook to?'

B: Petrovi. / Tobě. / **Ti.* / Jí.
Petr$_D$ / you$_{sgD}$ / you$_{sgD}$ / her$_D$.
'To Petr.' / 'To you.' / *'To you.' / 'To her.'

4.3.3 Reflexives

As (24-27) show, accusative *se* and dative *si* reflexive pronouns are constant clitics. The strong form *sebe* corresponds to *se*, and *sobě* corresponds to *si*. In addition, there are two contractions with the second-person singular present auxiliary (used to form past tense) – *ses* = *jsi* + *se* and *sis* = *jsi* + *si*. The contractions are not obligatory but are preferred: in the spoken corpus Oral2006, 84% of cases are contractions, in the private correspondence corpus KSK, it is 73%.

(24) *[*Final] a clitic cannot occur sentence finally:*

a. Marie chválila v posudku Petra / sebe / **se*.
Marie praised in review Petr / refl$_A$ / refl$_A$.
'Marie praised Petr$_R$ / herself$_R$ / *herself$_R$ in the review.'

b. Marie *se* chválila v posudku.
Marie refl$_A$ praised in review .
'Marie praised herself in the review.'

(25) *[*Alone] a clitic cannot stand alone:*

A: Koho chválila Marie v posudku?
'Whom did Marie praise in the review?'

B: Petra. / Sebe. / **Se.*
Petr$_A$ / refl$_A$ / refl$_A$
'Petr. / Herself / *Herself.'

(26) *[*Final] a clitic cannot occur sentence finally:*

a. Marie poslala e-mail Petrovi / sobě / **si.*
Marie sent e-mail Petr$_D$ / refl$_D$ / refl$_D$.
'Marie sent an e-mail to Peter / herself / *herself.'

b. Marie *si* poslala e-mail.
Marie refl$_D$ sent e-mail
'Marie sent an e-mail to herself.'

(27) *[*Alone] a clitic cannot stand alone:*

A: Komu poslala Marie e-mail?
'Who did Marie send an e-mail to?'

B: Petrovi. / Sobě. / **Si.*
Petr$_D$ / refl$_D$ / *refl$_D$
'Petr. / Herself. / *Herself.'

In addition to the reflexive anaphoric use, Czech reflexives are used in several other constructions: the so-called *reflexive passive* (28a), reciprocals (28b) and reflexive tantum verbs like *smát se* 'laugh' (28c). See (Králíková 1981; Panevová 1999) for more details. In all these cases, only the clitic form can be used.

(28) a. V Jičíně *by* *se* postavily dva kruhové objezdy.
In Jičín would$_3$ refl$_A$ built$_{pl}$ two roundabouts.
'In Jičín, they would build two roundabouts.'

b. Ani nevím, kdy *jsme* *si* naposledy psaly, tak ...
Even not-know$_{1sg}$ when aux$_{1pl}$ refl$_D$ last-time wrote$_{pl}$, so ...
'I even don't know, when was the last time we wrote to each other, so ...' [ksk]

c. Celou prohlídku *jsem* *se* musel smát.
Whole inspection aux$_{1sg}$ refl$_A$ must laugh$_{\textsf{inf}}$
'I had to laugh during the whole inspection.' [ksk]

As clitics, all reflexives, regardless of their meaning, have the same word-order properties.

		copula/passive auxiliary	past auxiliary	future auxiliary	conditional auxiliary
sg	1	jsem	jsem	budu	bych/bysem
	2	jsi/jseš	jsi/-s	budeš	bys/bysi/by+-s
	3	je		bude	by
pl	1	jsme	jsme	budeme	bychom/bysme
	2	jste	jste	budete	byste
	3	jsou		budou	by

Table 4.3: Copula in present tense and auxiliaries

4.3.4 Auxiliaries

The forms of the verb *být* 'to be', see Table 4.3, can serve as a copula or as an auxiliary in these periphrastic constructions (see also §A.1.5):

- past tense: auxiliary in present tense + past participle; the auxiliary is not present in the 3rd person. E.g., *psal jsem* 'I wrote/was writing$_{masc}$', *psal* 'he wrote'. Note that even the verb *být* 'to be' forms past tense periphrastically: *byl jsem* 'I was$_{masc}$', *byl* 'he was$_{masc}$'. Note that we use the term *past auxiliary* to refer to the auxiliary used to form the past tense, the verb *být* 'to be' in present tense.

- future tense: auxiliary in future tense + imperfective infinitive. E.g., *budu psát* 'I will write'. *být* forms future tense by the future auxiliary alone: *budu* 'I will be'.

- conditional: conditional auxiliary + past participle. E.g., *psala by* 'she would write$_{fem}$'. Similarly as with past tense, the verb *být* forms the conditional the same way: *byl bych* 'I would be'.

- past conditional: conditional auxiliary + auxiliary in past participle (possibly in frequentative) + past participle. E.g., *byla by psala* 'she would have written$_{fem}$', *bývala bych psala* 'I would use to write', *byla bych byla* 'I would have been'. The past conditional is rare in Common Czech, and the simple conditional is used instead.

- passive: copula in the appropriate tense and mood + passive participle. E.g., *jsem obdivován* 'I am adored$_{masc}$', *byl jsem obdivován*, 'I was adored$_{masc}$', *budeme obdivováni*, 'we will be adored$_{masc}$', *byl by obdivován*, 'he would be adored$_{masc}$', *byla bys bývala obdivována*, 'you would have been adored$_{fem}$'.

The different position of the auxiliaries in these examples is due to the fact that, as discussed below, some of the auxiliaries are or can be clitics, while others cannot. It is not natural for clitics to occur initially even in such fragments. The past tense and conditional auxiliary are constant clitics; the non-negated copula and passive auxiliary are inconstant clitics and the future auxiliary is never a clitic.

4.3.4.1 Future auxiliary

The future auxiliary (see Table 4.3) is not a clitic. Thus its position in the sentence is relatively unrestricted, it can be rhematic or contrasted, as in (29) or it can form a single-word sentences, as in (30). Contrast these sentences with similar sentences with the other auxiliaries below.

(29) *Unrestricted position:*

a. V pondělí *mu* bude Petr pomáhat.
On Monday him$_D$ will$_{3sg}$ Petr help$_{\mathsf{inf}}$
'On Monday, Peter will help him.'

b. V pondělí *mu* Petr bude pomáhat.

c. V pondělí *mu* Petr pomáhat bude. (*Final test fails)

d. Bude *mu* v pondělí Petr pomáhat?

(30) *[^{OK}Alone] – *Alone test fails:*

A: Budete *mu* pomáhat?
'Will you be helping him?'

B: Budeme.
will$_{1pl}$
'We will.'

4.3.4.2 Conditional auxiliary

The forms of the conditional auxiliary are listed in Table 4.3. The forms *bysem*, *bysi* and *bysme* are colloquial variants. The form *bysme* is closer to the official language than the other two forms. The 2sg form *by* is used with reflexives and is discussed below. The auxiliary is a constant clitic. Unlike the future auxiliary and other verbs, the conditional auxiliary cannot in general stand sentence finally – compare (31) with (29). And the auxiliary cannot form sentences by itself, for example as an answer to a question – compare (32) with (30).

(31) *[*Final] a clitic cannot occur sentence finally:*

a. * Petr *mu* pomáhal *by*.
Petr him$_D$ helped$_{m.sg}$ would$_3$

[1P-Cl]

b. Petr *by* *mu* pomáhal.
Petr would$_3$ him$_D$ helped$_{m.sg}$
'Petr would help him.'

(32) *[*Alone] a clitic cannot stand alone:*

A: Pomohl *bys mu to* udělat?
'Would help him to do it?'

B: * *Bych*.
Would$_{1sg}$

B: Pomohl.
helped$_{m.sg}$
'I would.'

Aby, kdyby. The auxiliary is also present in contractions with subordinate conjunctions in *aby* 'in order' (conj. of purpose/order/wish) and *kdyby* 'if', e.g., *abych*, *abys*, *abysme*, *kdybyste* – see example (33). These contractions are obligatory. See §4.4.6 on discussion on the position of the main clitic cluster relative to the complementizer contractions.

(33) Chce po nás, abychom *mu* koupaliště převedli bezúplatně.
wants$_{3sg}$ prep us so-that$_{1pl}$ him$_D$ swimming-pool transferred without-charge
'He wants us to transfer the swimming pool to him free of charge'. [syn5]

Diachrony and current reanalysis. Historically, the conditional auxiliary forms are aorist forms of the verb *být* 'to be' and the construction with past participle, now expressing conditional, had the meaning of past perfect tense (Rejzek 2001). Neither aorist nor past perfect are part of modern Czech. These idiosyncratic forms (from a present point of view) show the effect of reanalysis into particle *by* + past tense auxiliary. One and the same speaker can have both forms – whether two competing grammars or two competing forms is a different issue that is irrelevant here. The reanalysis is probably caused by the similarity of the 2nd and 3nd persons of both auxiliaries and by the presence of past participles in both periphrastic constructions. Many speakers have even taken the next logical step and write them as two words: *by jsme* for *bychom*, *aby jsme* for *abychom*, *kdyby jsme* for *kdybychom*, etc., see for example (34) (notice that in the second example, one conditional is reanalized, while the other is not). Table 4.4 shows that the reanalyzed forms of the 1st person plural

	Oral	PMK	KSK	
original: (a|kdy)bychom	57	66	312	
reanalyzed: (a|kdy)bysme, ...	541	355	185	(86, or 46% as two words)
percentage of reanalyzed	90	84	37	
original: (a|kdy)bych	2612	2084	3002	
reanalyzed: (a|kdy)bysem, ...	33	13	12	(12, or 100% as two words)
percentage of reanalyzed	1.2	0.6	0.4	

Table 4.4: Prevalence of reanalyzed forms in spoken and correspondence corpora

are clearly replacing the original forms, while they are rare in 1st person singular. Such reanalysis means that the original clitic is replaced by two clitics – the undeclined particle *by* and the finite past tense auxiliary. The finite auxiliary then governs the particle.

(34) a. Pokud *by* *jste* *se* setkal s nestandartním chováním aplikace …
If would aux$_{2pl}$ refl$_A$ met with nonstandard behavior application$_G$ …
'If you encountered any nonstandard application behavior … '
[mojebanka e-mail support 2007/05]

b. Chtěla *bych* Ti taky zavolat, aby *jsme* pokecaly.
Wanted would$_{1sg}$ You also call$_{\mathsf{inf}}$ so-that aux$_{1pl}$ chated.
'I would also like to call you to chat.' [ksk]

Reflexive contractions. Just as past tense auxiliaries form contractions with reflexives, *jsi* + *si* → *sis*, and *jsi* + *se* → *ses*, so do conditional auxiliaries: *bys* + *si* → *by sis* (35), *bys* + *se* → *by ses*, also *aby sis*, etc. This is another feature showing the similarity of morphological properties of both auxiliaries. While in the case of the past tense auxiliaries the contractions are optional (although preferred), in the case of the conditional auxiliaries they are obligatory (**bys si*, **abys si*), probably to avoid double *s*. Note however, that when the second person form *bys* is reanalyzed as the full form auxiliary *by jsi*, the contraction is also optional (36).

(35) a. A myslím, že *by* *sis* *ho* měla přečíst.
And think$_{1sg}$, that would aux-refl$_{2sg}$ him$_A$ should read$_{\mathsf{inf}}$
And I think, you should read it.' [ksk]

b. * A myslím, že *bys* *si* *ho* měla přečíst.
And think$_{1sg}$, that would$_{2sg}$ refl$_D$ him$_A$ should read$_{\mathsf{inf}}$

(36) No umíš *si* *to* představit, že *by* *[j]si* *si* postavil třeba chatu někde
Well can refl$_D$ it imagine$_{\mathsf{inf}}$ that would aux$_{2sg}$ refl$_D$ built$_{m.sg}$ say cottage somewhere
na hřbitově?
at cemetery

'Well, can you imagine, you would build say, a cottage, somewhere at a cemetery?' [Oral2006]

Dissyllabic clitics? One might argue that the bi-syllabicity of certain conditional auxiliary forms (*bychom*, *bysme*, etc.) means they are not clitics at all. However, they have exactly the same distribution as monosyllabic conditional auxiliaries, which in turn have distribution similar to other clitics. However, the bi-syllabicity might be another reason why the conditional clitics are being reanalyzed as a sequence of *by* + past auxiliary.

Stressed conditional auxiliary The conditional auxiliary can under certain circumstances be in contrastive theme – see (37). However, even then, surprisingly, they are still in 2P, not at the beginning of the sentence as contrastive themes usually are. The contrast is expressed purely prosodically; this is similar to marking certain other morphemes as rhematic/contrasted, e.g., past tense morpheme *-l*. One could thus say, that *by* is a syntactically constant clitic, but phonologically inconstant.[43] This is a different situation from Slovenian (Franks and King 2000) or Serbo-Croatian (Spencer 1991, p. 353), where the conditional auxiliaries are clearly inconstant clitics – only deaccented variants occur in 2P.

(37) A: Takže Petr *to* udělá?
'So Petr will do it?'

B: Říkal, že by$_C$ *to* udělal, kdyby ...
said that would it do if ...
'He said, he would$_C$ do it, if ...'

This is not possible with other clitics. This is not surprising, since for all of them there are other, less exceptional, options available. Most of the clitics have corresponding strong nonclitic forms that can be used (*ti* → *tobě*, *se* → *sebe*). Also there is no need to put contrast on the past tense auxiliary. It is more a marker of person than of "pastness" (the past morpheme *-l* of the past participle can indeed be stressed), and to put contrast on person, one simply puts it on the subject as in (38).

(38) a. Navrhoval *jsi*, abysme sem šli.
suggested aux$_{2sg}$ conj$_{1pl}$ here gone.
'You suggested going here.'

[43] We could also simply assume, following (Klavans 1995) that clitics do not need to be prosodically deficient.

b. *Navrhoval JSI$_R$, abysme sem šli.

c. Ty$_C$ *jsi* navrhoval, abysme sem šli. (Tak nenadávej.)
you aux$_{2sg}$ suggested conj$_{1pl}$ here gone.
It was you$_C$, who suggested going here. (So don't complain.)

4.3.4.3 Past and Passive auxiliary, copula

The present tense forms of the verb *být* 'to be', see Table 4.3, are used as (i) a copula, (ii) a passive auxiliary or (iii) past auxiliary. The 2sg copula form *jseš* is colloquial.

Past tense auxiliary The past tense auxiliary is a clitic and thus is restricted to 2P. In general, it cannot occur sentence finally[44] (39) and cannot stand isolated (40). Non-clitic auxiliaries do not have such restrictions – see for example the future auxiliary in (29) and (30), above, or the copula in (41) and (42) below.

(39) *[*Final] a clitic cannot occur sentence finally:*

a. * A museli *ho* dát zpátky [j]sme.
And must him$_A$ give$_{\mathsf{inf}}$ back aux$_{1pl}$

b. A museli *[j]sme ho* dát zpátky.
And must aux$_{1pl}$ him$_A$ give$_{\mathsf{inf}}$ back
'And we had to give him back.' [oral2006]

(40) *[*Alone] a clitic cannot stand alone:*

A: Nabídli *jste mu to*? (past auxiliary)
offered$_{pl}$ aux$_{2pl}$ him$_D$ it$_A$
'Have/Did you offered it to him?'

B: * Jsme.
aux$_{1pl}$

B: Nabídli (*jsme*).
offered$_{pl}$ aux$_{1pl}$
'We did.'

Copula and passive auxiliary On the other hand copula and passive auxiliary are inconstant clitics. This means they can be contrasted or rhematic. Outside of the clitic cluster their position is not restricted – they can stand sentence finally (41) or occur in isolation (42). However, they can

[44] Note that while the auxiliary stands finally in the (40)B, this is the special case where the final position is 2P at the same time – see the discussion of the test [*Final] in §4.3.1.

be also clitics, as Rosen (2001, p. 210) shows. In (43) the copula is a part of a larger clitic cluster. As shown in §4.4.4, a clitic cluster can be preceded by more than one constituent only when these constituents express path, period, stage or are contrasted. None of these is the case here, thus it is logical to call the copula a clitic in these sentences. Moreover, it is subject to the constraint on morpho-lexical ordering of clitics (§4.5) and occurs initially in the cluster, as auxiliary clitics do.

(41) *[^{OK}Final] can be final when non-clitic*

a. Já *si* myslim, že zrovna vy taková [j]ste. (copula)
I refl$_D$ think that just you such$_{fem}$ are$_{2pl}$

'I think that you$_C$ ARE$_R$ like that.' [Oral2006]

b. Pro ostatní kategorie limity stanoveny jsou. (passive)
For rest-of categories limits set are.

'The limits ARE$_R$ set for the rest$_C$ of the categories.' [syn5]

(42) *[^{OK}Alone] can occur alone when non-clitic*

a. A: Jste dneska doma? (copula)

'Are you at home today?'

B: Jsme.
are$_{1pl}$

'We are.'

b. A: Jsi pozván na pondělí? (passive auxiliary)

'Are you invited for Monday?'

B: Jsem.
am$_{1sg}$

'I am.'

(43) *[1P-Cl]*

a. [Jedinou radostí] *jsou mu* dopisy z domova, ...
Only joy are$_{3pl}$ him$_D$ letters from home

'The only joy for him are the letters from home, ...' [Rosen 2001 p. 210 / syn0]

b. [Nakonec] *je ti ho* skoro líto.
at-the-end is him$_D$ $_A$him nearly sorry

'At the end, you feel nearly sorry for him.' [Rosen 2001 p. 210]

c. [A teď] *je ho* tam taková spousta.
and now is him$_G$ there so much

'And now there is so much of him/it' [Rosen 2001 p. 210 / syn0]

d. [To] *je mu* podobný.
that is him_D similar
'That's exactly him.' [Rosen 2001 p. 210 / syn0]

Comparison The difference in clitic-hood between the copula/passive auxiliary and past tense auxiliary is not surprising – Toman (1980) lists several other aspects where the copula and the past tense auxiliary differ. They all show that the past tense auxiliary is more idiosyncratic than the copula, which behave more like a normal verb.

1. Negation prefix *ne-* attaches to the copula/passive auxiliary, but not to the past tense auxiliary. Sentences in past tense are negated by negating the past participle. Toman (1980) says this might be a consequence of the clitic-hood of the past auxiliary, assuming Czech clitics cannot be prefixed. Note that this is not a universal principle; Klavans (1985) mentions examples of affixes attaching to clitics.
2. The past tense auxiliary can form *-s* contractions in 2nd person singular. This is not possible for the copula or passive auxiliary.
3. The past tense auxiliary can be omitted in 1st person singular. Again, this is not possible with the copula or passive auxiliary.
4. Colloquially, *(j)seš*[45] is often used for the copula/passive auxiliary in the 2nd person singular. As Toman (1980) argues, the *jseš* form is probably by analogy with regular conjugation á la *píš-eš* 'write_{2sg}', *nes-eš* 'carry_{2sg}', etc. In many Moravian dialects, this goes even further with *(j)su* being used in 1st person singular, analogously to *píš-u* 'write_{1sg}', *nes-u* 'carry_{1sg}'. Again, this is not possible in the case of the past tense auxiliary.

It is worth noting that regarding the use of the past tense auxiliary, Czech is somewhere between Russian and Serbo-Croatian. In Russian, the past tense does not use any auxiliary, while in Serbo-Croatian the auxiliary is used in all persons. In Czech, the auxiliary is used in the first and second persons, while the third person is formed by a bare past participle. However, in Czech passive, the auxiliary occurs in all three persons.

[45] As with other forms of *být* 'to be', the initial *j* is usually not pronounced. In written Common Czech, the *j* is often omitted, too.

(44) a. Psal *jsem* dopis. (Czech)
wrote aux$_{1sg}$ letter$_A$

b. Ja pisal pismo. (Russian)
I wrote letter$_A$

c. Pisao *sam* pismo. (S-C)
wrote aux$_{1sg}$ letter$_A$
'I was writing a letter.'

(45) a. Psal dopis. (Czech)
wrote letter$_A$

b. On pisal pismo. (Russian)
He wrote letter$_A$

c. Pisao *je* pismo. (S-C)
wrote aux$_{3sg}$ letter$_A$
'He was writing a letter.'

4.3.5 tu 'here'

The adverb *tu* 'here' is a constant clitic, with *tady* or *zde* being nonclitic counterparts used in rheme or under contrast. However, the status of *tu* is less clear than that of the other constant clitics. The examples (46) with *tu* sentence-final or (47)) with *tu* isolated do not seem outright wrong (as, say, the corresponding sentences with the past tense auxiliary are), but instead sound hypercorrect or regional. Also there are a few expressions where *tu* is used sentence initially, for example (48), without having the strong colloquial flavor of other sentence initial clitics, as in (9). Also, there are some dialects where *tu* is clearly an inconstant clitic.

(46) *[*Final] a clitic cannot occur sentence finally:*

a. Kdyby se pořádně snažili, byl by ten zápas tady / v Praze / ?tu.
if refl$_A$ really tried been would that match here / in Prague / here
'If they really tried, the match would be HERE$_R$ / in PRAGUE$_R$ / HERE$_R$.'

b. Kdyby se pořádně snažili, byl *by* *tu* aspoň ten zápas.
if refl$_A$ really tried been would here at-least that match
'If they really tried, at least the MATCH$_R$ would be here.'

(47) *[*Alone] a clitic cannot stand alone:*

A: Kde bude ten zápas?
'Where is the the match going to take place?'

B: V Praze. / Tady. / ?*Tu*.
in Prague / here / here
'In Prague.' / 'Here.' / ?'Here.'

(48) Tu máš.
Here have$_{2sg}$
'Here you are.' [syn5]

Note that *tu* is also an adverb 'at that moment' (49) and a determiner 'this$_{fem.acc}$' (50), neither a clitic. While all three are etymologically related, we regard them as three separate homonymous words.

(49) Tu *se* Jirka zarazil.
Suddenly refl$_A$ Jirka paused
'Suddenly, Jirka paused/balked' [syn5]

(50) Tu knížku *jsem* *mu* četl.
That$_{fem.acc}$ book$_{fem.acc}$ aux$_{1sg}$ him$_D$ read.
'I read that book.'

4.3.6 Fringe clitics

The set of inconstant clitics is hard to clearly enumerate. Various short particles or adverbs with relatively little semantic content can be destressed and thus (seemingly?) function as clitics. An incomplete list of possible clitics, based on (Franks and King 2000, p. 103), is given in (51). Short (1993, p. 495) (similarly also (Karlík et al. 1996)) adds pronouns with prepositions to the list but he comments that "rules are impossible to give in this area of considerable subtlety".

(51) *tam* 'there', *však* 'though, but', *ale* 'though, but', *už* 'already', *prý/prej* 'allegedly',[46] *teda/tedy* 'so', *asi* 'probably', *snad* 'possibly (I hope)'

Note about translation: It is hard to find English expressions corresponding to these words in their clitic usage – it such usage they seem to have much less content and are much more backgrounded than their usual English counterparts. It many cases it seems that the speaker assumes the content communicated by the clitic is already known to the hearer. In addition, the words *prý/prej* are very close to being a modality marker – the speaker somehow distances himself from the statement, 'allegedly' the usually given translation, seems too strong in many cases. *už* 'already' is often subsumed by present perfect tense, while *však/ale* seem to be 'though' in clitic use while 'however' in their nonclitic use.

When clitics, these words usually follow the pronominal clitics in the clitic cluster (apart from being not the most typical, this is another reason why we label them as *fringe*). However, this implication does not go the other way – when a word from (51) is adjacent to clitics in a clitic cluster it can be either a clitic and be part of that cluster or be a non-clitic and be just adjacent to that cluster.

[46] *Prý* is a hypercorrection that replaced the original form *prej* in Official Czech: *praví* 'say$_{3sg/pl}$' → *praj* → *prej* → *prý*. See for example Rejzek (2001). In Common Czech, *prej* is more common.

All the tests suggested in §4.3.1 are useless in such case. One guide can be provided by phonology. Franks and King (2000, p. 113, ftn. 21) discuss this for *asi* in example (52) – it can be a clitic, with the initial vowel reduced or deleted, or it can be rhematic or contrasted and thus not be a clitic.

(52) Poslouchat$_2$ *ji*$_2$, | *by*$_0$ *ji*$_1$ asi nudilo.
to-listen her$_A$ would$_3$ her$_A$ probably bore.
'It would perhaps bore her (e.g., Ann) to listen to her$_C$ (e.g., Mary).

The words *však* and *prej/prý* are the easiest to classify as clitics because they can also occur at the beginning of the cluster following the host, as in (53), and therefore clearly part of the cluster (per the [1P-Cl] test).

(53) a. Delta *prý* *se*$_1$ snaží$_1$ udržovat "rodinné" ovzduší mezi zaměstnanci, . . .
Delta allegedly refl$_A$ strives maintain$_{\mathsf{inf}}$ family-like atmosphere among employees . . .
'Delta allegedly strives to maintain family-like atmosphere among employees . . . ' [syn6]

b. Chtěl *prý* *se* naučit ping-pong, ale . . .
wanted allegedly refl$_A$ learn ping-pong, but . . .
'He wanted to learn ping-pong, but . . . ' [syn6]

c. Osobně *však* *bych* považoval úplné zapomenutí těch událostí za nejlepší řešení.
Personaly though would$_{1sg}$ considered complete oblivion those events as best solution
Personaly though, I would consider a complete oblivion of those events to be the best solution. [syn6]

Note that the word *však* has at least two distinct meanings: either 'though/but', as in (54), or it a meaning similar to 'vždyť' that can be translated as *either*, *too*, sometimes *well*, etc., as in (55). It can be a clitic only in the former meaning.

(54) Zatím *se* *jim* *to* *však* nepodařilo.
so-far refl$_A$ them$_D$ it$_A$ though not-succeeded
'So far they did not succeed though.'

(55) Však ty víš, kde bydlím.
particle you know$_{2sg}$ where live$_{1sg}$
'Well, you know where I live.' or 'You do know where I live' [ksk]

4.3.7 *li* 'whether'

Traditionally (Karlík et al. 1996; Petr 1987), *li* 'whether/if' is considered to be a sentential clitic. However, Fried (1994) notices that, synchronically, *li* is a rather peripheral example of such a clitic.

Unlike other clitics and more like affixes it can be hosted only by certain syntactic categories. It mostly attaches to a finite verb (56), past participle and the particle *ne* 'not' (57). Other hosts, as the adverb *doma* 'home' in (58), are possible but very rare, sounding archaic and/or poetic.[47]

(56) V horším případě[,] má -*li* špatnou náladu a potřebuje *si ji* vybít, přijde osobně.
In worse case has$_{3sg}$ if bad mood and needs refl$_D$ her vent-on, comes in-person
'In a worse case, if he is in a bad mood and needs to vent it on, he comes in person.' [ksk]

(57) Navíc na výzo budu mít pět dvojek a za to *mě* rovnou přizabijou,
Moreover at final-report$_{colloq.}$ will have five twos and for that me$_A$ right-way nearly kill
ne -*li* zabijou !
not if kill
At the final report, I will have five [Bs] and for that they will nearly kill me right away, if not completely. [ksk]

(58) Dobrý den, doma -li pan Hordubal?
Good day home if Mister Hordubal
Hello, is Mister Hordubal at home? [syn5/K. Čapek: Hordubal; fiction 1933]

Avgustinova and Oliva (1995) do not consider *li* to be a sentential clitic at all. Instead, they claim, it is a word clitic attaching to the first word in the sentence. *li* appears to be the first member of the clitic cluster because the word it is usually hosted by, the finite verb, is a possible host for other clitics as well. They provide (a rather poetic, but still grammatical) example (59) showing that it can be detached from the cluster. The corresponding sentence where *li* does not split the NP *lásce své* 'your love' and immediately precedes *se* is worse, which would be highly unusual if *li* were a normal sentential clitic.

(59) Lásce -*li* své *se* v žití budeš protiviti, žebrákem půjdeš světem.
love$_D$ if own refl$_A$ in living will$_{2sg}$ oppose beggar$_I$ go$_{2sg}$ world$_I$
'If you oppose your love in your life, you will go through the world as a beggar.' [Avgustinova and Oliva 1995 (16)]

However, at least sometimes *li* can attach to multi-word phrases. In (62), it attaches to two coordinated verbs *poslouchám* 'listen$_{1sg}$' a *čtu* 'read$_{1sg}$'. Pragmatically it would be odd to interpret the

[47] Fried (1994) mentions only finite verbs as potential hosts, however *ne* 'not' (*ne-li* 'if not') is a common host, too. Syn2005, a balanced corpus of current *written* Czech, contains about 46,000 cases of finite verbs as hosts, about 3,100 cases of past participles, about 1,100 cases of *ne*, and some cases of *zda* 'if' and *než* 'than'. There are a few cases of other types of hosts in the corpus, but all that we checked were in fiction written in the first half of the 20th century (although the query produced about 700 such cases, many of them are tagging errors).

first verb as a separate clause. It seems that these cases are rather limited and we did not find any more complicated hosts in the corpora.[48]

(62) [Poslouchám a čtu] -*li* některé předvolební sliby kandidátů do Senátu, tak ....
listen and read if some pre-election promises of-candidates$_G$ to Senate then ...
'When I listen and read some of the pre-election promises of the candidates for Senate, ...' [Rosen (p.c.)/Syn2006pub]

li is actually very rare in Common Czech, as usually the conjunctions *jestli(že)* (originating from *jest*, an archaic form of 'is' + *li*), *pokud* and most frequently *když* are used instead.[49] So it is hard for a native speaker to make any robust judgments on the clitic. We thus exclude this clitic from

[48] Rosen (2001) provides even more interesting example given in (60) to support his claim that *li* can be a sentential clitic. We could analyze the sentence in two ways: either -*li* is hosted by the coordination of the two verbs *vstanu* 'get up$_{1sg}$' and *obléknu* 'dress$_{1sg}$' as in (61a) or only by the second verb as in (61b). Pragmatically, (61a) seems much more plausible. However, while this is an attested utterance, in our view it seems to be a performance error. All consulted speakers judged the sentence as incorrect or marginal (Some of the speakers did not want to judge the grammaticality with claims similar to "I know what the sentence is supposed to mean and there are probably no rules about these things".) or insisted it must have the meaning of (61b). Note also that (61a) is problematic for another reason: the clitic cluster contains *li*, a clitic related to the whole coordination, and *se*, a clitic related only to the second verb (*obléknu se* means 'I dress myself', there is no *vstanu se*), moreover separated from that verb by *li* – a highly unusual situation.

(60) Vstanu a obléknu -*li se*, je tím vyčerpán můj příděl energie pro zbývající den.
get-up and get-dressed if refl$_A$ is by-that spent my quota energy$_G$ for rest day
'If I get up and get dressed, my quota of energy for the rest of the day is spent.' [Rosen 2001 p. 210]

(61) a. [Vstanu a obléknu] -*li se*, je tím . . .
b. [Vstanu] a [obléknu -*li se*, je tím . . .]

[49] The following table shows that the preference is clearly different in different registers. It compares distribution of various (potentially) conditional complementizers in the syn2005 corpus (written, mostly Official Czech) and Oral2006 corpus (spoken, mostly Common Czech). While *li* accounts for 11% of those complementizers in syn2005, its share is negligible in Oral2006. Note that *když* is ambiguous between conditional 'if' and temporal meaning 'when'.

	syn5		Oral2006	
	tokens	%	tokens	%
li	51,588	11	18	0
když	293,459	63	4,287	73
jestliže	15,093	3	19	0
jestli	39,711	8	1,450	25
pokud	69,277	15	120	2
Total	469,128		5,894	

further consideration. However, if one decided that it *is* a sentential clitic, the modifications to the presented analysis would be only slight and straightforward.

4.3.8 Summary of §4.3

Overall, the set of Czech clitics is similar to that in many other Slavic languages. It can be divided into *constant* clitics and *inconstant* clitics. Constant clitics always behave as clitics; inconstant clitics can function as clitics but can also function as normal words. Enumerating the exact set of clitics is far from trivial and probably impossible. We have used the following tests to distinguish them from regular words:

- Clitics cannot occur in isolation ([*Alone]).

  Unlike normal words but similarly to affixes, they cannot occur in isolation.

- Clitics have restricted position ([*Final]).

  Their position is also more restricted than the position of normal words, although not as much as the position of affixes – they occur in so-called 2P in the sentence. Because, it is not easy to exactly identify that position, we use a slightly weaker test – they cannot stand sentence finally (unless it is 2P). Moreover, apart from a very colloquial register, they also cannot be sentence initial.

- A word followed by a clitic and preceded by 1P ([1P-Cl]) or another clitic (with no prosodic boundary between the clitics; [Cl-Cl]) is a clitic.

  Unlike the previous two tests, this test can identify inconstant clitics. The problem is that it fails short for clitics occurring on the right edge of the clitic cluster.

In addition there are other less, easily applicable tests – clitics are usually short monomorphemic units, they cannot bear contrastive accent by themselves, etc.

Using these tests, we obtained the following set of clitics.

1. Constant clitics:

   (a) all weak pronouns: *mi* ‘me$_D$’, *ti* ‘you$_{sgD}$, *ho* ‘$_A$him, etc. See Table 4.2.

   (b) weak reflexives: *se* (accusative), *si* (dative), and contractions with *jsi* aux$_{2sg}$: *ses*, *sis*.

   (c) past and conditional auxiliary

   (d) *tu* ‘here’ (however, in some dialects this is an inconstant clitic)

2. Inconstant clitics:

    (a) some personal pronouns: *jí* ‘her$_D$’, *nám* ‘us$_D$’, etc.

    (b) *to* ‘it’

    (c) non-negated copula, passive auxiliary

    (d) fringe clitics – various short particles or adverbs with a relatively little semantic content: *tam* ‘there’, *však* ‘though, but’, *ale* ‘though, but’, *už* ‘already’, *prý/prej* ‘allegedly’, … As the label suggests, fringe clitics are the most uncertain group.

4.4 Position of the main clitic cluster

As mentioned earlier in this chapter, the position of clitics is rather restricted. This applies both to the position of clitic clusters within the sentence and the relative position of clitics within a single clitic cluster. In this section, we address the possible positions of the whole clitic cluster, the next section discusses order of clitics within a single cluster.

Note: This dissertation discusses only position of the main clitic cluster, it does not address the position of clitics in embedded non-finite clauses. These clitics either precede or immediately follow their governor. There is very little work on the position of embedded clitic clusters; one exception is (Toman 2000).

Clitics usually follow the first clausal constituent in a phrase. However, there are many exceptions to this placement. The main cluster can be preceded by a partial constituent on the one hand or by several constituents on the other. In the following, we argue that these are not unusual clitic positions but instead, unusual frontings. We also argue that clitics can be positioned either relative to the first constituent or to the fronted expressions, which in most cases results into the same placement.

4.4.1 Following a clausal constituent

Usually, the main clitic cluster follows a single clausal constituent (a full sister of the head of the clause), as shown in (63). This constituent can be of various complexity ranging from a single word to a coordinated phrase, subordinate clause or a phrase modified by several clauses. The examples also show that both the head of the phrase and the word immediately preceding the clitic cluster can have any category.

(63) a. Noun:

Vražda *by* vzbudila zbytečný rozruch.
Murder would$_3$ cause unnecessary disturbance.

‘A murder would cause an unnecessary disturbance.’ [syn0]

b. Adverb:

Právě *jsem* *ti* chtěl volat.
just aux$_{1sg}$ you$_D$ wanted to-call.

‘I have just wanted to call you.’ [syn5]

c. Particle

Tak *si* na něj dávejte pozor.
So refl$_D$ at $_A$him pay$_{2pl}$ attention

‘So, be careful about him.’ [syn0]

d. Pronoun

Ono *by* *mu* *to* vadilo?
It$_{PP}$ would$_3$ him$_D$ it$_A$ minded.

‘He would mind it?’ [syn0]

e. PP

[Na koho jiného než na šéfa hádankářské rubriky] *by* se Konipas obrátil ?
To whom else than to chair of-quiz section would$_3$ refl$_A$ Konipas turned

‘To whom else than to the chair of the quiz section should Konipas turn?’ [syn5]

f. Coordinated NPs:

[Sociální demokraté a odbory] *se* domnívají, že ...
Social democrats and unions refl$_A$ think that ...

‘The Social Democrats and the unions think that ...’ [pdt]

g. Complex NP with a relative clause and an apposition:

Advokát, který zastupuje v České republice otce, JUDr. Hráský, *se* domnívá, že . . .
Attorney which represents in Czech Republic father JUDr. Hráský refl$_A$ thinks that . . .

‘The attorney representing my father in the Czech Republic, JUDr. Hráský thinks that . . . ’ [pdt]

4.4.2 Past participle

A well known exception to the above situation are sentences with an initial past participle – only the participle precedes the clitic cluster, while its complements follow it – see (65). One of the reasons for this could be that speakers probably perceive the past participle as the head of S rather than the finite auxiliary (the finite auxiliary being some kind of detached morpheme or a specifier of the participle).[50] This is especially true in the 3rd person, where there is no auxiliary, as (65) shows. Sometimes there is no auxiliary in the 1st person as well, see §4.3.4.3.

(64) Podíval$_1$ *jsem*$_0$ *se*$_2$ na hodinky.
Looked aux$_{1sg}$ refl$_A$ at watch
'I looked at my watch.' [syn5]

(65) Podíval$_1$ *se*$_2$ na hodinky.
Looked refl$_A$ at watch
'He looked at his watch.' [syn5]

Note that from the point of view of dependency grammar theories, finite verbs preceding 2P clitic cluster are a similar type of exception – the finite verb is the root of the dependency tree – see Figure 4.3.

(66) Nelíbí *se* *mi* jeho pes.
not-like refl$_A$ me$_D$ his dog$_N$
'I do not like his dog.' [syn5]

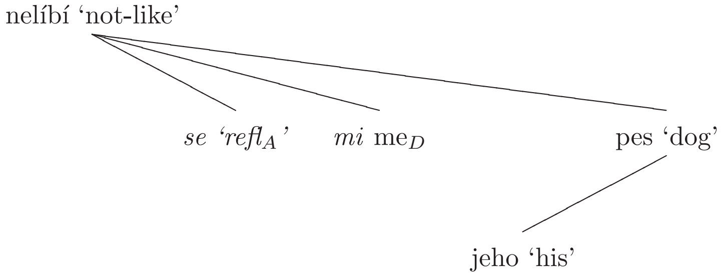


Figure 4.3: The dependency structure of (66)

[50] Actually, this is the way past tense is analyzed in Functional Generative Description (FGD; Sgall et al. 1986), the most prominent linguistic theory analyzing Czech. The auxiliary is considered to be similar to a morphological affix. The annotation in the Prague Dependency Treebank (Böhmová et al. 2001) follows this. Some other researchers, for example Ackerman and Webelhuth (1998), view auxiliaries similarly. However, in FGD, all auxiliaries are analyzed in this a way, including the future tense auxiliary or modals. In both of these cases, the main verb in infinitive can occur in the 1P with other dependents, excluding the auxiliaries.

4.4.3 Following a partial clausal constituent

While in most cases clitics are preceded by a full clausal constituent, sentences with clitics preceded by a partial clausal constituent are not rare.

The partial clausal constituent in 1P may be a full constituent at some level of embedding. For example *ten wordovský dokument* 'that Word document' in (67a) is a full object of the embedded infinitive *otevřít* 'open'. But the 1P expression may also be a true partial constituent, containing a head with only some of its daughters (the case of several daughters without a head is discussed in the next section). The head may be a head of a clausal constituent as in (67b) or a more embedded constituent (67c).[51]

(67) a. Full embedded constituent

[Ten wordovský dokument] *se*$_1$ *mu*$_1$ nepodařilo$_1$ otevřít$_2$.
that Word document refl$_A$ him$_D$ not-succeeded open$_{inf}$
'He did not manage to open that Word document.'

b. Partial clausal constituent

[Pohlídat$_2$ děti] *si*$_1$ možná troufnu$_2$ [Novákům] (ale určitě ne Hanům)
watch$_{inf}$ children refl$_D$ maybe dared Nováks$_D$
'I might dare to babysit$_C$ FOR THE NOVÁKS$_R$. (but certainly not for the Hanas)'

c. Partial embedded constituent

[Hlídat$_2$ děti] *bych*$_0$ *ti*$_1$ nepřál$_1$ [Novákům.] (ale Hanovi jsou OK)
watch$_{inf}$ children would$_{1sg}$ you$_D$ wished Nováks$_D$
'I would not wish you to watch children for the Nováks. (but the Hanas are fine)'

Not every partial constituent can precede the clitic cluster. For example, determiners seem to be out even when contrasted, as the example in (68) shows.

(68) * Tenhle mi slíbil peníze člověk.
this me$_D$ promissed man money
Intended: 'This$_C$ man promised me money.' [Rosen 2001 (191a)]

Rosen (2001) analyzes the constraints on possible partial constituents in such position as constraint on clitic placement. However, as the examples below show, the distribution of partial constituents is independent of clitics. Instead, it can be simply explained by constraints on split-fronting (§3.4.2), what-ever they are. The sentences in (69), parallel to (67) but with no clitics, show that the clitic

[51] In the following examples, *pohlídat* is a perfective variant of the imperfective verb *hlídat* 'watch$_{inf}$'.

simply follows the first part of an independently split constituent. The sentences in (70) show that the distribution also corresponds to possible long-fronted expressions. Note that in all examples below, we translate the fronted expression as contrasted. The reason is that they are the easiest to accept without a context. However in an appropriate context, the fronted expression may be interpreted as non-contrastive theme proper or as rheme proper; see §3.4 for more details.

(69) Short fronting, no clitics:

a. Full embedded constituent (no clitic)

[Ten	wordovský	dokument]	nešlo$_1$	otevřít$_2$.
that	Word	document	was-not-possible	open$_{inf}$

'It was impossible to open that Word document$_C$.'

b. Partial clausal constituent (no clitics)

[Pohlídat	děti]	můžu	[Novákům]
watch$_{inf}$	children	can$_{1sg}$	Nováks$_D$

'I can babysit$_C$ FOR THE NOVÁKS$_R$.'

c. Partial embedded constituent (no clitics)

[Pohlídat	děti]	budu	moct	[Novákům]
watch$_{inf}$	children	will$_{1sg}$	be-able$_{inf}$	Nováks$_D$

'I will be able to babysit$_C$ FOR THE NOVÁKS$_R$.'

(70) Long fronting:

a. Full embedded constituent

[Ten	wordovský	dokument]	vím,	že	*se*$_1$	*mu*$_1$	nepodařilo$_1$	otevřít$_2$.
that	Word	document	know$_{1sg}$	that	refl$_A$	him$_D$	not-succeeded	open$_{inf}$

'That Word document, I know that he did not manage to open.'

b. Partial clausal constituent

[Hlídat$_2$	děti]	říkal	Martin,	že	*si*$_1$	možná	troufne$_2$	[Novákům].
watch$_{inf}$	children	said	Martin	that	refl$_D$	maybe	dared	Nováks$_D$

'Martin said that he might dare to babysit$_C$ FOR THE NOVÁKS$_R$.'

c. Partial embedded constituent

[Hlídat$_2$	děti]	říkal	Martin,	že	*by*$_0$	*ti*$_1$	nepřál$_1$	[Novákům].
watch$_{inf}$	children	said	Martin	that	would$_3$	you$_D$	wished	Nováks$_D$

'Martin said that he he would not wish you to babysit$_C$ FOR THE NOVÁKS$_R$.'

(71) Impossible split:

a. * [Tenhle$_C$] slibuje peníze každému člověk.
this promises money everybody$_D$ man
Intended: ‘this$_C$ man is promising money to everybody.’

b. [Tenhle$_C$ člověk] slibuje peníze každému.
this promises money everybody$_D$ man
‘This$_C$ man is promising money to everybody.’

c. * [Tenhle$_C$] říkal Martin, že mu slíbil peníze člověk.
this said Martin that me$_D$ promised man money
Intended: ‘Martin said that this$_C$ man had promised him money.’

d. [Tenhle$_C$ člověk] říkal Martin, že mu slíbil peníze.
this man said Martin that me$_D$ promised money
‘Martin said that this$_C$ man had promised him money.’

4.4.3.1 Splitting a constituent

According to general grammar books, a clitic cannot split a constituent. For example, M. Grepl in (Karlík et al. 1996, §840) says:

> If the first position is occupied by a complex syntactic unit [i.e., by a multiword constituent], infinitival construction or a sentence, clitics are positioned in a way not to separate the expressions forming the [multiword constituent], infinitival construction or sentence, including an apposition or a subordinate clause.[52]

Similarly, Fried (1994, p. 158, ftn. 5), Toman (1986, p. 124) and others claim this is not possible (unlike in Serbo-Croatian). The examples used to prove this point are usually along the lines of (72). While Serbo-Croatian allows the clitic *mi* to either split the NP *taj pesnik* ‘that poet’ or to follow it, in Czech the NP cannot be split.

(72) Serbo-Croatian: [Comrie 1981 p.22]

a. [Taj pesnik] *mi* čita knjigu.
That poet me$_D$ reads book
‘That poet is reading a book to me.’

b. [Taj] *mi* [pesnik] čita knjigu.
That me$_D$ poet reads book
‘That poet is reading a book to me.’

[52] In original: “Pokud tedy první pozici obsazuje rozvitý větný člen, infinitivní konstrukce nebo věta, umisťují se příklonky tak, aby nerozdělily výrazy, které tvoří jeden větný člen, infinitivní konstrukci nebo větu, včetně přístavku a vedlejší věty.”

Czech: [Fried 1994 p.159]

c. [Ten básník] *mi* čte ze své knihy.
That poet me$_D$ reads from his book
'That poet is reading from his book to me.'

d. * [Ten] *mi* [básník] čte ze své knihy.

While it is true that in *this* Czech sentence the split is impossible, the generalization that clitics cannot split sentence initial constituents is incorrect. There are many possible cases of constituent split by clitics in Czech. A common case is a partial infinitival VP, as in (73) – the clitic *si* separates the contrastive theme *pohlídat děti* 'to watch children' from the theme *Novákům* 'for Nováks'. The difference between this sentence and a similar sentence in (67b) above is that here the constituent *pohlídat děti Novákům* would be continuous if it weren't for the clitic.

(73) [Pohlídat děti] *si* [Novákům] troufnu. (ale opravit auto ne.)
watch$_{\mathsf{inf}}$ children$_A$ refl$_D$ Nováks$_D$ dare$_{1sg}$
'I DARE$_R$ to watch children$_C$ for Nováks. (but not to repair their car)'

In (73), the material preceding the clitics is a partial constituent and includes its head. However the head can also follow the clitic. In such case, usually the clitic is preceded by a single full subconstituent of the interrupted constituent:

(74) a. *Context: Discussing what one can watch for the Nováks:*

[Děti] *si* [Novákům pohlídat] troufnu. (ale psa ne.)
children$_A$ refl$_D$ Nováks$_D$ watch$_{\mathsf{inf}}$ dare$_{1sg}$
'I DARE$_R$ to watch children$_C$ for Nováks. (but not the dog)'

b. *Context: Discussing for whom one can watch children:*

[Novákům] *si* [děti pohlídat] troufnu. (ale Císlerům ne.)
Nováks$_D$ refl$_D$ children$_A$ watch$_{\mathsf{inf}}$ dare$_{1sg}$
'I DARE$_R$ to watch children for Nováks$_C$. (but not for Císlers)'

Clitics can also split NPs in a similar fashion:

(75) a. *Context: In an answer to a letter talking about various topics, including a request for photographs of the other person's son: Pošli mi prosím nějaký fotky s Martinem, ať vidím, jak vyrostl. – 'Send me please some photos with Martin, so I can see how he is growing.'*

[Fotky] *ti* [nějaký] určitě pošlu, ale ...
Photos$_A$ you$_{sgD}$ some$_A$ definitely send, but ...
'I will send you some photos$_C$, but ...'

b. *Context: They speared horses with spears. I saw it myself.*

[Patricka] *jsem* [probodnutého] neviděl, ale nepochybuji, že *ho* probodli.
Patrick$_A$ aux$_{1sg}$ speared$_A$ not-seen but not-doubt that him$_A$ speared.
'I DID NOT SEE$_R$ speared Patrick$_C$, but no doubt they speared him.' [syn5]

(76) a. [Střízlivého] *jsem* [Patrika] neviděl, ani nepamatuju.
sober$_A$ aux$_{1sg}$ Patrik$_A$ not-seen not-even not-remember$_{1sg}$
'I do not remember when I saw Patrik sober$_C$ the last time'

b. *A comment to somebody showing his new shoes:*

[Hezké] *sis* [botky] koupil.
nice$_A$ refl$_D$-aux$_{2sg}$ shoes$_A$ bought
'You bought NICE$_R$ shoes.' (easiest to interpret in subjective ordering)

The clitics can even be preceded by several subconstituents of the split constituent – see (77). These cases are exactly parallel to cases covered in §4.4.4 and thus do not need any further discussion here.

(77) a. Path:

[[Z Chebu] [do Prahy]] *bych* [pěšky jít] nechtěl.
From Cheb to Prague would$_{1sg}$ by-foot go$_{\mathsf{inf}}$ not-wanted
'I would not like to walk from Cheb to Prague$_C$ by foot.'

b. Multiple contrasted:

[[Petra] [do Francie]] *bych* [poslat] ještě mohl, ale Martina do Maďarska ani náhodou.
Petr$_A$ to France would$_{1sg}$ send$_{\mathsf{inf}}$ still could but Martin$_A$ to Hungary not-even by-accident
'I could send Petr$_C$ to France$_C$, but never Martin$_C$ to Hungary$_C$.'

In all the sentences in (73-77), the clitic cluster follows a fronted part of a split constituent. From the point of clitic placement, it is only an accident that the rest of the constituent immediately follows the clitic cluster.[53]

Other properties follow from properties of fronting as well. The fact that the split by clitics is optional simply follows from the fact that split-fronting is optional, as discussed in §3.4.2. The

[53] This means the motivation for split constituents is different in Serbo-Croatian and Czech. In Serbo-Croatian, the clitic splitting a constituent in so-called 2W placement, is positioned by rules of prosody – the clitic follows the first prosodic word. In Czech it is information structure.

In addition, Serbo-Croatian clitics have the same option as Czech clitics – so-called 2D placement when its position is determined mainly by syntax – it roughly follows the first constituent. As Halpern (1996) argues that many cases of 2W can be analysed as 2D placement with 1D being an independently motivated partial constituent.

fact that the sentences in (73-77) seem to be less common than sentences where the clitics are not followed by the second part in (67) again follows from the properties of split fronting. A split is more likely when the two parts of the constituent have large difference in Information Structure. However, a fronted expression is usually thematic (it is rhematic in subjective ordering, but that is less frequent) and expressions following clitics immediately are usually thematic too. Finally, the resistance of most determiners to being split fronted also explains the impossibility of (72).

4.4.4 Following several constituents

Under certain circumstances, they can be also preceded by expressions that have been traditionally regarded as multiple constituents. This applies to path, period and stage adverbials and to multiple contrasted expressions, the same type of expressions that allow multiple fronting (§3.4.4).

4.4.4.1 Path, Period, and Stage Adverbials

Avgustinova and Oliva (1995) observed that the initial position can also contain several local or temporal adverbials expressing path (78a) or period (78b), or providing a "stage" for the sentence event (78c).

(78) a. [Od hrobky Caecilie Metelly na předměstí Říma] [přes vyprahlé roviny Apulie]
from tomb of-Caecilia Metella on suburb of-Rome over dried plateaus of-Apulia
[až po jižní pobřeží poloostrova] *se*$_1$ jako nikde nepřerušená rovná čára táhne$_1$
up to southern coast of-peninsula refl$_A$ as never interrupted straight line runs
nejznámější ze všech antických cest – Via Appia.
most-famous from all ancient roads – Via Appia.

'From the tomb of Caecilia Metella in the Rome suburbs over the dried plateaus of Apulia up to the southern coast of the peninsula runs the best known of all ancient roads, the Via Appia, in an uninterrupted straight line.' [Avgustinova and Oliva 1995 (41)]

b. [Od pátku] [do neděle] *se* zde narodilo pět miminek.
From Friday till Sunday refl$_A$ here born five babies.

'From Friday to Sunday, five babies were born here.' [syn5]

c. [Včera] [na Rudém náměstí] *se* stejná skupina starobolševických demonstrantů
Yesterday on Red Square refl$_A$ same group of-old-bolshevik demonstrants
opět střetla s milicí.
again clashed with militia

'Yesterday on the Red Square, the same group of old-bolshevik demonstrants again clashed with militia.' [Avgustinova and Oliva 1995 (55)]

Note that (79b) is incorrect. While the adverbials are identical to (79a), they cannot be interpreted as a path.

(79) a. [Z chalupy v Krkonoších] [do bytu na pražském sídlišti] *se* *mu*
From cottage in Krkonoše Mts. to apartment at Praguian neighborhood refl_A him_D
povedlo přivézt jen málo věcí.
managed take only few things
'From the cottage in Krkonoše Mountains to his apartment at a Prague housing development, he managed to take only few things.' [Avgustinova and Oliva 1995 (46a)]

b. * [Z chalupy v Krkonoších] [do bytu na pražském sídlišti] *se* *mu*
From cottage in Krkonoše Mts. to apartment at Praguian neighborhood refl_A him_D
hodilo jen málo věcí.
came-in-handy only few things
intended: 'From the cottage in Krkonoše Mountains, only few things were useful for his apartment at a Prague housing development.' [Avgustinova and Oliva 1995 (46b)]

Many speakers prefer the constituents in a particular order – the path and period in *from – through – to*, and the stage in *time – place.* we would also add, that the adverbials must have the same function in the Information-Structure.

4.4.4.2 Multiple contrasted constituents

As Avgustinova and Oliva (1995) show, the clitic cluster can be preceded by several contrasted constituents. Consider their example in (80). Although the expression *[na chatu] [v létě]* denotes place + time, it does not seem to be possible to argue that it is a similar case to the spatio-temporal adverbials in (78c) – the two PPs are contrasted with two independent PPs in the previous clause. However, even if such analysis were possible in this case, it is definitely impossible for the contrasted constituents in (82).

(80) [V našem pražském bytě] *jsme* příbuzné ze Saarbrückenu o vánocích ještě
In our Praguian apartment aux_{1pl} relatives from Saarbrücken during Christmas still
nějak snesli, ale [na chatu] [v létě] *jsme* *je* raději nepozvali.
somehow bore but to weekend-house in summer aux_{1pl} them_A better not-invited.
'In our Prague apartment, we bore the relatives from Saarbrücken during Christmas time somehow, but we decided it was better not to invite them to our weekend house in summer.'
[Avgustinova and Oliva 1995 (59)]

According to Avgustinova and Oliva (1995), the nature of the multiple constituents is rather restricted – the constituents must satisfy all the conditions in (81).

(81) Conditions on multiple contrasted constituents in 1P according to Avgustinova and Oliva (1995, pp. 36/37; my wording):

1. All the constituents must be adverbials.
2. Either all the constituents must be adjuncts or they must all be complements.
3. If the constituents are complements, they must form a single "semantic" modification – being of the same type, express path/period or stage (§4.4.4.1).

However, as the sentences in (82) show, the constraint is not correct. For example, in (82a), *Petra* is not an adverbial; in (82c) *Petra* is a complement while *na Smíchově* is an adjunct.

(82) a. *Context: I am a member of a travel-committee, reviewing requests for travel to different conferences. Petr requested France and Australia, Martin Hungary, etc. The money is limited so not everybody can go everywhere*

[Petra] [do Francie] *bych* ještě poslal, ale Martina do Maďarska ani náhodou.
Petr$_A$ to France would$_{1sg}$ still send but Martin$_A$ to Hungary not-even by-accident
'I would send Petr$_C$ to France$_C$, but never Martin$_C$ to Hungary$_C$.'

b. [Petrovi] [do Francie] *bych* *to* ještě poslal, ale Martinovi do Maďarska ani náhodou.
to Peter to France would$_{1sg}$ it$_A$ still send but to Martin to Hungary not-even by-accident
'I would send it to Peter$_C$ to France$_C$, but never to Martin$_C$ to Hungary$_C$.'

c. [Petra] [na Smíchově] *jsem* viděl, ale Martina na Václaváku ne.
Petr at Smíchov aux$_{1sg}$ saw but Honza$_A$ at Wenceslas Square not
'I saw Petr$_C$ at Smíchov$_C$, but I did not see Honza at Wenceslas Square.'

d. [Všechny sny] [najednou] *se* *mu* určitě nesplní.
All dreams at-once refl$_A$ him$_D$ definitely not-fulfill.
'There is no way all his dreams will come true at the same time.'

The restriction on possible multiple contrasted constituents preceding clitics appears to be again a restriction on fronting. Any multiple fronted constituents can be followed by clitics. In §3.4.4, we left the problem of restriction on multiple fronted constituents open, but in our opinion, the restrictions are rather of pragmatic than syntactic nature. Certain sentences with multiple frontings (and thus sentences with clitics preceded by multiple constituents) *seem* impossible simply because it is harder to imagine a context for them.

4.4.4.3 Splitting a fronted expression

There is another option: the clitic can split the string of multiple fronted elements and follow only the first contrasted constituent. In fact, this is the more common case. The sentences with both contrasted constituents preceding the clitic seem to put more stress on the contrast.

(83) [Petra] *bych* [do Francie] ještě poslal, ale Honzu do Maďarska ani náhodou.
Peter$_A$ would$_{1sg}$ to France still send but Honza$_A$ to Hungary not-even by-accident
'I would probably send Peter$_C$ to France$_C$, but never Honza$_C$ to Hungary$_C$.'

As (84) show, this option is available only for multiple short-fronting. A long-fronted expressions must stay continuous.[54]

(84) a. [Petra do Francie] poslal hned.
Petr$_A$ to France sent immediately
'He sent Petr to France$_C$ immediately.

b. [Petra] [do Francie] *bych*$_0$ poslal hned.
Petr$_A$ to France would$_{1sg}$ sent immediately
'I would send Petr to France$_C$ immediately.

c. [Petra] *bych*$_0$ [do Francie] poslal hned.
Petr$_A$ would$_{1sg}$ to France sent immediately
'I would send Petr to France$_C$ immediately.

d. [Petra do Francie] *si*$_1$ myslím$_1$, že Martin pošle hned.
Petr$_A$ to France refl$_D$ think$_{1sg}$ that Martin will-send immediately
'I think Martin will send Petr to France$_C$ immediately.

e. ?* [Petra] *si*$_1$ [do Francie] myslím$_1$, že Martin pošle hned.
Petr$_A$ refl$_D$ to France think$_{1sg}$ that Martin will-send immediately
'I think Martin will send Petr to France$_C$ immediately.

4.4.4.4 Summary of §4.4.4

In sentences with multiple fronting (stage/period/path adverbials and multiple contrastive themes), the main clitic cluster can either follow the whole fronted expression or the first, possibly partial, constituent.

In the case of multiple contrasted constituents, the contrast seems to be stronger when the whole fronted expression preceded the clitics than when only the first fronted constituent does and the

[54] This restriction is similar the similar restriction to multiple wh-long-movement discussed by Lenertová (2001, p. 297). However, we disagree with her conclusion that the position of clitics in short multiple wh-movement determines whether single versus multiple pair readings is possible.

others are marked for contrast prosodically. Some multiply fronted constituents are more ready to appear in such position than others (e.g., adjuncts), but in general the constraints seem to be pragmatic rather than syntactic.

4.4.5 Analysis, Version 1

The above data can be analyzed as clitics following two possible anchors:

1. the first constituent
2. the fronted expression

Because most sentences contain a fronted expression and because most fronted expressions consists of a single constituent (possibly partial), in most cases, these two choices results in the same clitic position. There is no fronting in rheme-only sentences in objective ordering and clitics simply follow the first constituent. On the other hand, in sentences with multiple fronting, there are two possible anchors – either the first fronted constituent or the whole fronted expression. We will revisit this view below.

4.4.6 After a Complementizer/Discourse particle

Clitics cannot follow coordinating conjunctions like *a* 'and', *i* 'even and', and they also cannot follow *ale* 'but'. However, in the case of subordinate conjunctions (e.g., *že* 'that', *jenže* 'but', *protože* 'because', *jestli* 'if'), there is a choice. One possibility is that clitics are adjacent to the complementizer as in (85a). The other possibility is that clitics are separated from the complementizer by the theme proper (usually contrasted) as in (85b), or, in subjective ordering, by rheme proper (with a proper intonation and in a proper context *Petr* in (85b) can be interpreted as either.)

(85) a. Helena říkala, že *se* Petr odstěhoval.
Helena said that refl$_A$ Petr moved
'Helena said that Petr had moved.' [Fried 1994 (9a)]

b. Helena říkala, že [Petr] *se* odstěhoval.
Helena said that Petr refl$_A$ moved
'Helena said, Petr$_C$ had moved.' [Fried 1994 (9b)]

The examples in (86) show that the constituent can be rather complex. As Uhlířová (1987, p. 91) mentiones, the complementizer can be even followed by a parenthetical as in (87).

(86) a. . . . nějaký ženský hlas *mi* sdělil, že [paní inženýrka ani pan inženýr] *se* zatím
. . . some female voice me$_D$ told that Ms. engineer$_F$ nor Mr. engineer$_M$ refl$_A$ so-far
domů nevrátili.
home not-returned

'. . . some female voice told me that neither Ms. engineer nor Mr. engineer have come back home yet.' [syn5]

b. Grégr včera sdělil, že [o přechodném období při liberalizaci energetického
Grégr yesterday said that about transitional period prep liberalization energy
trhu] *se* s EU stále jedná a ...
market refl$_A$ with E.U. still negotiate and ...

'Grégr said yesterday that the transitional period in energy markets liberalization$_C$ is still being negotiated with E.U. and ...' [syn5]

(87) . . . protože, [jak známo,] [mnozí lidé] *se* do konce života nenaučí správně mluvit . . .
. . . because as known many people refl$_A$ till end of-life not-learn correctly speak . . .

Usually all the cited examples use the complementizer *že* 'that'. Veselovská (1995, §9.3.5) even explicitly states that sentences with other complementizer, such as (88) with *jestli* 'whether' are ungrammatical (?# judgment is mine):

(88) ?# Ptal se, jestli [Petr] *mu* *to* nedal.
asked refl$_A$ whether Petr him$_D$ it$_A$ not-gave

'He asked whether Peter gave it to him.' [Veselovská 1995]

However, the sentence seems more pragmatically odd (in an out-of-the-blue context) than ungrammatical. A similar sentence in (89) is fine. And so are the sentences in (90) taken from corpora. Therefore, we can conclude that the construction is not limited to *že* 'that' but is possible with other complementizers as well.

(89) Ptal se, jestli [třeba Petr] *by* *mu* *to* nedal.
asked refl$_A$ whether perhaps Petr would$_3$ him$_D$ it$_A$ not-gave

'He asked whether perhaps Peter would not give it to him.'

(90) a. Nepamatuju se, jestli [tenhleten] *se* z toho vyvlíknul, nebo ne.
not-remember$_{1sg}$ refl$_A$ whether this-one refl$_A$ from that backed-out or not

'I do not remember if this one managed to back out of it.' [syn5]

b. Nejsem překvapen, že *se* na to ptáte, protože [Kanaďané] *mi* dávají tuhle
not-am surprised that refl$_A$ prep it ask because Canadians me$_D$ give this
otázku pořád dokola.
question all around

'I am not surprised you ask me about this because the Canadians ask me that question all the time.' [syn6]

Fried (1994, ftn. 7) also notices that matrix sentences introduced with a discourse particle pattern similarly:

(91) a. Vždyť se Petr odstěhoval!
Particle refl_A Petr moved
'But Petr moved away (so how can you be surprised that Helena is upset)!' [Fried 1994 p. 160]

b. Vždyť [Petr] se odstěhoval!
Particle Petr refl_A moved
'But Petr_C moved away (why are you therefore counting on his help?)!'[Fried 1994 p. 160]

The prevalence of the two constructions is hard to measure exactly with the current state of corpora annotation and search tools. However the numbers in (92) can give a rough idea, showing at least that neither of them is rare (the opposite of what Veselovská (1995, §4.6) claims).

(92) a. ... *že* 'that' *se* 'refl_A' noun ... – about 10,000 occurrences

b. ... *že* 'that' noun *se* 'refl_A' ... – about 6,000 occurrences

4.4.6.1 Verbs

Uhlířová (1987, p. 89) claims that a verb cannot occur between the complementizer and the clitic cluster. Veselovská (1995, §4.6) argues similarly, based on example in (93). However, their claim is simply not true. First, insertion of a constituent between the complementizer and the clitic cluster is used to express certain Information Structure of the clause, thus the context is extremely important. That the sentence fragment in (93) seems wrong out of the blue, does not mean it would not be judged as appropriate in some other context. The real sentences in (94) indeed show that the verb (incl. infinitive, finite verb, past participle) *can* occur between the complementizer and the clitic cluster.

(93) * ... že nedal *by mu to.* (judgment by Veselovská)
... that not-gave would_3 him_D it_A
'... that he would not give_C it to him.' [Veselovská 1995 (§4.6)]

(94) a. Petrová uvedla, že [jednat] *by se* mělo koncem druhého
Petrová put-forward that $\text{negotiate}_{\text{inf}}$ would_3 refl_A $\text{should}_{p.part}$ at-the-end second
zářijového týdne.
September week.
'Petrova put forward that the negotiation should take place in the end of the second week in September.' [syn6]

b. *Context: A and B do not share a common language. A: I have good wine at home. B: I don't drink.*

Špičkovou pantomimou jí vysvětlil, že [pil] *by* on.
perfect mime her$_D$ explained that drank would$_3$ he
'He explained miming perfectly, that HE$_R$ would drink$_C$.' [syn5]

c. Petr říkal, že [prodá] *mu to* určitě, a možná i dá.
Petr said that sells him$_D$ it$_A$ definitely and maybe even gives
'Petr said he will definitely sell$_C$ it to him it and maybe he will even give it to him.'

d. Nemluvě o tom, že [stačilo] *si* jednou za čas pustit zprávy
not-talking about that$_{loc}$ that$_{comp}$ was-enough refl$_D$ once prep time turn-on$_{inf}$ news
na Nově, aby mi došlo, že ..
on Nova,
'And it goes without saying that it was enough to turn on the Nova news sometime and it would come to my mind that ...' [syn5]

As (95) illustrates, the past participle can occur in this position only alone, which is similar to the restriction on past participle in main clauses discussed in §4.4.2.

(95) * Špičkovou pantomimou jí vysvětlil, že [pil víno] *by* on.
perfect mime her$_D$ explained that drank wine would$_3$ he
Intended: 'He explained miming perfectly, that HE$_R$ would drink wine$_C$.'

4.4.6.2 Multiple constituents

While all the linguistic sources available to us (e.g. Daneš et al. 1987, p. 619, Uhlířová 1987, p. 89, Veselovská 1995, §4.6) claim that there can be only one constituent between the complementizer and the clitic cluster, in fact the data show that there can be more of them as long as they are one of the following: path/period adverbials (96), stage adverbials (97) or they are all part of the contrastive theme (98). These constructions are analogous to the similar constructions in the matrix sentences, discussed above.

(96) a. Psali, že [od pátku] [do neděle] *se* zde narodilo pět miminek.
wrote$_{3pl}$ that from Friday till Sunday refl$_A$ here born five babies.
'They wrote from Friday to Sunday, five babies were born here.'

(97) a. Nechci, před vámi tajit pane Holmesi, že [u nás] [ve vyšetrovacím oddělení]
want$_{1sg}$ before you conceal Mr. Holmes that prep us in investigative department
si myslíme, že ...
refl$_D$ think$_{1pl}$ that ...

'I do not want to conceal from you, Mr. Holmes, that at our investigative department we think that ...' [syn5]

b. ... že [vocuď] [hned] *by* šel tamhle, ...
... that from-here righ-away would$_3$ went over-there ...

'... that from here, he would go there right away ...' [Oral2006]

(98) a. Helena říkala, že [Petr] [Pavlovi] *by* *to* dal, ale Honza Marii ne.
Helena said that Petr$_A$ Pavel$_D$ would$_3$ it gave but Honza Marie$_D$ not.

'Helena said that Petr$_C$ would give it to Pavel$_C$ but Honza$_C$ would not to Marie$_C$.'

b. Helena říkala, že [Honzu] [do Francie] *by* poslali, ale ...
Helena said that Honza$_A$ to France would$_3$ send but ...

'Helena said that they would send Honza$_C$ to France$_C$ but ...'

c. Předpokládá *se*, že [ropa] [do tuzemska] *by* mohla začít proudit již dnes.
assumes refl$_A$, that oil to inland would$_3$ could start$_{inf}$ flow$_{inf}$ already today

'It is assumed that oil could start to flow to our country already today.' [syn6]

4.4.6.3 Partial constituents

As (99) shows, the complementizer can be followed by various partial constituents parallel to the cases in §4.4.3 – compare examples (99) with the corresponding examples above: (99a) with (67b), (99b) with (73), (99c) with (75a), (99d) with (74b).

(99) a. (Partial clausal constituent)

Helena říkala, že [pohlídat děti] *si* troufne [Novákům].
Helena said that watch$_{inf}$ children refl$_D$ dare Nováks$_D$

'Helena said that she dares to watch children$_C$ FOR NOVÁKS$_R$.'

b. (Split constituent, Verbal head first)

Helena říkala, že [pohlídat děti] *si* [Novákům] troufne. (ale opravit auto ne.)
Helena said that watch$_{inf}$ children$_A$ refl$_D$ Nováks$_D$ dare$_{1sg}$

'Helena said that she DARES$_R$ to watch children$_C$ for Nováks. (but not to repair their car)'

c. (Split constituent, Nominal head first)

Helena říkala, že [fotky] *ti* [nějaký] určitě pošle, ale ...
Helena said that photos$_A$ you$_D$ some$_A$ definitely send, but ...

'Helena said that she would send you some photos$_C$, but ...'

d. (Split constituent, Verbal head later)

Helena říkala, že [Novákům] *si* [děti pohlídat] troufne. (ale Císlerům ne.)
Helena said that Nováks$_D$ refl$_D$ children$_A$ watch$_{\mathsf{inf}}$ dare$_{1sg}$

'Helena said that she DARE$_R$ to watch children for Nováks$_C$. (but not for Císlers)'

4.4.6.4 Aby

As (100) shows, the main clitic cluster surprisingly does not have to be adjacent to the contraction of complementizer with the conditional (*abychom*, *kdybychom*, etc. see §4.3.4). Although in most cases it is. This would mean that forms of *aby* are sometimes treated as contractions, i.e., the complementizer *aby* followed by an auxiliary clitic, and sometimes as a declined one-word complementizer similar to those in certain Germanic dialects (see for example, Bayer 1984; Kathol 2000*b*, and the references cited therein).

(100) a. Chceme, aby [stát] *se* k těmto závazkům přihlásil a vyplatil *nám*
want$_{1pl}$ that-should state refl$_A$ to these obligations acknowledged and paid us$_D$
ho například později v rámci státního rozpočtu.
him for-example later in scope state budget

'We want the state$_C$ to acknowledge these obligations and pay it to use later as, for example, a part of the budget.' [syn5]

b. Spíš chtějí, abych [já] *se* svěřoval jim.
rather want$_{1pl}$ that-should I refl$_A$ confided them$_D$

'They would prefer that I$_C$ confide to THEM$_R$' [syn5]

4.4.7 Analysis, Version 2

It is common to analyze sentences with complementizers in the following way: the complementizers that are able to host clitics occupy the first position (1P) and in addition, there is an optional position that can be occupied by a contrasted/stressed constituent. This route is followed, for example, by Veselovská (1995, §4.6) and Meyer (2005, p. 91).[55] However, such analysis is losing generalizations. As we have shown, the set of possible expressions between the complementizer and the clitic cluster is the same as the set of possible expressions occupying 1P in matrix sentences under the same conditions: it can contain partial constituents or multiple constituents, and when it contains a past participle it cannot contain anything else. We have also shown that the alleged restrictions on the so-called optional position (no verbs, no multiple constituent) that would differentiate it from the

[55] Svoboda (2000) puts complementizers into a position before 1P (*initial* and *pre-initial* field in his terminology). However, as far as we know, he does not provide any reasons for that.

pre-clitic position in matrix sentences, in fact, do not exist. Thus in this view, one has to restate the conditions on 1P for the new optional slot.

In §4.4.5 above, we concluded that in matrix sentences, 1P can be either the first constituent or the fronted expression. One way how to interpret the data in the previous section is that (a) the clitics are positioned relative to the whole complementized sentence ($\bar{\text{S}}$ or CP), and that (b) there is a third possible anchor for clitic blending the previous two cases: 1P can be also the first fronted constituent.

Consider the example in (101) which illustrates all three possibilities. The clitic might be placed after the first constituent (i.e., the complementizer), after the first fronted expression or after all fronted expressions (which is the actual attested case).

(101) Předpokládá *se*, že (*by*) ropa (*by*) do tuzemska *by* mohla začít proudit již dnes.
assumes refl_A, that oil to inland would_3 could $\text{start}_{\text{inf}}$ flow_{inf} already today
'It is assumed that oil could start to flow to our country already today.' [syn6]

In fact, a similar situation can be found in matrix sentences when a multiple fronted expression is preceded by certain particles such as *vždyť* (c.f. (91)):

(102) Vždyť (*by*) ropa (*by*) do tuzemska *by* mohla začít proudit již dnes.
Particle oil to inland would_3 could $\text{start}_{\text{inf}}$ flow_{inf} already today
'But oil could start to flow to our country already today.'

However, examples such as these are rather rare. In the majority of cases, all the three possibilities come to one. The reason is that:

1. Usually one and only one constituent is fronted; exceptions are rheme-only sentences where nothing is fronted, and multiple frontings.
2. Fronted expressions are usually initial, exceptions are complementizers and the infrequent cases of particles such as *vždyť*.

In example (103), the position of the clitic can be analyzed in either of the three ways: it follows the first constituent, all fronted expressions or the first fronted constituent.

(103) Hejtmana *by* navrhla ODS.
local-governor_A would_3 nominated ODS.
'The governor would be nominated by ODS.' [syn6]

4.4.8 Summary of §4.4

In this section, we have shown that while in a typical sentence the main clitic cluster follows the first clausal constituent, this is not the case in general. Clitics can be positioned in respect to three anchors:

1. the first constituent – this may be the first fronted constituent, the first constituent in rheme-only sentences without fronting, or the complementizer;
2. the first fronted constituent (possibly preceded by a complementizer)
3. the whole fronted expression

In an embedded clause with a complementizer, the clitics are positioned relative to the whole complementized clause. The constituents are partial in case of split-fronting, otherwise they are full constituents. In majority of cases, all these three possibilities come to one.

4.5 Morpholexical ordering

As mentioned in §4.2, sentential clitics not only have a fixed position relative to the rest of the clause; they also have a relatively fixed order relative to one another. A clitic cluster can be quite complex: clitics governed by different verbs (or even adjectives, etc.) can cluster together in one place due to clitic climbing (see §4.6). In the present section, we describe a constraint which orders clitics based on their morpholexical properties, so that certain clitics, and clitics in certain forms, must occur before certain other clitics. We present data and constraints that hold for Czech, but similar constraints are valid in other Slavic languages as well; for a comparison see, for example, (Franks and King 2000).

The examples in (104) illustrate the basic point: the order of clitics in (104a), reflexive – dative – accusative, is grammatical, while the order in (104b) is not.

(104) a. Martin *se*$_1$ *ti*$_2$ *ho*$_2$ nakonec rozhodl$_1$ koupit$_2$.
Martin$_N$ refl$_A$ you$_{sgD}$ him$_A$ finally decided buy$_{\mathsf{inf}}$
'Martin finally decided to BUY$_R$ it for you.'

b. * Martin *se*$_1$ *ho*$_2$ *ti*$_2$ nakonec rozhodl$_1$ koupit$_2$.
Martin$_N$ refl$_A$ him$_A$ you$_{sgD}$ finally decided buy$_{\mathsf{inf}}$

It is important to note that, for the relative acceptability of the sentences in (104), it is irrelevant whether or not the positioning of the verbs governing the relevant clitics (*rozhodl* 'decided' and *koupit* ‘buy$_{\text{inf}}$’) yields more or less discontinuous phrases. Consider the various possibilities in (105): the examples differ in their topic/focus structure, sometimes in very subtle ways, but all of them are grammatical.

(105) a. Martin *se*$_1$ *ti*$_2$ *ho*$_2$ koupit$_2$ nakonec rozhodl$_1$. (Ale Eva ještě váhá)
Martin$_N$ refl$_A$ you$_{sgD}$ him$_A$ buy$_{\text{inf}}$ finally decided
‘Martin finally DECIDED$_R$ to buy it for you. (But Eva is still hesitating.)’

b. Koupit$_2$ *se*$_1$ *ti*$_2$ *ho*$_2$ nakonec rozhodl$_1$ Martin.
buy$_{\text{inf}}$ refl$_A$ you$_{sgD}$ it$_A$ finally decided Martin
‘MARTIN$_R$ finally decided to buy$_C$ it for you.’

c. Rozhodl$_1$ *se*$_1$ *ti*$_2$ *ho*$_2$ nakonec koupit$_2$ Martin.
decided refl$_A$ you$_{sgD}$ it$_A$ finally buy$_{\text{inf}}$ Martin
‘MARTIN$_R$ finally decided$_C$ to buy it for you.’

The examples in (104) and (105) show that reflexives (the accusative reflexive *se* and the dative reflexive *si*) precede nonreflexive dative pronouns (like *jí* ‘her$_D$’, *mi* ‘me$_D$’, etc.), which in turn precede nonreflexive accusative pronouns (such as *ho* ‘him$_A$’, similarly *mě* ‘me$_A$’, etc.). Schematically then:

(106) reflexives < nonreflexive dative < nonreflexive accusative[56]

4.5.1 Reflexives

Only one of the four reflexive clitics (accusative, dative and contractions – see §4.3.3 above), can occur in the same clitic cluster, as (108) shows.[57] For cases of reflexives governed by different heads see §4.6.1.

(108) a. * Smál *se* *si*.
laughed refl$_A$ refl$_D$

[56] Slovak, Slovenian and Sorbian follow the same pattern, but Serbo-Croatian requires reflexives to follow accusatives.

[57] In this respect, Czech differs from Bulgarian, a South Slavic language, where only identical reflexives cannot co-occur in the same cluster.

(107) Barabanchikât si se usmixva.
drummer.the refl$_D$ refl$_A$ smiles
‘The drummer smiles at himself.’ [Rivero 2005 (27)]

b. Smál *se* (sám) sobě.
laughed refl$_A$ (alone) refl$_D$
'He laughed to himself.'

4.5.2 Datives

The situation with dative clitics is slightly more complicated, in that the ordering shown in (106) above holds only for complement dative clitics. There are two other types of nonreflexive dative clitics: ethical dative clitics and adjunct clitics. Second-person ethical dative clitics roughly corresponding to English phrase *you know* and the like.[58] Adjunct dative is used for somebody who benefits from or is affected by a process, in examples below, we translate it as *for me/her/...*

Ethical dative clitics can follow a reflexive like any other dative clitic, but they can also precede it. In (109a), the ethical dative *ti* follows the reflexive *se*, while in (109b), it precedes the reflexive. Some speakers prefer them to precede the complement datives (110a, 110b), but some allow also the opposite order (110c). It is necessary to mention that there is a great variety in speakers' constraints on the order of the ethical-dative clitics relative to the other dative clitics. However, all speakers perceive violations of their constraints on ethical dative placement as much less disturbing than violations of other constraints: e.g., violations of the relative ordering of dative and accusative clitics.

(109) a. On *se* *ti* vůbec nebál.
he refl$_A$ you$_D$ at-all not-scared
'You know, he wasn't scared at all.'

b. On *ti* *se* vůbec nebál.
he you$_D$ refl$_A$ at-all not-scared
'You know, he wasn't scared at all.'

(110) a. On *se* *ti* *jí* ani nepředstavil.
he refl$_A$ you$_D$ her$_D$ even not-introduced
'You know, he did not even introduce himself to her.'

b. On *ti* *se* *jí* ani nepředstavil.
he you$_D$ refl$_A$ her$_D$ even not-introduced
'You know, he did not even introduce himself to her.'

[58] As Rosen (2001) points out, in addition to the second person clitics *ti* 'you$_{Sg\,D}$' and *vám* 'you$_{Pl\,D}$', there is also a third-person plural ethical dative clitic *jim* 'them$_D$', formerly used in polite address. Such usage is now obsolete, and the second person plural pronoun is used instead.

c. ? On *se* *jí* *ti* ani nepředstavil.
he refl$_A$ her$_D$ you$_D$ even not-introduced
'You know, he did not even introduce himself to her.'

The position of adjunct datives is after ethical datives/reflexives, as seen in (111), and before complement datives, as seen in (112):

(111) a. Zbláznil *se* *jí* manžel.
Went-crazy refl$_A$ her$_D$ husband
'Her husband went crazy.' (Lit: The husband went crazy to her.)

b. * Zbláznil *jí* *se* manžel.
Went-crazy her$_D$ refl$_A$ husband

(112) a. On *se* *mi* *jí* ani nepředstavil.
He refl$_A$ me$_D$ her$_D$ even not-introduced
'He did not even introduce himself to her, for me.'
?'He did not even introduce himself to me, for her.'

b. On *se* *jí* *mi* ani nepředstavil.
He refl$_A$ her$_D$ me$_D$ even not-introduced
'He did not even introduce himself to me, for her.'
?'He did not even introduce himself to her, for me.'

4.5.3 Genitives

Although it is clear that genitive clitics occur close to the right edge of the clitic cluster, following for example reflexives (113a) or datives (113b), their position relative to accusative clitics is not entirely clear, as discussed for example by Franks and King (2000). One of the reasons is that sentences containing both accusative and genitive clitics are rather rare. Mostly, the genitive clitic is extracted from a numeral expression or an expression of amount (sometimes called numerative or partitive). The syn2005 corpus contains sentences exhibiting both orders, although a genitive following an accusative, e.g., (113c), is more frequent than a genitive preceding an accusative, e.g., (113d). The judgments are largely speaker dependent, some speakers judging both orders as incorrect or marginal. I prefer genitive following accusative, although in certain cases both possibilities seem equally acceptable to me, for example (113e) and (113f).

(113) a. Nemohl *jsem* *se* *jí* nabažit.
not-could aux$_{1sg}$ refl$_A$ her$_G$ get-tired-of
'I could not get tired of her.'

b. On *se* *ti* *mě* nebál.
He refl$_A$ you$_D$ me$_G$ not-scared
'You know, he wasn't scared of me.'

c. Kontaktovalo *nás* *jich* asi osm, ale ...
Contacted us$_A$ them$_G$ about eight but
'About eight of them [sport clubs] contacted us, but ... ' [syn5]

d. Ano, třicet *jich* *nás* přišlo zachránit, ...
Yes thirty them$_G$ us$_A$ came rescue$_{inf}$ ...
'Yes, thirty of them [scouts] came to our rescue, ...' [syn5]

e. Napadá *mě* *jich* *tu* vždycky spousta.
come-upon me$_A$ of-them$_G$ here always a-lot
'I always come upon a lot of them [e.g. jokes] here.'

f. Napadá *jich* *mě* *tu* vždycky spousta.

4.5.4 Auxiliaries

As explained in §4.3.4, some forms of the auxiliary verb *být* 'to be' (the past auxiliary, conditional auxiliary, non-negative passive auxiliary and non-negative copulas) are, or can be, clitics. They occur at the beginning of the clitic cluster, as for example in (114). Unsurprisingly, when the conditional auxiliary is reanalyzed as a conditional particle *by* + (past tense) auxiliary, the particle comes before the auxiliary, as (114c) shows.

(114) a. Martin *by* *se* *jí* *ho* nakonec rozhodl koupit.
Martin would$_3$ refl$_A$ her$_D$ him$_A$ finally decided to-buy
'Martin would decide to buy it for her at the end.'

b. Seznámila *jsem* *se* se zajímavým klukem.
Met aux$_{1sg}$ refl$_A$ with interesting boy
'I met an exciting boy.' [ksk]

c. Mohli *by* *jsme* *si* k tomu sehnat i různé věci a potřeby.
could would aux$_{1pl}$ refl$_D$ for that get even various things and requisities
'We could even get various things and requisities for that.' [ksk]

4.5.5 *to*

When clitic, *to* 'it$_A$' follows accusative/genitive personal pronouns[59] as (115) shows. In most cases it precedes *však, prý, prej, už* and the other inconstant clitics in (51) – see (116). In the corpus

[59] Recall, that *to* is a demonstrative pronoun, accusative singular neuter form of *ten*, with the meaning roughly as *this* and *that* without expressing closeness/distance. Usually, English personal pronoun *it* is the closest translation.

syn2005, sequences ⟨constant clitic⟩ + *to* + *však*|*prý*|*prej*|*už* are 25 times more frequent than sequences ⟨constant clitic⟩ + *však*|*prý*|*prej*|*už* + *to* (we require the sequences to start with a constant clitic to exclude most of the non-clitic uses of *však*, *prý* etc.).

(115) a. Šána kouká do země, jako *by* *se* *ho* *to* netýkalo.
Šána looks into ground as would refl$_A$ him$_G$ it$_A$ not-affected.
'Šána looks into the ground as if he weren't involved.' [syn5]

b. * Šána kouká do země, jako *by se to ho* netýkalo.

(116) Stalo *se* *mi* *to* *už* několikrát a vím, že ...
happened refl$_A$ me$_D$ it$_A$ already several-times and know$_{1sg}$ that ...
'It has happened to me several times and I know that ...' [syn5]

4.5.6 *však, prý, prej, ale, už*

Clitic *však* 'however/though' can occur at the beginning or preferably at the end of the clitic cluster following *to* 'it', as shown by the examples in (117), or the real examples in (118).

(117) a. Opravit *však* *jsem* *se* *mu* *to* včera snažil marně.
repair however aux$_{1sg}$ refl$_A$ him$_D$ it$_A$ yesterday tried fruitlessly
'However, I tried to repair it yesterday without success.'

b. Opravit *jsem se mu to však* včera snažil marně.

(118) a. V osobní komunikaci z očí do očí *by* *se* *vám* *to* *však*
In personal communication from eyes to eyes would$_3$ refl$_A$ you$_{plD}$ it$_A$ however
nemuselo podařit.
may-not succeed$_{\mathsf{inf}}$
'In personal eye to eye communication, you would not necessary succeed though.' [syn5]

b. Vůbec *se* *jí* *však* nelíbilo, když *jsem* *jí* donesla učení na
Not-at-all refl$_A$ her$_D$ however not-liked when aux$_{1sg}$ her$_D$ brought studying to
doplnění.
catch-up.
'She did not like at all though when I brought her study materials to catch up.' [ksk]

c. Naštěstí *však* *se* *mu* *to* nikdy nepodařilo a ...
Luckily however refl$_A$ him$_D$ it$_A$ never not-succeeded and ...
'Luckily he never succeeded though.' [syn5]

Some speakers allow *však* to occur anywhere within the clitic cluster, see (119) or (120). Other speakers judge these sentences as marginally acceptable, or even ungrammatical. The syn2005 corpus contains nearly 150,000 occurrences of *však*. 64-80% of them are not adjacent to a clitic; 19-33% occur at the end of the clitic cluster, around 1% occur at the beginning of the clitic cluster, occurrence in the middle of a clitic cluster is close to 0%.[60]

(119) a. Opravit *jsem však se mu to* včera snažil marně.

b. Opravit *jsem se však mu to* včera snažil marně.

c. Opravit *jsem se mu však to* včera snažil marně.

(120) a. Právě proto *jsem se však mu* snažil co nejvíce vyhnout.
Just therefore aux$_{1sg}$ refl$_A$ however him$_D$ tried what most avoid$_{\mathsf{inf}}$
'Exactly because of that, I tried to avoid him as much as possible.' [syn5]

b. Těsto pořádně promícháme, aniž *bychom však ho* silně hnětli.
Dough thoroughly mix without would$_{1pl}$ however him$_A$ hard kneaded
'We mix the dough thoroughly; however without kneading it hard.' [syn5]

A similar distribution can be observed for *ale*, also an inconstant clitic, but much less formal and much more frequently used as a non-clitic. Also *prý/prej* 'allegedly', see (121), and *už* 'already' also occur mostly at the beginning or the end of the cluster, rarely internally.

(121) a. Mluvil jsem s Rosensteinem a ten *mi* oznámil, že *jsem si tě*
talked aux$_{1sg}$ with Rosenstein and that me$_D$ informed that aux$_{1sg}$ refl$_D$ you$_A$
prý najal.
allegedly hired
'I talked with Rosensteinem and he told me, that allegedly I had hired you.' [syn5]

b. Mohlo$_1$ *by*$_0$ *se*$_2$ *to*$_2$ *prý*$_0$ snadno stát$_2$.
Could would$_3$ refl$_A$ it allegedly easily happen$_{\mathsf{inf}}$.
'It could allegedly easily happen.' [syn5]

c. Vaří *nám tu* zatím dobře, ale *prý se to* má zhoršit.
Cook$_{3pl}$ us$_D$ here so-far well but allegedly refl$_A$ it$_A$ should get-worse$_{\mathsf{inf}}$
'They cook for us well so far, but it should allegedly get worse.' [ksk]

d. Tím, že *jsem mu* přinesl celý rukopis, udobřil *jsem prý si*
By-that that aux$_{1sg}$ him$_D$ brought whole manuscript, reconciled aux$_{1sg}$ allegedly refl$_D$
ho.
him$_A$
'Allegedly, I reconciled with him by bringing the whole manuscript.' [syn0]

[60] The frequencies are provided as ranges because the corpus does not contain information about clitic-hood, and even the morphological and lexical information that could provide partial clues contains errors. The lower ends of the ranges are obtained by considering only unambiguous tokens as clitics (*bych* 'would$_{1sg}$, but not *se* 'refl'/preposition or *nás* 'us$_{G/A}$' an inconstant clitic), the higher ends by considering all tokens that can potential by clitics.

The relative position of these clitics to each other is probably mostly free. Although, based on the frequency in the corpus the order *však* < *prý/prej* < *už* seems to be preferred,[61] all consulted speakers judged any possible variations as equally acceptable.

(122) a. No, ale ve Španělsku *se* *prý* *už* opalují.
Well, but in Spain refl$_A$ allegedly already sun-bathe$_{3pl}$
'Well, but they say that it is already possible to sunbathe in Spain.' [ksk]

b. No, ale ve Španělsku se *už prý* opalují.

c. Daří *se* *mu* *to* *však* *prý* jen proto, že připravuje ...
succeeds refl$_A$ him$_D$ it$_A$ though allegedly only because prepares ...
'However he is allegedly successful only because he prepares ...' [syn5]

d. Na Žižkově *však* *prý* *už* podepsal smlouvu s platností od července 2004.
at Žižkov but allegedly already signed agreement with effectiveness from July 2004
'But allegedly he already signed an agreement at Žižkov effective July 2004.' [syn5]

e. *Její muž zatím během kampaně utratil okolo 48 mil. dolarů (zhruba 1,7 miliardy Kč), – 'Her husband spent about $48 million (roughly 1.7 billion CZK) during the campaign sofar,*

disponuje *však* *už* *prý* 70miliónovým fondem a ...
dispose however already allegedly 70-million fund and ...
however, he has allegedly 70-million fund at his disposal and ...' [syn5]

4.5.7 Summary of §4.5

In Czech, similarly as in other languages, clitics within a clitic cluster are ordered according to their morpholexical features.

(123) auxiliaries < reflexives < adjunct dative < complement dative < < accusative/genitive < *to*

Genitive usually follows accusative. In addition,

- ethical dative occurs anywhere after the position of auxiliaries and before the position of complement datives (or accusatives for some speakers);

[61] The corpus syn2005 contains only 3 sentences containing all 3 words in a sequence, two of them in (122), the syn2000 contains another 6 such sentences, ksk or pmk none (this is not surprising since one of them - *však* is quite infrequent in Common Czech). However taken by pairs (for syn2005), *vsak* preceds *prý/prej* in 78% cases, *prej* < *už* 68%, *vsak* < *už* 86%. It is worth noting, that some of the cases may include non-clitic usages of these words.

- other clitics, e.g., *tu*, *však*, *prý/prej*, *už*, *ale* follow the position of *to*. *však*, *prý/prej*, *už* can also precede the position of auxiliaries; for some speakers they can even be freely positioned anywhere within the clitic cluster. With a higher but still small frequency, they occur before *to*.

4.6 Clitic Climbing

In a clause, clitics governed by the highest non-clitic governor (usually a non-auxiliary finite verb, see below for other possibilities) obligatorily occur in Wackernagel position – in the main clitic cluster. However, there can be other clitic clusters in the domain of more embedded phrases. Clitics governed by those words can, or even tend, under certain circumstances to occur in the clitic clusters of less embedded governors, possibly in the main one. Within a finite clause, clitics governed by infinitives (124a), adjectives (124b), adverbs, and numerals (124c) can climb up into a higher clitic cluster.

An embedded cluster is within the phrase of its governor either preceding it or immediately following it. See (Toman 2000) for more details. Two adjacent clusters are potentially separated by a prosodic boundary. Thus impossibility to separate two clitics by a boundary means they are in the same cluster.

In this section, we discuss various rules on climbing. Some of them are strict rules and some are merely preferences. Most of the rules are well known, but some modification or corrections, we believe, are original.

(124) a.

Pomoct$_2$	najít$_3$	*by*$_0$	*se*$_1$	*mu*$_2$	*ho*$_3$	určitě	snažil$_1$	i	Martin.
to-help	to-find	would$_3$	refl$_A$	him$_D$	him$_A$	definitely	tried	even	Martin

‘Even Martin would try to help him to find it/him.’

b.

Marie	*mu*$_2$	byla$_1$	věrná$_2$.
Marie	mu$_D$	was	faithful

‘Mary was faithful to him.’ [rosen p.c.]

c.

Martinovi	*se*$_1$	*jich*$_3$	podařilo$_1$	ukrást$_2$	jen	pět$_3$.
Martin$_D$	refl$_A$	of-them$_G$	managed$_{neut.sg}$	steal$_{\mathsf{inf}}$	only	five

‘Martin managed to steal only five of them.’

4.6.1 Co-occurrence constraints

4.6.1.1 Restriction on Identical Clitics

A clitic cluster cannot contain two morphologically identical clitics with different governors. For example, in (125), the embedded clitic *mi* 'me$_D$' cannot climb to the main cluster when another token of that clitic is already there. As Avgustinova and Oliva (1995) show this is not a restriction on two clitics of the same case – a clitic cluster can contain for example two dative clitics (see §4.6.3.3 for more details).

(125) a. Kamila *mi*$_1$ slíbila$_1$ *mi*$_2$ *to*$_2$ vrátit$_2$.
Kamila me$_D$ promissed me$_D$ it$_A$ return$_{\mathsf{inf}}$
'Kamila promised me to return it to me.' [Rosen 2001 (221d)]

b. * Kamila *mi*$_1$ *mi*$_2$ *to*$_2$ slíbila$_1$ vrátit$_2$.
Kamila me$_D$ me$_D$ it$_A$ promissed return$_{\mathsf{inf}}$
[Rosen 2001 (221b)]

c. Kamila *mi*$_2$ *to*$_2$ slíbila$_1$ vrátit$_2$.
Kamila me$_D$ it$_A$ promissed return$_{\mathsf{inf}}$
'Kamila promised to return it to me.' [Rosen 2001 (221c)]

A clitic cluster can contain two identical clitics if they have the same governor, even if they climbed, as (126) shows. However, it is necessary to note that none of the searched corpora contain such a sentence, and some speakers, although accepting (126), suggested replacing the second *ji* by demonstrative *to*.[62]

(126) (Už umí Marie násobilku?)
('Has Marie mastered multiplication (tables)'?)

Ne, ale Martin *by*$_0$ *ji*$_2$ *ji*$_2$ mohl$_1$ naučit$_2$ rychle.
No but Martin would$_3$ her$_A$ her$_A$ could teach$_{\mathsf{inf}}$ fast
'No, but Martin could teach it to her fast.'

A similar constraint was formulated by Rosen (2001, p. 227), however his formulation is unnecessary restrictive: "Two phonologically identical clitics cannot co-occur in a single clitic cluster as a result of clitic climbing." First, his constraint incorrectly rules out the sequence *si si* 'aux$_{2sg}$ refl$_D$', as in (127).

[62] Sentences with two feminine pronouns *ji* sound better than sentences with two, say, masculine pronouns *ho ho*. In our view, this is because the *ji* can be pronounced both with short or long vowel (see §4.3.2) and thus in the sequence *ji ji* the vowels can dissimilate and be pronounced as [jiːjɪ]. This option is not available with other clitics.

(127) a. Ty *[j]si*$_0$ *si*$_1$ pokecal$_1$ ponožku.
You aux$_{2sg}$ refl$_D$ sloshed sock
roughly: 'You spilled on your sock.' [oral2006]

b. Ty *[j]si*$_0$ *si*$_2$ chtěl$_1$ hrát$_2$?
You aux$_{2sg}$ refl$_D$ wanted play$_{\mathsf{inf}}$
'Did you want to play?'

Second, it incorrectly rules out two identical clitics that climb but are governed by the same verb, as in (126), where *ji ji* are governed by an embedded infinitive and climbed to the main clitic cluster. Sentences with multiple identical clitics are not always accepted by speakers, but whether the clitics climbed or not does not influence the acceptability.

4.6.1.2 Haplology of reflexives

While a clitic cluster can contain at most one reflexive (§4.5.1), certain combinations of reflexive clitics can undergo so-called *haplology* – only the more embedded reflexive is realized (see e.g., Avgustinova and Oliva 1995, §2.1.2, Rosen 2001, §7.3).

Note that phonological identity of clitics is neither a necessary nor a sufficient condition for haplology as some authors claim (e.g., Avgustinova 2000, Rosen 2001, p. 229[63]).

First, haplology does not need to occur when clitics are phonologically identical: (127) shows the reflexives can be immediately preceded by *jsi* 'aux$_{2sg}$', usually pronounced as [sɪ], thus homophonous with *si*. *(j)si* + *si* 'aux$_{2sg}$ + refl$_D$' can be replaced by the contraction *sis*, but this is not obligatory. The perception of this repetition is clearly different from the perception of (126) and similar cases of repeated pronominal clitics – all speakers accept examples like (127). Similarly, as in (128), reflexives can be followed by a preposition *se* 'with', a proclitic, which can be homophonous with accusative reflexive *se* when the reflexive proclitizises (§4.2.2). This should not be surprising – as Stemberger (1981, p. 802) documents by examples from various languages, haplology may be present with one affix, but is absent with another, homophonous one.

(128) Ti, co *mě* neznají, | *se* se mnou začnou hádat, . . .
Those what me$_{acc}$ not-know refl$_A$ with me start argue$_{\mathsf{inf}}$. . .
'Those that do not know me start to argue with me, . . . ' [syn6]

Second, haplology can occur when clitics are not phonologically identical: *si* can stand for *se* + a more embedded *si* as (129) shows. The fact that it is the higher *se* and not the lower *si* that is lost,

[63] According to Rosen, the phonological identity is not a necessary condition for haplology to occur, but it is still a sufficient one. We agree with the first part of his claim, but disagree with the second one.

is in line with Stemberger (1981, p. 802) cross-language observation that the morpheme that is lost in haplology dominates the other morpheme. For some reason the opposite haplology (*si* + a more embedded *se*) is not attested as Rosen (2001, p. 232)'s (130) shows.

(129) a. Jan *se*$_1$ bál$_1$ vzít$_2$ *si*$_2$ kravatu.
Jan refl$_A$ was-afraid take$_{\mathsf{inf}}$ refl$_D$ tie
'Jan was afraid to take a tie.'

b. Jan *si*$_2$ bál$_1$ vzít$_2$ kravatu.
Jan refl$_D$ was-afraid take$_{\mathsf{inf}}$ tie
'Jan was afraid to take a tie.' [Rosen 2001 (233) / K. Oliva]

c. *Jan *se*$_1$ bál$_1$ vzít$_2$ kravatu.
Jan refl$_A$ was-afraid take$_{\mathsf{inf}}$ tie

(130) a. Troufla$_1$ *si*$_1$ usadit$_2$ *se*$_2$ v první řadě.
dared refl$_D$ to-sit refl$_A$ in first row
'She dared to sit in the first row.' [Rosen 2001 (233)]

b. * Troufla$_1$ *si*$_1$ usadit$_2$ v první řadě.
dared refl$_D$ to-sit refl$_A$ in first row

c. * Troufla$_1$ *se*$_2$ usadit$_2$ v první řadě.
dared refl$_A$ to-sit in first row

4.6.2 Constraints on the climbing path

- Clitics can climb only from infinitive phrases (124a), predicative adjective (124b), and in case of quantified genitives from quantified phrases (124c) (NPs, APs, or AdvPs); see for example (Rosen 2001, pp. 226f). Thus climbing is impossible from finite clauses, nominal, nonpredicative adjectival and adverbial participles and non-quantifying nominal phrases.

  This might be explained, in our view, by a requirement on a single path of climbing – there is only one sequence of embedded infinitives and one predicative nominal, but there can be several NPs or clauses with embedded clitics. Such requirement thus limits the possible governors of climbing clitics, therefore significantly decreasing the cognitive load on hearer processing a sentence with climbing clitics.

  However, there is an exception – quantified genitives can climb from subject and objects at the same time. Consider example (131). The genitive clitic *nás* 'us$_G$' belongs to the subject NP *většina nás* 'most of us', and the other genitive clitic *jich* 'them$_G$' belongs to the object NP *pět jich* 'five of them' (the order of the two clitics is probably free). Such sentences are rare but possible. We are not ready to explain this deviation and will have to leave it for further research.

(131) Je jich sedm, ale včera *nás jich* většina viděla jen pět.
is them$_G$ seven but yesterday us$_G$ them$_G$ most$_{fem}$ saw$_{fem}$ only five$_{non-oblique}$
'There are seven of them, but yesterday most of us saw only five of them.'

Also, Dotlačil (2006) nicely shows why it is logical that climbing out of CPs is impossible: As mentioned in §3.4.3, only contrastive or focused items can climb out of Czech CPs, clitics cannot be contrasted nor focused.

- Clitics nearly always climb out of phrases governed by auxiliaries, especially in case of past tense and conditional.

- Clitics tend to climb out of a phrase governed by a modal verb (Karlík et al. 1996, p. 651). Thus (132a) is usually preferred to (132b).

(132) a. V pondělí *mu*$_3$ *to*$_3$ budu$_1$ muset$_2$ konečně vrátit$_3$.
On Monday him$_D$ it$_A$ will$_{1sg}$ must$_{inf}$ finally return$_{\mathsf{inf}}$
'On Monday, I will have to return it finally to him'

b. V pondělí budu$_1$ muset$_2$ konečně *mu*$_3$ *to*$_3$ vrátit$_3$.
On Monday will$_{1sg}$ must$_{inf}$ finally him$_D$ it$_A$ return$_{\mathsf{inf}}$
'On Monday, I will have to return it finally to him'

- (Karlík et al. 1996, p. 651) claim that a clitic usually does not climb from phrases governed by nonmodal verbs, if its governor has other non-clitic dependents. However, this does not seem true. As Rosen (p.c.) notes, the example (133a), with climbing *ho* is better than with nonclimbing *ho* in (133b), even though *dát* 'give$_{\mathsf{inf}}$' has another complement: *Petrovi* 'Petr$_D$'.

(133) a. Marie *ho*$_2$ slíbila$_1$ dát$_2$ Petrovi.
Marie him$_A$ promissed give$_{\mathsf{inf}}$ Petr$_D$
'Marie promised to give it to Petr.' [rosen p.c.]

b. Marie slibila$_1$ dát$_2$ *ho*$_2$ Petrovi.
Marie promissed give$_{\mathsf{inf}}$ him$_A$ Petr$_D$
'Marie promised to give it to Petr.' [rosen p.c.]

4.6.3 Structural constraints

4.6.3.1 Climbing is monotonic

A clitic cannot climb over another clitic. More precisely:

(134) A clitic can climb to a particular cluster only if all clitics with a less embedded governor climbed to that or a higher cluster as well.

In (135a), clitics stay with their verbs so that the only clitic in Wackernagel position is *se* ‘refl$_A$’. In (135b), *mu* ‘him$_D$’ climbs from the verb *pomoci* ‘to-help’ to Wackernagel position, and *ho* ‘him$_A$’ climbs one level up, to the verb *pomoci.* Sentence (135d) is ill-formed, because the clitic *ho* ‘him$_A$’ is more embedded than the clitic *mu* ‘him$_D$’ (i.e., in *ho*’s governor is more embedded than *mu*’s governor), yet it occurs in a less embedded cluster than *mu* – *ho* is in the main cluster and *mu* is in the cluster of *pomoci.*

(135) a. Všichni *jsme*$_0$ *se*$_1$ snažili$_1$ [*mu*$_2$ pomoci$_2$ [*ho*$_3$ najít$_3$.]]
all aux$_{1pl}$ refl$_A$ tried him$_D$ to-help him$_A$ to-find
‘All of us tried to help him to find it.’

b. Všichni *jsme*$_0$ *se*$_1$ *mu*$_2$ snažili$_1$ [*ho*$_3$ pomoci$_2$ najít$_3$.]

c. Všichni *jsme*$_0$ *se*$_1$ *mu*$_2$ *ho*$_3$ snažili$_1$ pomoci$_2$ najít$_3$.

d. * Všichni *jsme*$_0$ *se*$_1$ *ho*$_3$ snažili$_1$ [*mu*$_2$ pomoci$_2$ najít$_3$.]

Note that the surface ordering of verbs does not have to correspond to their embeddedness; c.f. (136a) &

(136) a. Pomoci$_2$ najít$_3$ *jsme*$_0$ *se*$_1$ *mu*$_2$ *ho*$_3$ snažili$_1$ všichni.

b. [Pomoci$_2$ *mu*$_2$ *ho*$_3$ najít$_3$] *jsme*$_0$ *se*$_1$ snažili$_1$ všichni.

The (rather artificial) examples in (137) show that this applies even to more embedded clusters. While the sentence in (137a) with all clitics climbing to the main cluster is preferred, only the over the examples with partially climbing clitics, only (137de) violating the monotonicity constraint is are clearly out.

(137) a. [Zítra] *se*$_2$ *mu*$_3$ *ho*$_4$ určitě všichni začnou$_1$ snažit$_2$ pomoct$_3$ najít$_4$.
Tomorrow refl$_A$ him$_D$ him$_A$ definitely all start try$_{\mathsf{inf}}$ help$_{\mathsf{inf}}$ find$_{\mathsf{inf}}$
‘Tomorrow, all will definitely start to try to help him to find him/it.’

b. ? [Zítra] určitě všichni začnou$_1$ *se*$_2$ snažit$_2$ pomoct$_3$ *mu*$_3$ *ho*$_4$ najít$_4$.

c. ? [Zítra] určitě všichni začnou$_1$ *se*$_2$ *mu*$_3$ *ho*$_4$ snažit$_2$ pomoct$_3$ najít$_4$.

d. * [Zítra] určitě všichni začnou$_1$ *se*$_2$ *ho*$_4$ snažit$_2$ pomoct$_3$ *mu*$_3$ najít$_4$.

e. * [Zítra] *se*$_2$ *ho*$_4$ určitě všichni začnou$_1$ snažit$_2$ pomoct$_3$ *mu*$_3$ najít$_4$.

4.6.3.2 Control Constraints

Subject Control Constraint?

Thorpe (1991) has argued that clitics cannot climb from object-controlled infinitives. The clitic *ho* 'him$_A$' in (139a) governed by a subject-controlled infinitive may climb to the main clitic cluster, as in (139b). On the other hand, the clitic *ho* 'him$_A$' in (140) governed by a object-controlled infinitive cannot climb.[64]

(139) a. Alena *ho*$_2$ slíbila$_1$ navštívit$_2$, jakmile to bude možný.
Alena him$_A$ promised visit$_{\mathsf{inf}}$ as-soon-as it will be-possible$_{\mathsf{inf}}$
'Alena promised to visit him as soon as possible.'

b. Alena slíbila$_1$ navštívit$_2$ *ho*$_2$, jakmile to bude možný.
Alena promised visit$_{\mathsf{inf}}$ him$_A$ as-soon-as it will be-possible$_{\mathsf{inf}}$
'Alena promised to visit him as soon as possible.'

(140) a. * Alenu *ho*$_2$ nutili$_1$ navštívit$_2$.
Alena$_A$ him$_A$ forced$_{3pl}$ visit$_{\mathsf{inf}}$
intended: 'They were forcing Alena to visit him.'

b. Alenu nutili$_1$ navštívit$_2$ *ho*$_2$.
Alena$_A$ forced$_{3pl}$ visit$_{\mathsf{inf}}$ *him*$_2$
'They were forcing Alena to visit him.'

[64] Veselovská (1995, §9.5) argues that a similar constraint applies to all Exceptional Case Marking structures, including perception verbs that can be analyzed as object raising verb such as *vidět* 'see$_{\mathsf{inf}}$':

(138) a. Viděl$_1$ ji$_1$ dát$_2$ ho$_2$ Marušce.
saw$_{3sg}$ her$_A$ him$_A$ give$_{\mathsf{inf}}$ Maruška$_D$
'He saw her give it to Maruška. [Veselovská 1995]

b. * Viděl$_1$ ji$_1$ ho$_2$ dát$_2$ Marušce. (judgment by Veselovská)
saw$_{3sg}$ her$_A$ him$_A$ give$_{\mathsf{inf}}$ Maruška$_D$
'He saw her give it to Maruška. [Veselovská 1995]

However all questioned speakers accepted the sentence in (138b) with *ho* climbing out of the domain of the verb *dát* with object raised subject.

(141) **Subject Control Constraint (SCC)**

Clitics do not climb from object-controlled VPs.

However, this constraint is too strong. Consider the example in (142). The embedded infinitive *vyhodit* 'fire$_{\mathsf{inf}}$' is controlled by the indirect object *šéfovi* 'boss$_D$' of the verb *doporučila* 'recommended', yet *ho* 'him$_A$' governed by *vyhodit* climbs to the main cluster.

(142) Martinovi se v práci moc nedařilo, a když *ho*$_2$ i perzonalistika doporučila$_1$ šéfovi vyhodit$_2$, byl v háji.

'Martin was not very successful at his job and when even human resources recommended his boss to fire him, he was screwed.'

. . .	a	když	*ho*$_2$	i	perzonalistika	doporučila$_1$	šéfovi	vyhodit$_2$,	. . .
. . .	and	when	him$_A$	even	human-resources	recommended	boss$_D$	fire$_{\mathsf{inf}}$	. . .

'. . . and when even human resources recommended his boss to fire him, . . . '

Moreover, George and Toman (1976) show that a clitic can climb from an infinitive headed by a causative. Also, it can climb from (at least some) infinitives that are neither subject-controlled nor causatives, when it has non-animate referent (in the non-linguistic sense).

Reflexives and Control Constraint?

In (Hana 2004), unaware of the work by Thorpe (1991) and Veselovská (1995), we formulated the constrain in (145), weaker than (141). This was motivated by the fact that while the reflexive can climb from the subject controlled infinitives in (144), it cannot climb from the object controlled infinitives in (143).

(143) a.

*	Martin	*se*$_2$	zakázal$_1$	Petrovi	dívat$_2$	na	televizi.
	Martin	refl$_A$	forbid	Peter$_D$	to-watch	on	TV

'Martin forbid Peter to watch TV.'

b.

Martin	zakázal$_1$	Petrovi	dívat$_2$	*se*$_2$	na	televizi.
Martin	forbid	Peter$_D$	to-watch	refl$_A$	on	TV

'Martin forbid Peter to watch TV.'

c.

*	Neviděl$_1$	*jsem*$_0$	*si*$_2$	ještě	Martina	mýt$_2$	ruce.
	not-seen	aux$_{1sg}$	refl$_D$	yet	Martin$_A$	to-wash	hands$_A$

'I haven't seen Martin wash his hands yet.'

d.

Neviděl$_1$	*jsem*$_0$	ještě	Martina	mýt$_2$	*si*$_2$	ruce.
not-seen	aux$_{1sg}$	yet	Martin$_A$	to-wash	refl$_D$	hands$_A$

'I haven't seen Martin wash his hands yet.'

e. * Vláda *se*$_2$ občanům doporučila$_1$ pojistit$_2$.
government refl$_A$ citizens$_D$ recommended to-insure
'The government recommended the citizens get insurance.'

f. Vláda občanům doporučila$_1$ *se*$_2$ pojistit$_2$.
government citizens$_D$ recommended refl$_A$ to-insure
'The government recommended the citizens get insurance.'

(144) a. Při výběru *si*$_2$ zákazník musí$_1$ všímat$_2$ i ceny.
during selection refl$_D$ customer must to-pay-attention also price$_G$
'During selection, the customer must pay attention also to price.'

b. Ekonomika *se*$_2$ začíná$_1$ zlepšovat$_2$.
economy refl$_A$ starts to-improve
'The economy starts to improve.'

c. Martin *se*$_2$ potřebuje$_1$ zeptat$_2$, jak ...
Martin refl$_A$ needs to-ask how ...
'Martin needs to ask how ....'

d. Martin *se*$_2$ snažil$_1$ dokončit$_2$ všechno včas.
Martin refl$_A$ tried to-finish everything on-time
'Martin tried to finish everything on time.'

(145) **Reflexives and Control Constraint (RCC)**

Reflexive clitics do not climb from object-controlled VPs.

It seems clear that for non-reflexive clitic a more fine grained distinction of verbs than that based on control is needed. We leave this for further research.

4.6.3.3 Ordering by Governors' Degree of Embeddedness (GDEC)

Rosen (2001, p. 233) points out that if multiple dative clitics occur in a single clitic cluster they have to be ordered according to the relative embedding of their governors – a clitic governed by a more embedded verb follows a clitic with a less embedded verb. This can be seen in the example (146) containing two dative clitics *mi* 'me$_D$' and *mu* him$_D$. Since *mi* precedes *mu* in (146a), *mi*'s governor must be less embedded than *mu*'s governor – the opposite interpretation, as in (146b) is impossible. The other order of the dative clitics requires the opposite interpretation.[65]

[65] This could be analyzed in terms of crossing dependencies, which would mean the negation of Pesetskys (1982) Path Containment Condition holds. Note however that a clitic with a more embedded verb is required to come later in word order even when its verb is fronted.

(146) a. Poslat$_2$ kurýrem *se*$_1$ *mi*$_1$ *mu*$_2$ *ho*$_2$ dnes nepodařilo$_1$.
to-send by-courier refl$_A$ me$_D$ him$_D$ him$_A$ today not-succeeded
'I did not succeed in sending it to him by a courier today'

[Avgustinova and Oliva 1995 (20)]

b. ?? Poslat$_2$ kurýrem *se*$_1$ *mi*$_2$ *mu*$_1$ *ho*$_2$ dnes nepodařilo$_1$.
to-send by-courier refl$_A$ me$_D$ him$_D$ him$_A$ today not-succeeded
'He did not succeed in sending it to me by a courier today.'

c. Poslat$_2$ kurýrem *se*$_1$ *mu*$_1$ *mi*$_2$ *ho*$_2$ dnes nepodařilo$_1$.
to-send by-courier refl$_A$ him$_D$ me$_D$ him$_A$ today not-succeeded
'He did not succeed in sending it to me by a courier today.'

d. ?? Poslat$_2$ kurýrem *se*$_1$ *mu*$_2$ *mi*$_1$ *ho*$_2$ dnes nepodařilo$_1$.
to-send by-courier refl$_A$ him$_D$ me$_D$ him$_A$ today not-succeeded
'I did not succeed in sending it to him by a courier today.'

Similarly, in (147a), the dative pronoun *mu* 'him$_D$' goes before the dative pronoun *jí* 'her$_D$', therefore *mu* is governed by the highest verb – *zakázal* 'forbade' and *jí* by the embedded verb *kupovat* 'to-buy'. In (147b), the situation is reversed. Sentence (147c) shows that the linear order of the verbs is irrelevant, only their embedding is important.

(147) a. Martin *mu*$_1$ *jí*$_2$ včera zakázal$_1$ kupovat$_2$ takové dárky.
Martin him$_D$ her$_D$ yesterday forbade to-buy such presents
'Martin forbade him to buy her such presents yesterday.'
?'Martin forbade her to buy him such presents yesterday.'

b. Martin *jí*$_1$ *mu*$_2$ včera zakázal$_1$ kupovat$_2$ takové dárky.
Martin her$_D$ him$_D$ yesterday forbade to-buy such presents
'Martin forbade her to buy him such presents yesterday.'
?'Martin forbade him to buy her such presents yesterday.'

c. Kupovat$_2$ takové dárky *mu*$_1$ *jí*$_2$ včera Martin zakázal$_1$.
to-buy such presents him$_D$ her$_D$ yesterday Martin forbade
'Martin forbade him to buy her such presents yesterday.'
?'Martin forbade her to buy him such presents yesterday.'

Although co-occurrence of two accusatives in a single cluster is rather rare, the same constraint seem to apply, as (148) shows.

(148) a. Martin *jí*$_1$ *ho*$_2$ učil$_1$ napsat$_2$.
Martin her$_A$ him$_A$ taught write$_{\mathsf{inf}}$
'Martin taught her to write it. (e.g., *článek* 'article$_{masc}$')'

b. ? Martin *jí*$_2$ *ho*$_1$ učil$_1$ napsat$_2$.
Martin her$_A$ him$_A$ taught write$_{\mathsf{inf}}$
Intended: ‘Martin taught her to write it.’

c. Martin *ho*$_1$ *jí*$_2$ učil$_1$ napsat$_2$.
Martin him$_A$ her$_A$taught write$_{\mathsf{inf}}$
‘Martin taught him to write it.’ (e.g., *povídku* ‘novel$_{fem}$’)

d. ? Martin *ho*$_2$ *jí*$_1$ učil$_1$ napsat$_2$.
Martin him$_A$ her$_A$ taught write$_{\mathsf{inf}}$
Intended: ‘Martin taught him to write it.’

(149) **Ordering by Governors’ Degree of Embeddedness Constraint (GDEC)**
All (nonreflexive) dative clitics in the same cluster with the same case are ordered by the degree of embedding of their governors: namely, a clitic governed by a less deeply embedded verb precedes a clitic governed by a more deeply embedded verb. The surface order of the governors is irrelevant. The same probably holds also for personal accusative clitics.

4.6.3.4 Bonet’s Person-Case Constraint

(Bonet 1991, 1994) presents so-called Person-Case Constraints, a universal constraint[66] that disallows co-occurence of 1st and 2nd person accusatives with dative pronominal arguments of the same verb. It appears that in Czech such constraint holds only for some speakers, if at all. Rezac (2005) claims that that sequence of dative + non-3rd accusative is indeed impossible, except with ethical dative. For example, according to him, (150) is not grammatical.

(150) Ukážu *mu* *tě* zítra.
show$_{1sg}$ him$_D$ you$_A$ tomorrow
‘I will show you to him tomorrow’ [Rezac 2005]

However, for all questioned speakers, the sentence is fully acceptable and so are other sentences violating this constraint, including these two corpus examples:

(151) Chci *mu* *tě* ukázat.
want$_{1sg}$ him$_D$ you$_A$ show$_{\mathsf{inf}}$
‘I want to show you to him.’ [syn0]

(152) Pořád *mi* říkal, jak *je* *mu* *tě* líto.
all-the-time me$_D$ told how is him$_D$ you$_A$ sorry
‘He was telling me all the time how he felt sorry for you’ [syn5]

[66] She formulates the constraint in Optimality Theory, where all constraints are universal and only their ranking is language specific.

CHAPTER 5

CZECH IN HOG

In this chapter, we gradually develop a simple grammar of Czech in HOG. First, we provide the tectogrammar, then, after developing the necessary underlying framework, the corresponding phenogrammar. Finally, we focus on a particular phenomenon – clitics.

5.1 Simple Czech Tectogrammar

In this section, we develop a simple tectogrammar of Czech. Until the relation to phenogrammar is provided later in this chapter, we do not really know how to pronounce the tectogrammatical expressions. However, even before that, it should be intuitively clear that the tectogrammatical terms "make sense".

For the readers' convenience, we use English glosses as names for the lexical tectogrammatical terms. Thus instead of chrápe, we write snores. Note that these are just labels, we could also write 123-17-B or eats or charles-bridge for the same term, as long as we were consistent and made sure that it were pronounced by the phenogrammar as /xraːpɛ/. In §5.5, we will assign the tecto terms the expected pheno terms, for example the pheno term with phonology /ɛvjɛ/ to tecto term eva_D. Also note that $\mathsf{our}_{\mathsf{f.sg.A}}$ is a primitive term without any structure; the subscripts are necessary to capture distinctions required by Czech tectogrammar in a reader convenient way. Formally the subscripts have no status, we could use the term xyz for $\mathsf{our}_{\mathsf{f.sg.A}}$ and abc for $\mathsf{our}_{\mathsf{m.sg.N}}$. On the other hand, subscripts on types, for example $\mathsf{NP}_{\mathsf{acc}}$, have a precisely defined meaning (see §2.2).

The Czech grammar in this thesis assumes rather flat structures. For example, a sentence is an expression headed by a verb and there is no category of finite VPs. This provides word-order flexibility within larger domains of the heads, e.g., subject can occur in between complements without discontinuous structures. The other possibility is to use the standard (for English) hierarchical structures, but keep arguments and adjuncts separate up to the maximal projection and then order them all at once. This approach was taken for example by (Kupść 2000): while in syntax she distinguishes VPs and Ss, in phenogrammar (HPSG's domain objects) they form a single word-order domain.

5.1.1 Corpus to cover

The tectogrammar aims at covering the sentences below. We focus on several phenomena. At the end, the grammar covers simple sentences in present tense, past tense and conditional and handles subject-predicate agreement and agreement within noun phrases.

(1) Sentences, verb valency, adjuncts

a. Adam (často) (hrozně) chrápe.
Adam$_N$ often terribly snores
'Adam is (often) snoring (terribly).'

b. Adam (zase) krmí (naší) kozu (na louce).
Adam$_N$ again feeds our$_{f.sg.A}$ goat$_{f.A}$ on meadow$_{f.sg.L}$
'Adam is again feeding (our) goat (on the meadow).'

c. Adam (zase) dal (včera) Evě (pod stromem) hrušku.
Adam$_N$ again gave yesterday Eva$_D$ under tree$_{i.sg.I}$ pear$_{f.sg.A}$
'Adam gave Eva (again) a pear (yesterday) (under a tree).'

(2) Noun Phrases

a. (náš) (malý) Adam
our$_{m.sg.N}$ little$_{m.sg.N}$ Adam$_N$
'(our) (little) Adam '

b. (tu) (naší) (starou) kozu (od dědy).
that$_{f.sg.A}$ our$_{f.sg.A}$ old$_{f.sg.A}$ goat$_{f.A}$ from grandpa$_{m.sg.G}$
'(that/our) (old) goat from grandpa'

(3) complementizers, auxiliaries

a. Adam ví, že Petr dá Evě hrušku.
Adam$_N$ knows that Petr will-give$_{3sg}$ Eva$_D$ pear$_{f.A}$
'Adam knows that Petr will give Eva a pear.'

b. Adam by chrápal.
Adam$_N$ would$_3$ snored$_{m.sg}$
'Adam would snore.'

c. Adam by dal Evě hrušku.
Adam$_N$ would$_3$ gave$_{m.sg}$ Eva$_D$ pear$_{f.A}$
'Adam would give Eva a pear.'

d. Adam chrápal.
Adam$_N$ snored$_{m.sg}$
'Adam was snoring.'

(4) subject-predicate agreement[67]

a. Adam chrápe / *chrápu / *chrápou.
Adam$_N$ snores$_{3sg}$ / snore$_{1sg}$ / snore$_{3pl}$
'Adam snores (is snoring).'

b. Kluci chrápou / *chrápe / *chrápu.
Boys$_N$ snore$_{3pl}$ / snore$_{3sg}$ / snore$_{1sg}$
'Boys snore.'

c. Adam chrápal / *chrápala / *chrápali.
Adam$_N$ snored$_{m.sg}$ / snored$_{f.sg}$ / snored$_{m.pl}$
'Adam was snoring.'

d. Eva chrápala / chrápal* / *chrápali.
Eva$_N$ snored$_{f.sg}$ / snored$_{m.sg}$ / snored$_{m.pl}$
'Eva was snoring.'

e. Kluci chrápali / *...
Boys$_N$ snored$_{m.pl}$
'Boys snored.'

f. Holky chrápaly / *...
Girls$_N$ snored$_{f.pl}$
'Girls snored.'

g. Adam by chrápal.
Adam$_N$ would$_3$ snored$_{m.sg}$
'Adam would snore.'

h. Kluci by chrápali / *chrápaly / *chrápal
Boys$_N$ would snored$_{m.pl}$ / snored$_{f.pl}$ / snored$_{m.sg}$
'Boys would snore.'

i. Hruška by shnila / *shnil / *shnili / *shnily.
Pear$_{f.N}$ would rot$_{f.sg}$ / rot$_{m.sg}$ / rot$_{m.pl}$ /rot$_{f.pl}$
'A pear would rot.'

In the following text, we annotate tecto terms licensed by the grammar relative to this corpus in the following way: *term – the term is wrong and the grammar correctly does not licence it, !*term – the

[67] As mentioned in §A.2.1, only Official Czech distinguishes gender for plural participles, moreover *-ly* and *-li* have the same pronunciation. Similar situation applies to plural adjectives (§A.1.2).

term is wrong but is licensed (the grammar is overgenerating), !term – the term should be licensed by the grammar (the grammar is undergenerating).

5.1.2 Basic Valency

Governors are modeled as functions. Finite verbs are functions from the governed expressions to finite sentences (S for now; later we distinguish other types of sentences, too). Thus their type has the form of *valency* → S. Valencies are then captured as an indexed tuples.[68] The tuples are indexed by a set of indexes corresponding to syntactic functions like subject, object, indirect object, etc. An intransitive verbs is then a function taking tuple containing a single noun phrase indexed as SUBJ as its argument and returning a sentence. The type of intransitive verbs is thus:

(5) $[\textsc{subj}\ \mathsf{NP}] \to \mathsf{S}$

A transitive verb is a function of two arguments:

(6) $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$

or written as an AVM:

(7) $\begin{bmatrix} \textsc{subj} & \mathsf{NP} \\ \textsc{comps} & \mathsf{NP} \end{bmatrix} \to \mathsf{S}$

Note that the order in which the product components are written is irrelevant. For example, $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$ and $[\textsc{comps}\ \mathsf{NP}, \textsc{subj}\ \mathsf{NP}]$ are two different ways to write the same type.

This enables us to define the first simple tectogrammar of Czech or in fact of any language with subjects and complements:

(8) **Grammar:**

a. Indexes for products used for valencies: $\mathsf{subj}, \mathsf{obj}$

[68] The valency assumed in this grammar is very simple. A detailed analysis of Czech valency can be found for example in (Panevová 1980, 1994). The theory distinguishes 5 so-called actants (roughly non-adverbial complements) and a high number semantically-classified types of adverbials. Vallex (http://ufal.mff.cuni.cz/vallex/), a lexicon of approximately 2,700 lexemes has been created.

b. Basic types: NP, S

c. Lexicon defining four primitive terms, two NPs and two verbs:

$\mathsf{adam}_{\mathsf{N}}, \mathsf{goat}_{\mathsf{A}} : \mathsf{NP}$

$\mathsf{snores} : [\textsc{subj}\ \mathsf{NP}] \rightarrow \mathsf{S}$

$\mathsf{feeds} : [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \rightarrow \mathsf{S}$

The grammar licences the intuitively correct expressions in (9) and does not licence the intuitively incorrect expressions (10).

(9) a. $\mathsf{snores}(\textsc{subj}\ \mathsf{adam}_{\mathsf{N}}) : \mathsf{S}$

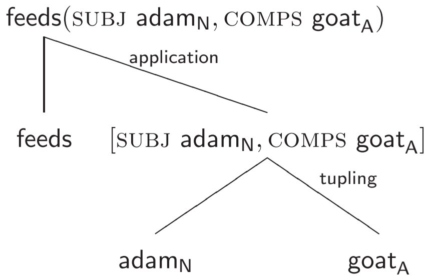
b. $\mathsf{feeds}(\textsc{subj}\ \mathsf{adam}_{\mathsf{N}}, \textsc{comps}\ \mathsf{goat}_{\mathsf{A}}) : \mathsf{S}$

(10) a. $*\,\mathsf{snores}(\textsc{subj}\ \mathsf{adam}_{\mathsf{N}}, \textsc{comps}\ \mathsf{goat}_{\mathsf{A}})$

b. $*\,\mathsf{feeds}(\textsc{subj}\ \mathsf{adam}_{\mathsf{N}})$

We say *intuitively correct* for two reasons. First, until the phenogrammar and its relation to tectogrammar is provided in the following section, we do not really know how to pronounce the terms. Second, while the format of the terms suggests the way in which it could have been derived, formally the terms are indivisible and they do not record history of the way they were created. Instead of $\mathsf{snores}(\textsc{subj}\ \mathsf{adam}_{\mathsf{N}})$ we could have written $\mathsf{term237114}$, because given a term, there is no way how to get the components it was created from. Similarly, given the number 4, there is no way to tell whether it is the result of $2 + 2$ or $3 + 1$. In (11), we show a derivation of (9b), corresponding to (12).

(11)



(12) Adam krmí kozu.
Adam_N feeds $\text{goat}_{f.A}$
‘Adam is feeding a/the goat.’

5.1.3 Case

As it stands, the grammar above licences also expressions with the subject in accusative, object in nominative, etc. In fact the grammar does not know about the category of case, at all.

(13) a. $*!\,\mathsf{snores}(\textsc{subj}\ \mathsf{goat_A}) : \mathsf{S}$

b. $*!\,\mathsf{feeds}(\textsc{subj}\ \mathsf{goat_A}, \textsc{comps}\ \mathsf{adam_N}) : \mathsf{S}$

To solve this problem, we add a primitive type Case of the possible 7 case terms and a function case assigning such terms to NPs (later, we make the function polymorphic to assign case to other types of expressions, too):[69]

(14) $\mathsf{Case} := \{\mathsf{nom}, \mathsf{gen}, \mathsf{dat}, \mathsf{acc}, \mathsf{voc}, \mathsf{loc}, \mathsf{ins}\}$

(15) $\mathsf{case} : \mathsf{NP} \rightarrow \mathsf{Case}$

Because HOG supports predicate subtyping (§2.2), we can refer not just to NPs but to NPs in a particular case. For example $\mathsf{NP}_{\lambda x:\mathsf{case}(x)=\mathsf{nom}}$, usually written simply as $\mathsf{NP_{nom}}$, is the type of all nominative NPs. Obviously, the lexicon then must be updated so that the primitive terms distinguish case, and verbs ask for the right case in their valencies.

(16) Lexicon (update):

$\mathsf{adam_N} : \mathsf{NP_{nom}}$ (*Adam*)

$\mathsf{goat_A} : \mathsf{NP_{acc}}$ (*kozu*)

$\mathsf{snores} : [\textsc{subj}\ \mathsf{NP_{nom}}] \rightarrow \mathsf{S}$ (*chrápe*)

$\mathsf{feeds} : [\textsc{subj}\ \mathsf{NP_{nom}}, \textsc{comps}\ \mathsf{NP_{acc}}] \rightarrow \mathsf{S}$ (*krmí*)

Now, the terms in (13) are no longer licenced, while the desirable terms in (9) are.

[69] If there were case neutralisation/syncretism in Czech (e.g., an NP would act as in nominative and accusative in the same utterance), one could introduce more fine-grained values as in (Daniels 2001) or (Pollard and Hana 2003, §3). Another possibility is to follow Pollard (2006): primitive predicates (instead of case values) for each of the case would be introduced, e.g., $\mathsf{nom} : \mathsf{NP} \rightarrow \mathsf{Bool}$. An NP syncretic between nominative and accusative would be of the type $\mathsf{NP}_{\lambda x\,.\,\mathsf{nom}(x)\&\mathsf{acc}(x)}$.

5.1.4 Verbal Adjuncts

We treat verbal adjuncts as being on the same level with complements. There are several reasons for this:

1. This is the usual way adjuncts are treated in Czech syntax (most analyses of Czech are within dependency grammar theories, and there is in fact no other option).
2. Unlike in English, Czech adjuncts freely mix with arguments in the surface strings. Adjuncts can separate a verb from its direct object.
3. As mentioned at the beginning of this chapter, this provides word-order flexibility within larger domains and many phenomena can be analyzed as scrambling instead of involving discontinuities. This is the view of most theories of Czech, including Functional Generative Description (FGD; e.g., Sgall et al. 1986). FGD is a dependency theory, which means the flat structures are inherent.

Thus verbs are functions not only from their complements to sentences but from their complements and potential adjuncts. We add ADJS to the possible valency indexes and allow every verb to combine with a (possibly empty) list of adjuncts.[70] Assuming that Adj is the type of verbal adjuncts (defined below), a list of adjuncts has the type Adj^* and we can redefine the two verbs in our grammar in the following way:

(17)
$$\begin{array}{ll} \mathsf{snores}: & [\textsc{subj}\ \mathsf{NP}_{\mathsf{nom}}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S} \\ \mathsf{feeds}: & [\textsc{subj}\ \mathsf{NP}_{\mathsf{nom}}, \textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S} \end{array}$$

To capture the generalization that every verb can combine with adjuncts, and most verbs (all verbs considered here) have subjects, we define a type operator that for a verb's valency (just complements, no subjects or adjuncts) gives the type of such verb:

(18) $\mathsf{FinVerb}(\mathit{Val} : \mathsf{VALENCY}) = [\textsc{subj}\ \mathsf{NP}_{\mathsf{nom}}, \textsc{adjs}\ \mathsf{Adj}^*] \oplus \mathit{Val} \to \mathsf{S}$

[70] At this point a multiset would suffice. But once we provide phenogrammar for this tectogrammar, we need to know which tecto adjunct corresponds to which pheno adjunct.

$\mathsf{VALENCY}$ is the kind of all possible valencies, excluding the subject. It is really simple in the present grammar – it contains only empty tuple or complements:

(19) $\mathsf{VALENCY} = \{[], [\textsc{comps}\ \mathsf{Tecto}]\}$

In a more elaborated grammar, additional distinctions can be made. The operator $\oplus$ is a type constructor merging two record types with distinct indexes. In §5.1.9 below, we modify the operator to handle subject-verb agreement properly and allow other subjects than nominative NPs.

When we are at it, we could also give simple names to the type of intransitive and transitive verbs with accusative objects as:

(20)
$$\begin{aligned} &\mathsf{IV} &&:= \mathsf{FinVerb}([]) \\ &\mathsf{TVa} &&:= \mathsf{FinVerb}([\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}]) \end{aligned}$$

We will not do it, to keep the type of verbs more apparent.

Finally, we need to specify what the type Adj is. Let's assume for the moment that a verbal adjunct can be either an adverbial phrase (AdvP) or a prepositional phrase (PP), both primitives for now. To express this, in (21) we define Adj to be a supertype (§2.2) of both. Note, that the type Adj is just a convenient name, the type $\mathsf{AdvP} + \mathsf{PP}$ exists whether we define Adj or not.

(21) $\mathsf{Adj} := \mathsf{AdvP} + \mathsf{PP}$

In (22), the full tectogrammar is listed as it is at this point. Henceforth, we omit $\textsc{adjs}\ \langle\rangle$ in terms denoting combination with no adjuncts.

(22) **Grammar:**

a. Indexes for products used for valencies:

$\textsc{subj}, \textsc{comps}, \textsc{adjs}$

b. Basic types, and type operators:

$\mathsf{NP}, \mathsf{S}, \mathsf{AdvP}, \mathsf{PP}$

$\mathsf{Case} := \{\mathsf{nom}, \mathsf{gen}, \mathsf{dat}, \mathsf{acc}, \mathsf{voc}, \mathsf{loc}, \mathsf{ins}\}$

$\mathsf{FinVerb}(\mathit{Val} : \mathsf{VALENCY}) = [\textsc{subj}\ \mathsf{NP}_{\mathsf{nom}}, \textsc{adjs}\ \mathsf{Adj}^{*}] \oplus \mathit{Val} \rightarrow \mathsf{S}$

$\mathsf{VALENCY} := \{[], [\textsc{comps}\ \mathsf{Tecto}]\}$

c. Type abbreviations:

$\mathsf{Adj} := \mathsf{AdvP} + \mathsf{PP}$

d. Nonlexical primitive terms:

$\mathsf{case} : \mathsf{NP} \to \mathsf{Case}$

e. Lexicon defining four primitive terms, two NPs and two verbs:

adam_N :	NP_nom
goat_A :	NP_acc
snores :	$\mathsf{FinVerb}([])$
	i.e. $[\textsc{subj}\ \mathsf{NP}_\mathsf{nom}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S}$
feeds :	$\mathsf{FinVerb}([\textsc{comps}\ \mathsf{NP}_\mathsf{acc}])$
	i.e. $[\textsc{subj}\ \mathsf{NP}_\mathsf{nom}, \textsc{comps}\ \mathsf{NP}_\mathsf{acc}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S}$
$\mathsf{often}, \mathsf{horribly}, \mathsf{again}$:	AdvP
$\mathsf{on\text{-}meadow}$:	PP

The grammar licences terms like (23) corresponding to sentences in (24).

(23) a. $\mathsf{snores}(\textsc{subj}\ \mathsf{adam}_\mathsf{N}, \textsc{adjs}\ \langle \mathsf{often}, \mathsf{horribly} \rangle) : \mathsf{S}$

b. $\mathsf{snores}(\textsc{subj}\ \mathsf{adam}_\mathsf{N}, \textsc{adjs}\ \langle \rangle) : \mathsf{S}$

c. $\mathsf{feeds}(\textsc{subj}\ \mathsf{adam}_\mathsf{N}, \textsc{comps}\ \mathsf{goat}_\mathsf{A}, \textsc{adjs}\ \langle \mathsf{often}, \mathsf{on\text{-}meadow} \rangle) : \mathsf{S}$

(24) a. Adam často hrozně chrápe.
Adam$_N$ often terribly snores
'Adam often snores terribly.'

b. Adam chrápe.
Adam$_N$ snores
'Adam is snoring.'

c. Adam často krmí kozu na louce.
Adam$_N$ often feeds goat$_{f.A}$ on meadow$_{f.sg.L}$
'Adam often feeds a/the goat on a/the meadow.'

As an example, we show a derivation of (23c) in Figure 5.1. The derivation proves that there is a tecto term $[\textsc{subj}\ \mathsf{adam}_\mathsf{N}, \textsc{comps}\ \mathsf{goat}_\mathsf{A}, \textsc{adjs}\ \langle \mathsf{often}, \mathsf{on\text{-}meadow} \rangle]$ of the type S, in other words that the grammar licences the sentence

(25) $[\textsc{subj}\ \mathsf{adam}_\mathsf{N}, \textsc{comps}\ \mathsf{goat}_\mathsf{A}, \textsc{adjs}\ \langle \mathsf{often}, \mathsf{on\text{-}meadow} \rangle]$.

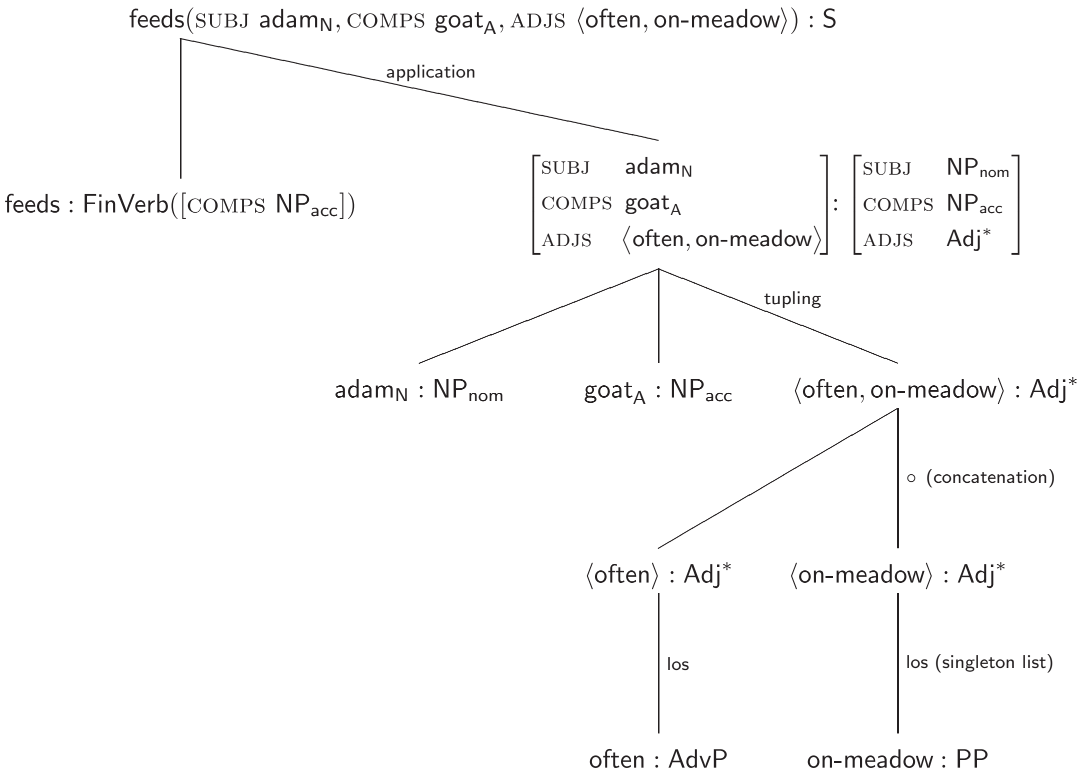

feeds(SUBJ adam$_N$, COMPS goat$_A$, ADJS ⟨often, on-meadow⟩) : S
application
feeds : FinVerb([COMPS NP$_{acc}$])
[SUBJ adam$_N$ COMPS goat$_A$ ADJS ⟨often, on-meadow⟩] : [SUBJ NP$_{nom}$ COMPS NP$_{acc}$ ADJS Adj*]
tupling
adam$_N$: NP$_{nom}$
goat$_A$: NP$_{acc}$
⟨often, on-meadow⟩ : Adj*
∘ (concatenation)
⟨often⟩ : Adj*
⟨on-meadow⟩ : Adj*
los
los (singleton list)
often : AdvP
on-meadow : PP

Figure 5.1: Sample tecto derivation of a sentence

1. The derivation/proof starts with the facts known from the tecto lexicon, which are non-logical axioms. For example $\mathsf{often} : \mathsf{AdvP}$, i.e., there is a term often of the type AdvP.

2. If $a : \mathsf{A}$ and $\mathsf{A} \sqsubseteq \mathsf{B}$ then $a : B$.[71] If a grammar licences a term, it licences its embedding into a supertype.

   Therefore, from $\mathsf{often} : \mathsf{AdvP}$ we know $\mathsf{often} : \mathsf{Adj}$, because $\mathsf{Adj} = \mathsf{AdvP} + \mathsf{PP}$ is a supertype of AdvP.

3. If $a : \mathsf{A}$ then $\langle a \rangle : \mathsf{A}^*$. If a grammar licences a term, it licences a singleton list of that term (similarly a set or a multiset).

   Therefore from $\mathsf{often} : \mathsf{Adj}$ we know $\langle \mathsf{often} \rangle : \mathsf{Adj}^*$

4. Similarly, we can show that $\langle \mathsf{on\text{-}meadow} \rangle : \mathsf{Adj}^*$

[71] Formally, this is a little bit more complicated. As discussed in §C.4.4, the logic used in HOG requires that every term belongs to exactly one type (so-called *monotyping* property). Therefore, a term of a particular type must be 'packaged' by the appropriated embedding function to be a term of a supertype. For two types A and B, where $\mathsf{A} \sqsubseteq \mathsf{B}$, the 'packaging' or embedding function is written as $\mathsf{ker}_{\mathsf{A},\mathsf{B}}$. Therefore, a precise formulation of the above statement is:

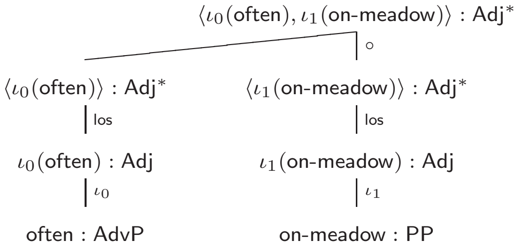
(26) if $a : \mathsf{A}$ and $\mathsf{A} \sqsubseteq \mathsf{B}$ then $\mathsf{ker}_{\mathsf{A},\mathsf{B}}(a) : B$

For supertypes defined via coproducts (see §C.4.1), the packaging function is the appropriate injections. For the case above, where often is of type AdvP and we want a term of type $\mathsf{Adj} = \mathsf{AdvP} + \mathsf{PP}$, the embedding function is

(27) $\mathsf{ker}_{\mathsf{AdvP},\mathsf{Adj}} = \mathsf{ker}_{\mathsf{AdvP},\mathsf{AdvP}+\mathsf{PP}} = \iota_{0,\mathsf{AdvP}+\mathsf{PP}}$

Except in the most meticulous formulations, $\iota_{0,\mathsf{A}+\mathsf{B}}$ is written simply as ι_0 If we wanted to be precise and show the injection, the lower right part of the proof in (5.1) would look as follows:

(28)



Because, (1) the 'packaging' can be inferred from context (except types of the form $\mathsf{A} + \mathsf{A}$, but such types have no linguistic motivation), and (2) it is a purely technical requirement, without any linguistic significance, we consistently omit it.

5. If $a : \mathsf{A}^*$ and $b : \mathsf{A}^*$, then $a \circ b : \mathsf{A}^*$. If a grammar licences two lists of the same type, it licences their concatenation.

   Thus from $\langle \mathsf{often} \rangle : \mathsf{Adj}^*$ and $\langle \mathsf{on\text{-}meadow} \rangle : \mathsf{Adj}^*$ we know $\langle \mathsf{often}, \mathsf{on\text{-}meadow} \rangle : \mathsf{Adj}^*$

6. The tecto lexicon guarantees that $\mathsf{adam_N} : \mathsf{NP_{nom}}$ and $\mathsf{goat_A} : \mathsf{NP_{acc}}$.

7. The logic licences indexed tuples of terms.

   Thus from $\mathsf{adam_N} : \mathsf{NP_{nom}}$, $\mathsf{goat_A} : \mathsf{NP_{acc}}$, and $\langle \mathsf{often}, \mathsf{on\text{-}meadow} \rangle : \mathsf{Adj}^*$ we know that

   $$\begin{bmatrix} \textsc{subj} & \mathsf{adam_N} \\ \textsc{comps} & \mathsf{goat_A} \\ \textsc{adjs} & \left\langle \mathsf{often}, \mathsf{on\text{-}meadow} \right\rangle \end{bmatrix} : \begin{bmatrix} \textsc{subj} & \mathsf{NP_{nom}} \\ \textsc{comps} & \mathsf{NP_{acc}} \\ \textsc{adjs} & \mathsf{Adj}^* \end{bmatrix}$$

8. If $f : \mathsf{A} \rightarrow \mathsf{B}$ and $a : \mathsf{A}$ then $f(a) : \mathsf{B}$. If a grammar licences a function and a term that can serve as an argument to the function, it licences the result of applying the function to the argument.

   In this case, the function is the transitive verb feeds and the argument is the tuple from the previous step. Because we know from the tecto lexicon that $\mathsf{feeds} : \mathsf{FinVerb}([\textsc{comps}\ \mathsf{NP_{acc}}])$, which means:

   $$\mathsf{feeds} : \begin{bmatrix} \textsc{subj} & \mathsf{NP_{nom}} \\ \textsc{comps} & \mathsf{NP_{acc}} \\ \textsc{adjs} & \mathsf{Adj}^* \end{bmatrix} \rightarrow \mathsf{S_{fin}},$$

   and from the previous step that

   $$\begin{bmatrix} \textsc{subj} & \mathsf{adam_N} \\ \textsc{comps} & \mathsf{goat_A} \\ \textsc{adjs} & \left\langle \mathsf{often}, \mathsf{on\text{-}meadow} \right\rangle \end{bmatrix} : \begin{bmatrix} \textsc{subj} & \mathsf{NP_{nom}} \\ \textsc{comps} & \mathsf{NP_{acc}} \\ \textsc{adjs} & \mathsf{Adj}^* \end{bmatrix},$$

   we also know that

   $\mathsf{feeds}(\textsc{subj}\ \mathsf{adam_N}, \textsc{comps}\ \mathsf{goat_A}, \textsc{adjs}\ \langle \mathsf{often}, \mathsf{on\text{-}meadow} \rangle) : \mathsf{S}$

It would be possible to prove a simple schematic lemma that would allow us to do the proof above in a one-step “big” application and hide the technical steps of creating singleton lists, concatenation and tupling. The proof trees would then look as usual (flat) syntactic structures.

5.1.5 NPs

5.1.5.1 Structure

We also follow FGD in assuming that the head of the Czech NP is the noun. L. Zlatić argues in favor of such treatment of NPs in Slavic languages in general (Zlatić to appear), and in Serbian in particular (Zlatić 1997). As first approximation, we can treat all modifiers (attributes) as nominal adjuncts.

(29) $\mathsf{N}_c := [\textsc{adjs}\ \mathsf{Attr}^*] \rightarrow \mathsf{NP}_c$

(30) $\mathsf{Attr} := \mathsf{AP} + \mathsf{Det} + \mathsf{Poss} + \mathsf{PP} + \mathsf{NP}_{\mathsf{gen}} + \mathsf{NP}_{\mathsf{dat}} \ldots$

(31) Lexicon (update):

$\mathsf{adam}_{\mathsf{N}} : \mathsf{N}_{\mathsf{nom}}$ (*Adam*)

$\mathsf{goat}_{\mathsf{A}} : \mathsf{N}_{\mathsf{acc}}$ (*kozu*)

In this simple grammar, we ignore the fact that occurrence of many of the attributes is not entirely free – for example, only one or two genitive NPs can modify the noun.

5.1.5.2 Agreement

Some nominal adjuncts, usually called agreeing attributes, agree with the head noun in gender, number and case (see §A.2.1.2 for more details). For expository reasons, we consider only adjectives, possessive pronouns and determiners.

First, we need to redefine the case function introduced in (14) to allow case on other expressions than just NPs. It is defined as a polymorphic function from the kind $\{\mathsf{NP}, \mathsf{AP}, \mathsf{Det}, \mathsf{Poss}\}$ to the type Case:

(32) $\mathsf{case} : \{\mathsf{NP}, \mathsf{AP}, \mathsf{Det}, \mathsf{Poss}\} \rightarrow \mathsf{Case}$

Similar functions are introduced for gender and number:

(33)
$$\begin{array}{ll} \mathsf{Nr} = & \{\mathsf{sg}, \mathsf{pl}\} \\ \mathsf{nr}: & \{\mathsf{NP}, \mathsf{AP}, \mathsf{Det}, \mathsf{Poss}\} \rightarrow \mathsf{Nr} \\ \mathsf{Gender} = & \{\mathsf{m}, \mathsf{i}, \mathsf{f}, \mathsf{n}\} \\ \mathsf{gender}: & \{\mathsf{NP}, \mathsf{AP}, \mathsf{Det}, \mathsf{Poss}\} \rightarrow \mathsf{Gender} \end{array}$$

Below, we modify the domains of the three functions because, for example, past participles have gender and number but no case, while finite verbs have number but no case or gender.

Now, we need to properly constrain the gender, number and case values within the NP. The easiest way to handle this is to split nominal attributes into agreeing attributes and non-agreeing attributes. The agreeing attribute then has the same values for gender, number and case as the head noun, and so does the whole NP. We define a schematic dependent type[72] $\mathsf{N}_{gen,nr,c}$ that ensures this:

(35) $\mathsf{N}_{gen,nr,c} := [\textsc{agr}\ \mathsf{AttrAgr}^{*}_{gen,nr,c}, \textsc{nagr}\ \mathsf{AttrNon}^{*}] \rightarrow \mathsf{NP}_{gen,nr,c}$

Obviously, it is just a matter of personal preference whether one writes the type schema that way or in an AVM notation as in (36).

[72] A dependent type is a type which depends on a term. It is a type operator receiving terms as parameters and returning types depending on the parameters. HOG does not have the full power of dependent types, but some of the possibilities can be expressed by predicate subtyping and some can be thought in terms of schemas. In this case, the expression that is abbreviated by N, to be precise

(34) $\mathsf{N} = \lambda gen : \mathsf{Gender}, nr : \mathsf{Nr}, c : \mathsf{Case}\,.\,[\textsc{agr}\ \mathsf{AttrAgr}^{\circ}_{gen,nr,c}, \textsc{nagr}\ \langle\mathsf{AttrNon}\rangle] \rightarrow \mathsf{NP}_{gen,nr,c}$

receives three parameters – gender, number and case, and for each combination returns a normal nondependent type. Because there are countably many combination (4 genders, 2/3 numbers and 7 cases), this could be thought of as a schema defining several nondependent types.

(36) $$\begin{bmatrix} \mathsf{N} & \\ \mathsf{gender} & \boxed{\text{gen}} \\ \mathsf{number} & \boxed{\text{nr}} \\ \mathsf{case} & \boxed{\text{c}} \end{bmatrix} := \begin{bmatrix} \textsc{agr} & \begin{bmatrix} \mathsf{AttrAgr} & \\ \mathsf{gender} & \boxed{\text{gen}} \\ \mathsf{number} & \boxed{\text{nr}} \\ \mathsf{case} & \boxed{\text{c}} \end{bmatrix}^{*} \\ \textsc{nagr} & \mathsf{AttrNon}^{*} \end{bmatrix} \rightarrow \begin{bmatrix} \mathsf{NP} & \\ \mathsf{gender} & \boxed{\text{gen}} \\ \mathsf{number} & \boxed{\text{nr}} \\ \mathsf{case} & \boxed{\text{c}} \end{bmatrix}$$

The types of agreeing and non-agreeing attributes are defined in (37).

(37) $$\begin{aligned} \mathsf{AttrAgr} &= \mathsf{AP} + \mathsf{Det} + \mathsf{Poss} \\ \mathsf{AttrNon} &= \mathsf{PP} + \mathsf{NP}_{\mathsf{gen}} + \mathsf{NP}_{\mathsf{dat}} \end{aligned}$$

where $(A + B)_{\phi}$ means $A_{\phi} + B_{\phi}$, therefore:

(38) $$\mathsf{AttrAgr}_{gen,nr,c} = \mathsf{AP}_{gen,nr,c} + \mathsf{Det}_{gen,nr,c} + \mathsf{Poss}_{gen,nr,c}$$

And finally, the lexicon must be updated so that the nouns “know” not only their case, but also their gender and number. We also add terms for several agreeing and non-agreeing attributes.

(39) Lexicon (update):

$\mathsf{adam}_{\mathsf{N}} : \mathsf{N}_{\mathsf{m,sg,nom}}$	(*Adam*)
$\mathsf{goat}_{\mathsf{A}} : \mathsf{N}_{\mathsf{f,sg,acc}}$	(*kozu*)
$\mathsf{that}_{\mathsf{f.sg.A}} : \mathsf{Det}_{\mathsf{f,sg,A}}$	(*tu*)
$\mathsf{our}_{\mathsf{f.sg.A}} : \mathsf{Poss}_{\mathsf{f,sg,A}}$	(*naší*)
$\mathsf{our}_{\mathsf{m.sg.N}} : \mathsf{Poss}_{\mathsf{m,sg,N}}$	(*náš*)
$\mathsf{little}_{\mathsf{m.sg.N}} : \mathsf{AP}_{\mathsf{m,sg,N}}$	(*malý*)
$\mathsf{old}_{\mathsf{f.sg.A}} : \mathsf{AP}_{\mathsf{f,sg,A}}$	(*starou*)
on-meadow : PP	(*na louce*)
under-tree : PP	(*pod stromem*)
from-grandpa : PP	(*od dědy*)

Now, we can derive the terms in (40) corresponding to phrases in (41).

(40) a. $\mathsf{adam}_{\mathsf{N}}(\textsc{agr}\ \langle \mathsf{our}_{\mathsf{m.sg.N}}, \mathsf{little}_{\mathsf{m.sg.N}} \rangle) : \mathsf{NP}_{\mathsf{m,sg,nom}}$

b. $\mathsf{goat}_{\mathsf{A}}(\textsc{agr}\ \langle \mathsf{that}_{\mathsf{f.sg.A}}, \mathsf{our}_{\mathsf{f.sg.A}}, \mathsf{old}_{\mathsf{f.sg.A}} \rangle, \textsc{nagr}\ \langle \mathsf{from\text{-}grandpa} \rangle) : \mathsf{NP}_{\mathsf{f,sg,acc}}$

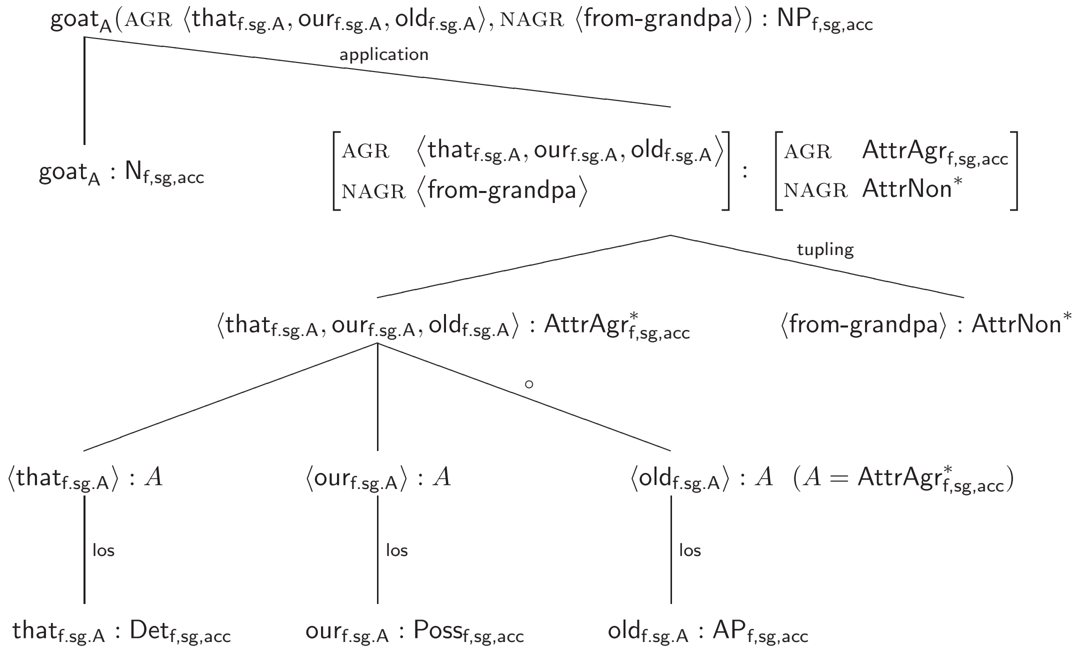

$\mathsf{goat_A}(\textsc{agr}\ \langle \mathsf{that_{f.sg.A}}, \mathsf{our_{f.sg.A}}, \mathsf{old_{f.sg.A}} \rangle, \textsc{nagr}\ \langle \mathsf{from\text{-}grandpa} \rangle) : \mathsf{NP_{f,sg,acc}}$
application
$\mathsf{goat_A} : \mathsf{N_{f,sg,acc}}$
$\begin{bmatrix} \textsc{agr} & \langle \mathsf{that_{f.sg.A}}, \mathsf{our_{f.sg.A}}, \mathsf{old_{f.sg.A}} \rangle \\ \textsc{nagr} & \langle \mathsf{from\text{-}grandpa} \rangle \end{bmatrix} : \begin{bmatrix} \textsc{agr} & \mathsf{AttrAgr_{f,sg,acc}} \\ \textsc{nagr} & \mathsf{AttrNon}^* \end{bmatrix}$
tupling
$\langle \mathsf{that_{f.sg.A}}, \mathsf{our_{f.sg.A}}, \mathsf{old_{f.sg.A}} \rangle : \mathsf{AttrAgr^*_{f,sg,acc}}$
$\langle \mathsf{from\text{-}grandpa} \rangle : \mathsf{AttrNon}^*$
$\circ$
$\langle \mathsf{that_{f.sg.A}} \rangle : A$
$\langle \mathsf{our_{f.sg.A}} \rangle : A$
$\langle \mathsf{old_{f.sg.A}} \rangle : A \quad (A = \mathsf{AttrAgr^*_{f,sg,acc}})$
los
los
los
$\mathsf{that_{f.sg.A}} : \mathsf{Det_{f,sg,acc}}$
$\mathsf{our_{f.sg.A}} : \mathsf{Poss_{f,sg,acc}}$
$\mathsf{old_{f.sg.A}} : \mathsf{AP_{f,sg,acc}}$

Figure 5.2: Sample tecto derivation of an NP

(41) a. náš malý Adam
our$_{m.sg.N}$ little$_{m.sg.N}$ Adam$_N$
'our little Adam'

b. tu naší starou kozu od dědy
that$_{f.sg.A}$ our$_{f.sg.A}$ old$_{f.sg.A}$ goat$_{f.A}$ from grandpa$_{m.sg.G}$
'that our old goat from grandpa'

The derivation of (40b) corresponding to (41b) is captured in Figure 5.2.

The term in (42) corresponding to non-grammatical noun phrase in (43) is not licensed.

(42) $*\,\mathsf{adam_N}(\textsc{agr}\ \langle \mathsf{our_{f.sg.A}} \rangle)$

(43) * naší Adam
our$_{f.sg.A}$ Adam$_N$

5.1.6 PPs

To model prepositional phrases, we introduce a primitive type PP. Prepositions are then functions from getting an NP as an argument and returning a PP. Verbs and nouns then subcategorize for a particular subtype of PP – for each preposition, there is a predicate defining such a subtype of PP. For example, one of the constructions with the verb *donutit* 'force' requires an object with prepositions *k* as in (44). The tecto term corresponding to the verb is in (45), it requires an object of the type $\mathsf{PP_k}$. $\mathsf{PP_k}$ is the type of prepositional phrases with *k* as the preposition, it is a subtype of PP, the type of all prepositional phrases. k is a PP-predicate corresponding to the tecto term k (they are written the same way).

(44) Martin donutí Petra k odchodu.
Martin will-forces Petr prep leaving$_{dat}$
'Martin will force Petr to leave.'

(45) $\mathsf{forces} : \mathsf{FinVerb}([\textsc{comps}\ \mathsf{NP_{acc}} \circ \mathsf{PP_k}])$ (*donutí*)

(46) $\mathsf{k} : [\textsc{comps}\ \mathsf{NP_{dat}}] \rightarrow \mathsf{PP_k}$ (*k*)

Note that many prepositions are ambiguous and can combine with various cases. For example, the verb *dívat se* 'watch$_{\mathsf{inf}}$' requires the object to be a PP with preposition *na* and an NP in accusative, while the verb *záviset* 'depend$_{\mathsf{inf}}$' requires there to be a PP with preposition *na* and an NP in locative – see (47) for examples. Because case cannot be neutralized for such preposition (e.g., a preposition requiring accusative or locative cannot take a conjunction of these two), we treat such prepositions as distinct in tecto and the verbs subcategorize for formally unrelated PPs as the lexical entries in (48) show.

(47) a. Petr se dívá na televizi.
Petr refl$_A$ watch prep TV$_{acc}$
'Petr is watching TV.'

b. Náš osud závisí především na demokratizaci Srbska.
Our fate depends mostly prep democratization Serbia$_{gen}$
'Our fate depends mostly on the democratization of Serbia.' [syn6]

(48) Lexicon (update):

$\mathsf{na_A} : [\textsc{comps}\ \mathsf{NP_{loc}}] \rightarrow \mathsf{PP_{na\text{-}acc}}$ (*na*)

$\mathsf{na_L} : [\textsc{comps}\ \mathsf{NP_{loc}}] \rightarrow \mathsf{PP_{na\text{-}loc}}$ (*na*)

$\mathsf{pod_A} : [\textsc{comps}\ \mathsf{NP_{acc}}] \rightarrow \mathsf{PP_{pod\text{-}acc}}$ (*pod*)

$\mathsf{pod_I} : [\textsc{comps}\ \mathsf{NP_{ins}}] \rightarrow \mathsf{PP_{pod\text{-}ins}}$ (*pod*)

$\mathsf{od_G} : [\textsc{comps}\ \mathsf{NP_{gen}}] \rightarrow \mathsf{PP_{od\text{-}gen}}$ (*od*)

$\mathsf{meadow_L} : \mathsf{N_{f,sg,loc}}$ (*louce*)

$\mathsf{tree_I} : \mathsf{N_{m,sg,ins}}$ *(stromem)*

$\mathsf{grandpa_G} : \mathsf{N_{f,sg,gen}}$ *(dědy)*

$\mathsf{looks} : \mathsf{FinVerb}([\textsc{comps}\ \mathsf{PP_{na\text{-}acc}}])$ *(dívá se)*

$\mathsf{depends} : \mathsf{FinVerb}([\textsc{comps}\ \mathsf{PP_{na\text{-}loc}}])$ *(závisí)*

(and we drop the primitive terms of type PP, e.g., $\mathsf{from\text{-}grandpa}$)

5.1.7 Complementizers

In this subsection, we expand the tectogrammar to handle complementized clauses like the one in (49).

(49) Adam vabout, že Petr dá Evě hrušku.
Adam$_N$ knows that Petr will-give$_{3sg}$ Eva$_D$ pear$_{f.A}$
'Adam knows that Petr will give Eva a pear.'

Adding a new type $\bar{\mathsf{S}}$ for complementized clauses, we can update the lexicon to contain tecto terms for the complementizer *že* 'that' and the verb *ví* 'knows$_{3sg}$':

(50) Lexicon (update):

$\mathsf{that} : [\textsc{spec}\ \mathsf{S}] \to \bar{\mathsf{S}}$ *(že)*

$\mathsf{knows} : \mathsf{FinVerb}([\textsc{comps}\ \bar{\mathsf{S}}])$ *(ví)*

Such grammar licences the following term corresponding to the sentence in (49):

(51) $\mathsf{knows}(\textsc{subj}\ \mathsf{adam_N}, \textsc{comps}\ \mathsf{that}(\textsc{spec}\ \mathsf{will\text{-}give_{3sg}}(\textsc{subj}\ \mathsf{petr_N}, \textsc{comps}\ \mathsf{pear_A} \circ \mathsf{eva_D})))$

5.1.8 Auxiliaries

Czech verbal periphrastic constructions are formed by combining the auxiliary or copula (forms of the verb *být* 'to be') with participles or infinitives. In this section, we address conditional and past tense. Leaving future tense, past conditional and periphrastic passive aside.

First, we need to distinguish among different clause types. These will all be subtypes of the type S. We can introduce an S-predicate for each some type:

1. $\mathsf{fin} : \mathsf{S} \to \mathsf{Bool}$ determines $\mathsf{S_{fin}}$, the type of finite clauses (up to this point simply S)

2. $\mathsf{inf} : \mathsf{S} \rightarrow \mathsf{Bool}$ determines $\mathsf{S_{inf}}$ clauses headed by an infinitive

3. $\mathsf{pp} : \mathsf{S} \rightarrow \mathsf{Bool}$ determines $\mathsf{S_{pp}}$ clauses headed by past participles

Conditional is formed by combining the conditional auxiliary with past participles, e.g., *psala by* 'she would write$_{fem}$', *byl bych* 'I would be'. Past tense is formed by combining the past auxiliary with a past participle. In past tense, there is no auxiliary in the 3rd person.

Argument raising The arguments of conditional and paste tense act as being arguments of the whole periphrastic verb. First, they are freely scrambled in the domain of the auxiliary. Any of the arguments can occur before, in between or after the auxiliary and the participle, as sentences in (52) show.

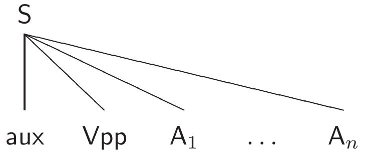
(52) Adam by dal Evě hrušku.
Adam$_N$ would gave$_{m.sg}$ Eva$_D$ pear$_{f.A}$
'Adam would give Eva a pear.'

Hrušku by dal Evě Adam.
pear$_{f.A}$ would gave$_{m.sg}$ Eva$_D$ Adam$_N$
'Adam would give Eva a pear.'

Adam by Evě dal hrušku.
Adam$_N$ would Eva$_D$ gave$_{m.sg}$ pear$_{f.A}$
'Adam would give Eva a pear.'

Assuming the subject is an argument of the auxiliary and the complements of the participle would result in frequent discontinuities. While other discontinuities (e.g., split-topicalization, see §3.4.2) are considered as somehow marked by speakers, this does not seem to be true here. Second, it does not make sense to distinguish whether adjuncts modify the auxiliary or the past participle. Therefore, we assume a flat structure and analyze both the conditional and the past tense via argument raising (Hinrichs and Nakazawa 1994). The auxiliary "steals the valency" of the past participle; schematically:

(53)



Type of the conditional The conditional auxiliary has the following polymorphic type, ($\oplus$ fuses two record types):

(54) $\mathsf{would}_{3sg} : ([\textsc{gov}\ (A \to \mathsf{S_{pp}})] \oplus A) \to \mathsf{S_{fin}}$ (by)

The auxiliary specifies only one dependent in its valency – it governs a past participle.[73] It accepts the participle without satisfied valency and fills the valency for it (including the subject).

Type of the past auxiliary. As mentioned above, in the past tense, there is no auxiliary in the 3rd person and the past is expressed simply by the past participle. There are several ways to address it: (1) assume that the participle in the third person is in fact a finite verb; (2) the auxiliary is always present in tecto, but it is phonetically realized only in first and second persons; (3) assume there is a term that transforms a participle into a finite verb, an analog of a unary rule in rewriting grammars. In HOG, the (2) and (3) items are equivalent. We choose to have an auxiliary in all persons. Unlike some other authors taking a similar route (e.g., Veselovská 1995), we do not claim that the language system or our mental grammars are actually patterned that way.

(55) $\mathsf{past}_{3sg} : ([\textsc{gov}\ (A \to \mathsf{S_{pp}})] \oplus A) \to \mathsf{S_{fin}}$

In fact, the future auxiliary can combine only with imperfective verbs (see §A.1.5), thus to analyze future tense adequately, verbs and nonfinite clauses would have to distinguish aspect. We ignore this detail here.

Participles. We add two participles to the lexicon in (56); $\mathsf{snore_{pp.m.sg}}$ corresponds to *chrápal*, past participle of an intransitive verb, $\mathsf{give_{pp.m.sg}}$ corresponds to *dal*, past participle of a ditransitive verb.

(56) $\mathsf{snore_{pp.m.sg}} : \mathsf{PstParticiple}([\textsc{subj}\ \mathsf{NP_{nom}}])$
$\mathsf{give_{pp.m.sg}} : \mathsf{PstParticiple}([\textsc{subj}\ \mathsf{NP_{nom}}, \textsc{comps}\ \mathsf{NP_{acc}} \circ \mathsf{NP_{dat}}])$

This uses a type operator PstParticiple, similar to the operator FinVerb defined in (18):

(57) $\mathsf{PstParticiple}(\mathit{Val} : \mathsf{VALENCY}) = [\textsc{subj}\ \mathsf{NP_{nom}}, \textsc{adjs}\ \mathsf{Adj}^*] \oplus \mathit{Val} \to \mathsf{S_{pp}}$

We can now combine the participles with the conditional auxiliary to derive terms like those in (58) which correspond in an obvious way to sentences in (59).

[73] The name of the valency index, GOV , is taken from (Chung 1998).

(58) a. $\mathsf{would}_{3sg}([\textsc{gov}\ \mathsf{snore}_{\mathsf{pp.m.sg}}, \textsc{subj}\ \mathsf{NP}_{\mathsf{nom}}]) : \mathsf{S}_{\mathsf{fin}}$

b. $\mathsf{would}_{3sg}([\textsc{gov}\ \mathsf{give}_{\mathsf{pp.m.sg}}, \textsc{subj}\ \mathsf{NP}_{\mathsf{nom}}, \mathsf{pear}_{\mathsf{A}} \circ \mathsf{eva}_{\mathsf{D}}]) : \mathsf{S}_{\mathsf{fin}}$

(59) a. Adam by chrápal.
Adam$_N$ would snored$_{m.sg}$
'Adam would snore.'

b. Adam by dal Evě hrušku.
Adam$_N$ would gave$_{m.sg}$ Eva$_D$ pear$_{f.A}$
'Adam would give Eva a pear.'

5.1.9 Subject-Predicate Agreement

5.1.9.1 Personal Pronouns

First we need to distinguish NPs that are pronominal and that are not. There are at least two ways how to do it:

1. One possibility is that the NP is the primitive type (as it is above), and there is an NP-predicate pron:

   (60) $\mathsf{pron} : \mathsf{NP} \rightarrow \mathsf{Bool}$

   This means that the two types $\mathsf{NP}_{\mathsf{pron}}$ and $\mathsf{NP}_{\neg\mathsf{pron}}$ are subtypes partitioning the type NP:

   (61) $\mathsf{NP}_{\mathsf{pron}} \cup \mathsf{NP}_{\neg\mathsf{pron}} = \mathsf{NP}_{\mathsf{pron}\vee\neg\mathsf{pron}} = \mathsf{NP}$

   The grammar ensures that the proper expressions are of the proper subtype of NP. This means that we need to change the definition of the (family of) noun types in (35) in such a way that the result is non-pronominal:

   (62) $\mathsf{N}_{gen,nr,c} := [\textsc{agr}\ \mathsf{AttrAgr}^{\circ}_{gen,nr,c}, \textsc{nagr}\ \langle\mathsf{AttrNon}\rangle] \rightarrow \mathsf{NP}_{gen,nr,c,\neg\mathsf{pron}}$

   For convenience, we can write PPron for $\mathsf{NP}_{\mathsf{pron}}$ and NNP for $\mathsf{NP}_{\neg\mathsf{pron}}$.

2. The other possibility is that the type of personal pronouns PPron and the type of non-pronominal NPs NNP are primitives and the type NP is defined as their supertype:

   (63) $\mathsf{NP} = \mathsf{PPron} + \mathsf{NNP}$

The predicate pron is then

(64) $\mathsf{pron} = \lambda x : \mathsf{NP} \,.\, x :: \mathsf{PPron}$

In either case, we have two types PPron and NNP that are partitioning the type NP.

Morphologically, personal pronouns distinguish gender, person, number and case. In (14) we introduced a function returning case of an NP and in (33), functions returning gender and number. We need to add a similar function returning the person:

(65) $\mathsf{person} : \mathsf{NP} \to \mathsf{Person}$

where the type of person values is simply

(66) $\mathsf{Person} := \{1, 2, 3\}$

All non-pronominal NPs have 3rd person:

(67) $\forall n : \mathsf{NNP} \,.\, n.\mathsf{person} = 3$

Finally, we can define a schematically parametric type of individual personal pronouns:

(68) $\mathsf{PPron}_{gen,p,nr,c} := [\, x : \mathsf{PPron} \,|\, x.\mathsf{person} = p \,\&\, x.\mathsf{gender} = gen \,\&\, x.\mathsf{nr} = nr \,\&\, x.\mathsf{case} = c \,]$

Now, we can add some pronouns to the lexicon (note that non-3rd person pronouns are schematized over gender):

(69) Lexicon (update):

$\mathsf{I}_g : \mathsf{PPron}_{g,\mathsf{1},\mathsf{sg},\mathsf{nom}}$ (*já*)

$\mathsf{you}_{g,\mathsf{sg},\mathsf{nom}} : \mathsf{PPron}_{g,\mathsf{2},\mathsf{sg},\mathsf{nom}}$ (*ty*)

$\mathsf{he} : \mathsf{PPron}_{\mathsf{m},\mathsf{3},\mathsf{sg},\mathsf{nom}}$ (*on*)

$\mathsf{she} : \mathsf{PPron}_{\mathsf{f},\mathsf{3},\mathsf{sg},\mathsf{nom}}$ (*ona*)

$\mathsf{they}_{\mathsf{m}} : \mathsf{PPron}_{\mathsf{m},\mathsf{3},\mathsf{pl},\mathsf{nom}}$ (*oni*)

$\mathsf{they}_{\mathsf{f}} : \mathsf{PPron}_{\mathsf{f},\mathsf{3},\mathsf{pl},\mathsf{nom}}$ (*ony*)

5.1.9.2 Subject – finite verb agreement

As discussed in §A.2.1, the finite verb agrees with the subject in person and number. Nominative noun-phrases require third person in the appropriate number, and all nonstandard things (e.g., partitives, certain numerative phrases) require third person singular, which can be seen as some kind of default.[74]

Agreement is enforced via the subcategorization requirements of the finite verb. For every person and number, the type of the subject is $\mathsf{NP}_{person,number,\mathsf{nom}}$, and in the 3rd person singular it can also be one of the nonstandard things (simplified here as QP). We can thus redefine the $\mathsf{FinVerb}$ type operator to take care of this in the following way:[75]

(70) $\mathsf{FinVerb}(p : \mathsf{Person}, n : \mathsf{Nr}, \mathit{Val} : \mathsf{VALENCY}) =$
$$[\textsc{subj}\ \mathsf{Subj}(p, n), \textsc{adjs}\ \mathsf{Adj}^*] \oplus \mathit{Val} \to \mathsf{S}_{\mathsf{fin}}$$

The operator accepts person, number and specification of complements as parameters and returns the corresponding type of finite verbs. It uses operator Subj to get the proper type of the subject depending on the person and number:

(71) $\mathsf{Subj}(p : \mathsf{Person}, n : \mathsf{Nr}) =$
$$\mathsf{if}\ [p, n] = [3, \mathsf{sg}])\ \mathsf{then}\ (\mathsf{NP}_{p,n,\mathsf{nom}} + \mathsf{QP})\ \mathsf{else}\ \mathsf{NP}_{p,n,\mathsf{nom}}$$

This gives the expected types, for example

(72) $\mathsf{Subj}(1, \mathsf{sg}) = \textsc{subj}\ \mathsf{NP}_{\mathsf{1,sg,nom}}$
$\mathsf{Subj}(3, \mathsf{sg}) = \textsc{subj}\ \mathsf{NP}_{\mathsf{3,sg,nom}} + \mathsf{QP}$

therefore:

(73) $\mathsf{FinVerb}(1, \mathsf{sg}, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}]) = [\textsc{subj}\ \mathsf{NP}_{\mathsf{1,sg,nom}}, \textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S}_{\mathsf{fin}}$
$\mathsf{FinVerb}(3, \mathsf{sg}, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}]) = [\textsc{subj}\ \mathsf{NP}_{\mathsf{3,sg,nom}} + \mathsf{QP}, \textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S}_{\mathsf{fin}}$

Thus we have the tecto terms corresponding to the four finite verbs from the corpus in §5.1.1 have the following types:

[74] This is the case also for verbs like *pršet* ‘rain’ which are usually analyzed as subjectless. Such verbs are not considered here.

[75] This is another (schematically) polymorphic and (schematically) dependent type.

(74) $\mathsf{snore}_{p,n} : \mathsf{FinVerb}(p, n, [])$
$\mathsf{feed}_{p,n} : \mathsf{FinVerb}(p, n, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}])$
$\mathsf{give}_{p,n} : \mathsf{FinVerb}(p, n, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}} \circ \mathsf{NP}_{\mathsf{dat}}])\ \mathsf{know}_{p,n} : \mathsf{FinVerb}(p, n, [\textsc{comps}\ \bar{\mathsf{S}}])$

5.1.9.3 Subject – participle agreement

Participles in periphrastic constructions (and predicative adjectives) agree in number and gender with the subject. Again there is a default form for things like partitives; they require the participle in neuter singular. More details can be found §A.2.1. Similarly to the discussion above, we focus only on past tense and conditional.

Past participle. We will first adjust the types of participles and then of the auxiliaries. The participles, which inflect for gender and number combine only with subjects in the corresponding gender and number. Partitives require neuter singular. We define an operator PstParticiple analogous to the operator FinVerb which requires the proper form of the subject:

(75) $\mathsf{PstParticiple}(g : \mathsf{Gender}, n : \mathsf{Nr}, \mathit{Val} : \mathsf{VALENCY}) =$
$\quad [\textsc{subj}\ \mathsf{Subj}(g, n), \textsc{adjs}\ \mathsf{Adj}^*] \oplus \mathit{Val} \to \mathsf{S}_{\mathsf{pp}}$

The type operator Subj is similar to the above Subj except that it restricts subjects by gender and number and not by person and number:

(76) $\mathsf{Subj}(g : \mathsf{Gender}, n : \mathsf{Nr}) =$
$\quad \mathsf{if}\ [g, n] = [\mathsf{n}, \mathsf{sg}]\ \mathsf{then}\ (\mathsf{NP}_{g,n,\mathsf{nom}} + \mathsf{QP})\ \mathsf{else}\ \mathsf{NP}_{g,n,\mathsf{nom}}$

For example:

(77) $\mathsf{Subj}(\mathsf{m}, \mathsf{sg}) = \mathsf{NP}_{\mathsf{m,sg,nom}}$
$\mathsf{Subj}(\mathsf{n}, \mathsf{sg}) = \mathsf{NP}_{\mathsf{n,sg,nom}} + \mathsf{QP}$

Then

(78) $\mathsf{PstParticiple}(\mathsf{m}, \mathsf{sg}, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}]) = [\mathsf{NP}_{\mathsf{m,sg,nom}}, \textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S}_{\mathsf{pp}}$
$\mathsf{PstParticiple}(\mathsf{n}, \mathsf{sg}, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}]) = [\mathsf{NP}_{\mathsf{n,sg,nom}} + \mathsf{QP}, \textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}, \textsc{adjs}\ \mathsf{Adj}^*] \to \mathsf{S}_{\mathsf{pp}}$

Past participles:

(79) $\mathsf{snored}_{pp,p,n} : \mathsf{PstParticiple}([])$

$\mathsf{fed}_{pp,p,n} : \mathsf{PstParticiple}([\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}])$

$\mathsf{give}_{pp,p,n} : \mathsf{PstParticiple}([\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}} \circ \mathsf{NP}_{\mathsf{dat}}])$

Auxiliary. Now we can turn to the auxiliary. Morphologically, the past auxiliary distinguishes person and number, and the past participles distinguish gender and number. In (55), we assigned the following type to the past auxiliary:

(80) $([\textsc{gov}\ (A \to \mathsf{S}_{\mathsf{pp}})] \oplus A) \to \mathsf{S}_{\mathsf{fin}}$

The subject of the auxiliary must simultaneously satisfy:

1. the agreement requirements of the auxiliary in the same way as subjects of a usual finite verb do;
2. the requirements of the participle, whatever they are. This includes agreement requirements and possibly some other systematic or idiosyncratic restrictions of the subject.

Let's simplify the situation for a moment: ignore the complements and adjuncts of participles. Then, informally, the type of the auxiliary is:

(81) $\mathsf{past}_{p,n} : [\textsc{gov}\ ([\textsc{subj}\ \mathsf{P}] \to \mathsf{S}_{\mathsf{pp}}), \textsc{subj}\ \mathsf{S}] \to \mathsf{S}_{\mathsf{fin}}$

The above restrictions on the subject S can be (again informally) written as:

(82) $\mathsf{S} \sqsubseteq \mathsf{P}\ \&\ \mathsf{S} \sqsubseteq \mathsf{Subj}(p, n)$

This means that the type of the subject S must satisfy the requirements of the participle (S is P or is a subtype of P, $\mathsf{S} \sqsubseteq \mathsf{P}$) but also the agreement requirements of the finite auxiliary ($\mathsf{S} \sqsubseteq \mathsf{Subj}(p, n)$). Adding the adjuncts and complements, the type of the auxiliary would then be a polymorphic dependent type (schematizing over types S, P, X and terms p and n; as discussed in §C.3, the type subscripts on terms are omitted):

(83) $\mathsf{past}_{p,n} : [\ \textsc{gov}\ ([\textsc{subj}\ \mathsf{P}] \oplus \mathsf{X} \to \mathsf{S}_{\mathsf{pp}}),\quad \textsc{subj}\ \mathsf{S}\ ] \oplus \mathsf{X} \to \mathsf{S}_{\mathsf{fin}}$

(where $\mathsf{S} \sqsubseteq \mathsf{P}\ \&\ \mathsf{S} \sqsubseteq \mathsf{Subj}(p, n)$)

Unfortunately, bounding the type S by such condition is clearly beyond the limits of the simple mechanic schematic polymorphism of HOG. There are two options – adopt a more powerful type system (with more complex models) or give up some of the generality of the above type and redefine the auxiliary so that the polymorphism is truly schematic. While we think the former is a better choice, we show how the latter can be done.

We need to explicitly specify the type of subject in the auxiliary. That means several things: (i) the auxiliary must be schematized over gender; (ii) the solution is not truly modular, the agreement for the participle is specified in the auxiliary; (iii) we are giving up the possibility of individual participles to restrict the type of their subjects.[76] Then the type of the past auxiliary can be written as:

(84) $\mathsf{past}_{g,p,n} : [\textsc{gov}\ ([\textsc{subj}\ \mathsf{Subj}(g,n)] \rightarrow \mathsf{S}_{\mathsf{pp}}), \textsc{subj}\ \mathsf{Subj}(g,p,n)] \rightarrow \mathsf{S}_{\mathsf{fin}}$

where $\mathsf{Subj}(g,p,n)$ is informally $\mathsf{Subj}(p,n) \cap \mathsf{Subj}(g,n)$,[77] is defined as:

(85) $\mathsf{Subj}(g : \mathsf{Gender}, p : \mathsf{Person}, n : \mathsf{Nr}) =$
$\quad \mathsf{if}\ [g,p,n] = [\mathsf{n},3,\mathsf{sg}]\ \mathsf{then}\ (\mathsf{NP}_{g,p,n,\mathsf{nom}} + \mathsf{QP})\ \mathsf{else}\ \mathsf{NP}_{g,p,n,\mathsf{nom}}$

and adding the adjuncts and complements, we get the final definition of the past auxiliary:

(86) $\mathsf{past}_{g,p,n} : [\textsc{gov}\ ([\textsc{subj}\ \mathsf{Subj}(g,n)] \oplus \mathsf{X} \rightarrow \mathsf{S}_{\mathsf{pp}}), \textsc{subj}\ \mathsf{Subj}(g,n,p)] \oplus \mathsf{X} \rightarrow \mathsf{S}_{\mathsf{fin}}$
(schematizing over all types X)

The conditional auxiliary can be handled in the exactly same way:

(87) $\mathsf{would}_{g,p,n} : [\textsc{gov}\ ([\textsc{subj}\ \mathsf{Subj}(g,n)] \oplus \mathsf{X} \rightarrow \mathsf{S}_{\mathsf{pp}}), \textsc{subj}\ \mathsf{Subj}(g,n,p)] \oplus \mathsf{X} \rightarrow \mathsf{S}_{\mathsf{fin}}$
(schematizing over all types X)

Obviously, we could also define the type of auxiliaries combining with past participles:

(88) $\mathsf{PapaAux}_{g,p,n} := [\textsc{gov}\ ([\textsc{subj}\ \mathsf{Subj}(g,n)] \oplus \mathsf{X} \rightarrow \mathsf{S}_{\mathsf{pp}}), \textsc{subj}\ \mathsf{Subj}(g,n,p)] \oplus \mathsf{X} \rightarrow \mathsf{S}_{\mathsf{fin}}$
(schematizing over all types X)

[76] Subtyping of functions is contravariant in their arguments, i.e., $(T \rightarrow X) \sqsubseteq (U \rightarrow X)$ iff $U \sqsubseteq T$. Informally, where a function accepting U is expected, we need a function accepting at least U (if we need a function accepting even integers, a function accepting integers will do). Therefore a type of a participle making some idiosyncratic requirements on its subjects is not a subtype of a type without such requirements.

[77] In fact, we could have defined the other two operators in terms of this one.

In the following section we show an alternative possibility for handling agreement – agreement is not expressed as part of subcategorization in tectogrammar but instead as constraints over whole signs.

5.1.10 Full simple tectogrammar of Czech

5.1.10.1 Nouns, Pronouns, NPs

1. Basic types: NP, AP, Det, Poss
2. Agreement and similar types:

   $\mathsf{Gender} := \{\mathsf{m}, \mathsf{i}, \mathsf{f}, \mathsf{n}\}$

   $\mathsf{Person} := \{1, 2, 3\}$

   $\mathsf{Nr} := \{\mathsf{sg}, \mathsf{pl}\}$

   $\mathsf{Case} := \{\mathsf{nom}, \mathsf{gen}, \mathsf{dat}, \mathsf{acc}, \mathsf{voc}, \mathsf{loc}, \mathsf{ins}\}$
3. Agreement and similar features:

   $\mathsf{gender} : \{\mathsf{NP}, \mathsf{AP}, \mathsf{Det}, \mathsf{Poss}\} \rightarrow \mathsf{Gender}$

   $\mathsf{person} : \mathsf{NP} \rightarrow \mathsf{Person}$

   $\mathsf{nr} : \{\mathsf{NP}, \mathsf{AP}, \mathsf{Det}, \mathsf{Poss}\} \rightarrow \mathsf{Nr}$

   $\mathsf{case} : \{\mathsf{NP}, \mathsf{AP}, \mathsf{Det}, \mathsf{Poss}\} \rightarrow \mathsf{Case}$
4. Pronouns and nonpronouns:

   $\mathsf{pron} : \mathsf{NP} \rightarrow \mathsf{Bool}$

   $\mathsf{PPron} := \mathsf{NP}_{\mathsf{pron}}$

   $\mathsf{NNP} := \mathsf{NP}_{\neg\mathsf{pron}}$
5. NP structure and agreement:

   indexes: $\textsc{agr}, \textsc{nagr}$

   $\mathsf{N}_{g,n,c} := [\textsc{agr}\ \mathsf{AttrAgr}^{*}_{g,n,c}, \textsc{nagr}\ \mathsf{AttrNon}^{*}] \rightarrow \mathsf{NNP}_{g,n,c}$

   $\mathsf{AttrAgr} := \mathsf{AP} + \mathsf{Det} + \mathsf{Poss}$

   $\mathsf{AttrNon} := \mathsf{PP} + \mathsf{NP}_{\mathsf{gen}} + \mathsf{NP}_{\mathsf{dat}}$

5.1.10.2 PPs

1. Basic type: PP
2. PP predicates for each preposition: $\mathsf{na\text{-}acc} : \mathsf{PP}$ (na)

5.1.10.3 Verbs

1. Indexes for products used for valencies:

   $\textsc{subj}, \textsc{comps}, \textsc{adjs}, \textsc{gov}$

2. Clauses:

   Basic type: S

   Various types of clauses ($\mathsf{S}_{\mathsf{fin}}, \mathsf{S}_{\mathsf{pp}}, \mathsf{S}_{\mathsf{inf}}$) are defined by S-predicates: fin, pp, inf

3. Valency and agreement:

   $\mathsf{Adj} := \mathsf{AdvP} + \mathsf{PP}$

   $\mathsf{VALENCY} = \{\ [], [\textsc{comps}\ \mathsf{Tecto}]\}$

   $\mathsf{FinVerb}(p : \mathsf{Person}, n : \mathsf{Nr}, \mathit{Val} : \mathsf{VALENCY}) :=$
   $\quad [\textsc{subj}\ \mathsf{Subj}(p,n), \textsc{adjs}\ \mathsf{Adj}^*] \oplus \mathit{Val} \rightarrow \mathsf{S}_{\mathsf{fin}}$

   $\mathsf{PstParticiple}(g : \mathsf{Gender}, n : \mathsf{Nr}, \mathit{Val} : \mathsf{VALENCY}) :=$
   $\quad [\textsc{subj}\ \mathsf{Subj}(g,n), \textsc{adjs}\ \mathsf{Adj}^*] \oplus \mathit{Val} \rightarrow \mathsf{S}_{\mathsf{fin}}$

   $\mathsf{PapaAux}(g : \mathsf{gender}, p : \mathsf{Person}, n : \mathsf{Nr}) :=$
   $\quad [\textsc{gov}\ ([\textsc{subj}\ \mathsf{Subj}(g,n)] \oplus X \rightarrow \mathsf{S}_{\mathsf{pp}}), \textsc{subj}\ \mathsf{Subj}(g,p,n)] \oplus X \rightarrow \mathsf{S}_{\mathsf{fin}}$

   $\mathsf{Subj}(p : \mathsf{Person}, n : \mathsf{Nr}) =$
   $\quad \mathsf{if}\ [p,n] = [3,\mathsf{sg}]\ \mathsf{then}\ (\mathsf{NP}_{p,n,\mathsf{nom}} + \mathsf{QP})\ \mathsf{else}\ \mathsf{NP}_{p,n,\mathsf{nom}}$

   $\mathsf{Subj}(g : \mathsf{Gender}, n : \mathsf{Nr}) =$
   $\quad \mathsf{if}\ [g,n] = [\mathsf{n},\mathsf{sg}]\ \mathsf{then}\ (\mathsf{NP}_{g,n,\mathsf{nom}} + \mathsf{QP})\ \mathsf{else}\ \mathsf{NP}_{g,n,\mathsf{nom}}$

   $\mathsf{Subj}(g : \mathsf{Gender}, p : \mathsf{Person}, n : \mathsf{Nr}) =$
   $\quad \mathsf{if}\ [g,p,n] = [\mathsf{n},3,\mathsf{sg}]\ \mathsf{then}\ (\mathsf{NP}_{g,p,n,\mathsf{nom}} + \mathsf{QP})\ \mathsf{else}\ \mathsf{NP}_{g,p,n,\mathsf{nom}}$

5.1.10.4 The Rest

1. types (treated as primitives) AdvP, QP

5.1.10.5 Lexicon

1. Nouns:

   $\mathsf{adam}_{\mathsf{N}} : \mathsf{N}_{\mathsf{m},\mathsf{sg},\mathsf{nom}}$ (*Adam*)

   $\mathsf{petr}_{\mathsf{N}} : \mathsf{N}_{\mathsf{m},\mathsf{sg},\mathsf{nom}}$ (*Petr*)

$\mathsf{eva}_{\mathsf{D}} : \mathsf{N}_{\mathsf{f,sg,dat}}$ (*Evě*)

$\mathsf{goat}_{\mathsf{A}} : \mathsf{N}_{\mathsf{f,sg,acc}}$ (*kozu*)

$\mathsf{pear}_{\mathsf{A}} : \mathsf{N}_{\mathsf{f,sg,acc}}$ (*hrušku*)

$\mathsf{meadow}_{\mathsf{L}} : \mathsf{N}_{\mathsf{f,sg,loc}}$ (*louce*)

$\mathsf{tree}_{\mathsf{I}} : \mathsf{N}_{\mathsf{m,sg,ins}}$ (*stromem*)

$\mathsf{grandpa}_{\mathsf{G}} : \mathsf{N}_{\mathsf{f,sg,loc}}$ (*dědy*)

2. Adjectives, ...:

$\mathsf{that}_{\mathsf{f.sg.A}} : \mathsf{Det}_{\mathsf{f,sg,acc}}$ (*tu*)

$\mathsf{our}_{\mathsf{f.sg.A}} : \mathsf{Poss}_{\mathsf{f,sg,acc}}$ (*naší*)

$\mathsf{our}_{\mathsf{m.sg.N}} : \mathsf{Poss}_{\mathsf{m,sg,nom}}$ (*náš*)

$\mathsf{little}_{\mathsf{m.sg.N}} : \mathsf{AP}_{\mathsf{m,sg,nom}}$ (*malý*)

$\mathsf{old}_{\mathsf{f.sg.A}} : \mathsf{AP}_{\mathsf{f,sg,acc}}$ (*starou*)

3. Personal Pronouns:

$\mathsf{I}_{\mathsf{m}} : \mathsf{PPron}_{\mathsf{1,m,sg,nom}}$ (*já*)

$\mathsf{I}_{\mathsf{f}} : \mathsf{PPron}_{\mathsf{1,f,sg,nom}}$ (*já*)

$\mathsf{you}_{\mathsf{m,sg}} : \mathsf{PPron}_{\mathsf{2,m,sg,nom}}$ (*ty*)

$\mathsf{he} : \mathsf{PPron}_{\mathsf{3,m,sg,nom}}$ (*on*)

$\mathsf{she} : \mathsf{PPron}_{\mathsf{3,f,sg,nom}}$ (*ona*)

$\mathsf{they}_{\mathsf{m}} : \mathsf{PPron}_{\mathsf{3,m,pl,nom}}$ (*oni*)

$\mathsf{they}_{\mathsf{f}} : \mathsf{PPron}_{\mathsf{3,f,pl,nom}}$ (*ony*)

4. Auxiliary verbs:

$\mathsf{past}_{g,p,n} : \mathsf{PapaAux}(g,p,n)$

$\mathsf{would}_{g,p,n} : \mathsf{PapaAux}(g,p,n)$

5. Finite verbs:

$\mathsf{snore}_{p,n} : \mathsf{FinVerb}(p,n,[])$ (*chrápu,..*)

$\mathsf{feed}_{p,n} : \mathsf{FinVerb}(p,n,[\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}])$ (*krmím,..*)

$\mathsf{give}_{p,n} : \mathsf{FinVerb}(p,n,[\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}} \circ \mathsf{NP}_{\mathsf{dat}}])$ (*dám,..*)

$\mathsf{know}_{p,n} : \mathsf{FinVerb}(p,n,[\textsc{comps}\ \bar{\mathsf{S}}])$ (*vím,..*)

$\mathsf{force}_{p,n} : \mathsf{FinVerb}(p, n, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}} \circ \mathsf{PP}_{\mathsf{k}}])$ (*donutím,..*)

$\mathsf{depend}_{p,n} : \mathsf{FinVerb}(p, n, [\textsc{comps}\ \mathsf{PP}_{\mathsf{na\text{-}loc}}])$ (*závisím*)

6. Past participles:

   $\mathsf{snored}_{pp,g,n} : \mathsf{PstParticiple}(g, n, [])$ (*chrápal,..*)

   $\mathsf{fed}_{pp,g,n} : \mathsf{PstParticiple}(g, n, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}])$ (*krmil,..*)

   $\mathsf{give}_{pp,g,n} : \mathsf{PstParticiple}(g, n, [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}} \circ \mathsf{NP}_{\mathsf{dat}}])$ (*dal,..*)

   $\mathsf{know}_{pp,g,n} : \mathsf{PstParticiple}(g, n, [\textsc{comps}\ \bar{\mathsf{S}}])$ (*věděl,..*)

7. Prepositions

   $\mathsf{k}_{dat} : [\textsc{comps}\ \mathsf{NP}_{\mathsf{dat}}] \rightarrow \mathsf{PP}_{\mathsf{k\text{-}dat}}$ (*k*)

   $\mathsf{na}_{acc} : [\textsc{comps}\ \mathsf{NP}_{\mathsf{loc}}] \rightarrow \mathsf{PP}_{\mathsf{na\text{-}acc}}$ (*na*)

   $\mathsf{na}_{loc} : [\textsc{comps}\ \mathsf{NP}_{\mathsf{loc}}] \rightarrow \mathsf{PP}_{\mathsf{na\text{-}loc}}$ (*na*)

   $\mathsf{pod}_{acc} : [\textsc{comps}\ \mathsf{NP}_{\mathsf{acc}}] \rightarrow \mathsf{PP}_{\mathsf{pod\text{-}acc}}$ (*pod*)

   $\mathsf{pod}_{ins} : [\textsc{comps}\ \mathsf{NP}_{\mathsf{ins}}] \rightarrow \mathsf{PP}_{\mathsf{pod\text{-}ins}}$ (*pod*)

   $\mathsf{od}_{gen} : [\textsc{comps}\ \mathsf{NP}_{\mathsf{gen}}] \rightarrow \mathsf{PP}_{\mathsf{od\text{-}gen}}$ (*od*)

8. Various

   $\mathsf{that} : [\textsc{spec}\ \mathsf{S}] \rightarrow \bar{\mathsf{S}}$ (*že*)

   $\mathsf{often}, \mathsf{horribly}, \mathsf{again} : \mathsf{AdvP}$

5.2 Combining signs II.

As explained in Chapter 2, the set of signs, i.e., the possible tuples of pheno and tecto (and semantic) terms, is specified recursively. The lexicon lists the basic signs and then there are constraints determining possible combination of signs. In some cases, the combination in one grammar component corresponds to the same combination in another component. For instance in §2.8, we assume that tuples in tecto correspond to tuples in pheno. In some cases, only one component is changed. For example, type embedding in tecto does not affect the corresponding pheno in any way. However, the relation can be more complex. This is especially true in the case of function application. Specifying a pheno object corresponding to application of a functor to its arguments is non-trivial. In this section, we develop a simple framework for constraining such combinations of signs.

The English toy grammar in Chapter 2 had only two application constraints:

(89) $\vdash \forall h : \mathsf{Sign}([\textsc{spec}\ \mathsf{N}] \to \mathsf{NP}, \mathsf{Pheno})$
$\quad\quad \forall a : \mathsf{Sign}([\textsc{spec}\ \mathsf{N}], [\textsc{spec}\ \mathsf{Pheno}])$
$\quad\quad \exists m : \mathsf{Sign}(\mathsf{NP}, \mathsf{Pheno})\,.$
$\quad m.\textsc{tecto} = (h.\textsc{tecto})(a.\textsc{tecto})\ \&$
$\quad m.\textsc{pheno} = h.\textsc{pheno} \circ a.\textsc{pheno}.\textsc{spec}$

(90) $\vdash \forall h : \mathsf{Sign}([\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}, \mathsf{Pheno})$
$\quad\quad \forall a : \mathsf{Sign}([\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}], [\textsc{subj}\ \mathsf{Pheno}, \textsc{comps}\ \mathsf{Pheno}])$
$\quad\quad \exists m : \mathsf{Sign}(\mathsf{S}, \mathsf{Pheno})\,.$
$\quad m.\textsc{tecto} = (h.\textsc{tecto})(a.\textsc{tecto})\ \&$
$\quad m.\textsc{pheno} = a.\textsc{pheno}.\textsc{subj} \circ h.\textsc{pheno} \circ a.\textsc{pheno}.\textsc{comps}$

The first specifies that in NPs, the determiner comes before the noun. The second says that the combination of a transitive verb with its arguments in tecto corresponds to a concatenation of the pheno of the subject, the pheno of the verb and the pheno of the object.

While the constraints might look complicated, their structure is rather simple. Schematically, they can be written as (I is the set of indexes, e.g., $I = \{\textsc{subj}, \textsc{comps}\}$):

(91) $\vdash \forall h : \mathsf{Sign}([\textsc{i}\colon A_i]_{\textsc{i}\in I} \to B, \mathsf{Pheno})$
$\quad\quad \forall a : \mathsf{Sign}([\textsc{i}\colon A_i]_{\textsc{i}\in I}, [\textsc{i}\colon \mathsf{Pheno}]_{\textsc{i}\in I})$
$\quad\quad \exists m : \mathsf{Sign}(B, \mathsf{Pheno})\,.$
$\quad m.\textsc{tecto} = (h.\textsc{tecto})(a.\textsc{tecto})\ \&$
$\quad m.\textsc{pheno} = p(h.\textsc{pheno}, a.\textsc{pheno})$

where p is a function specifying m's pheno in terms of concatenation of the pheno of the head ($h.\textsc{pheno}$) and the phenos of the individual arguments (i.e., the projections of $a.\textsc{pheno}$, e.g., $a.\textsc{pheno}.\textsc{subj}$). In (90), $[\textsc{i}\colon A_i]_{\textsc{i}\in I}$ is $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$ and B is S. The structure of such constraint can be depicted in the following way, reminiscent of the HPSG phrase structure schemata:

(92)

$$\mathsf{m} = \begin{bmatrix} \textsc{tecto} & \mathsf{h}.\textsc{tecto}(\mathsf{a}.\textsc{tecto}) \\ \textsc{pheno} & p(\mathsf{h}.\textsc{pheno}, \mathsf{a}.\textsc{pheno}) \end{bmatrix} : \begin{bmatrix} \textsf{Sign} & \\ \textsc{tecto} & B \\ \textsc{pheno} & \mathsf{Pheno} \end{bmatrix}$$

$$\mathsf{h} : \begin{bmatrix} \textsf{Sign} & \\ \textsc{tecto} & [\textsc{i}\colon A_i]_{\textsc{i}\in I} \to \mathsf{B} \\ \textsc{pheno} & \mathsf{Pheno} \end{bmatrix} \qquad \mathsf{a} : \begin{bmatrix} \textsf{Sign} & \\ \textsc{tecto} & [\textsc{i}\colon A_i]_{\textsc{i}\in I} \\ \textsc{pheno} & [\textsc{i}\colon \mathsf{Pheno}]_{\textsc{i}\in I} \end{bmatrix}$$

The instantiation of the schema is uniquely determined by the tecto type of the arguments ($[\textsc{i}\colon A_i]_{\textsc{i}\in I}$), the tecto type of the result (B) and the function p. In the case of our simple grammar, this corresponds to the following information:

(93)

	arguments' type	result's type	$p = \lambda hp, ap\,.$
a	$[\textsc{spec}\ \mathsf{N}]$	NP	$hp \circ ap.\textsc{spec}$
chased	$[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$	S	$ap.\textsc{subj} \circ hp \circ ap.\textsc{comps}$

Such schematic presentation makes the constraints significantly more transparent. However, in real grammars we need far more flexibility. Usually, the possible pheno combinations will not be determined by the sub-categorization and type of the head in such a simple way. For example, the ordering constraints of the Czech grammar presented in this chapter are more complex and depend on various factors. These include information structure (theme comes before rheme regardless of its tecto type), structure of the derivation (clitics with the same governor act differently than clitics with different governors), lexical properties of words (weak pronouns are positioned differently than strong pronouns), etc. Moreover, we may want a single combination in tecto to correspond to several different phenos, thus the expressing mother's pheno as function of daughters may be too restrictive.

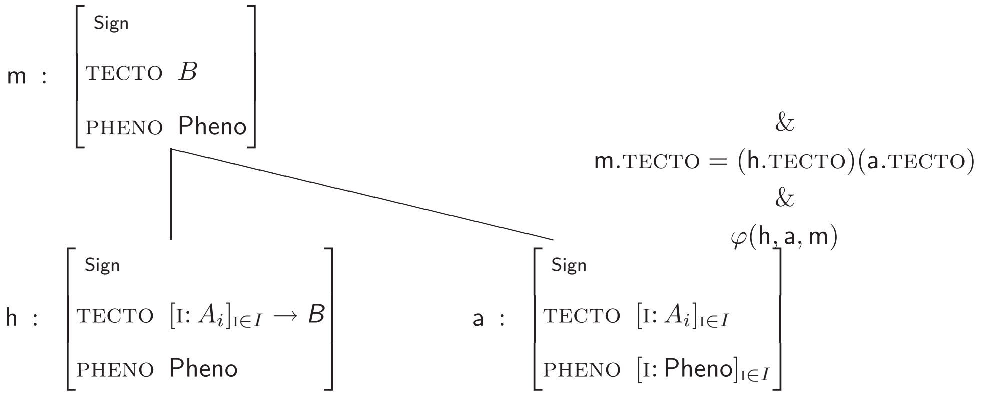
Because of these reasons, we generalize the schematic constraint in (91) in the following way:

(94)

$$\begin{aligned} \vdash\ & \forall h : \mathsf{Sign}([\textsc{i}\colon A_i]_{\textsc{i}\in I} \to \mathsf{B}, \mathsf{Pheno})\ \forall a : \mathsf{Sign}([\textsc{i}\colon A_i]_{\textsc{i}\in I}, [\textsc{i}\colon \mathsf{Pheno}]_{\textsc{i}\in I}) \\ & \qquad \forall m : [\textsc{tecto}\ \mathsf{B}, \textsc{pheno}\ \mathsf{Pheno}]\,. \\ & m.\textsc{tecto} = (h.\textsc{tecto})(a.\textsc{tecto})\ \& \\ & \varphi(h, a, m) \Rightarrow m :: \mathsf{Sign} \end{aligned}$$

This constraint states that the head h and its arguments a can be combined in all possible ways into complex signs, as long as each of the new sign's tectogrammar corresponds to tecto application of h on a and the predicate φ is satisfied. The predicate φ is parametrized by whole signs, not just their phenos, it may thus refer to their tecto (or semantic) components as well.

(95)



Therefore, to constrain the order within an NP, we have to specify a constraint for the instantiation of the schema where

(96) $A = [\text{I: } A_i]_{\text{I} \in I} = [\text{SPEC } \mathsf{N}]$ and $B = \mathsf{NP}$

If we specify the predicate φ as (we omit typing on the three parameters, since it is given by A and B):

(97) $\lambda h, a, m \,.\, m.\text{PHENO} = h.\text{PHENO} \circ a.\text{SPEC.PHENO}$

a determiner would precede its noun. If however, the predicate were as follows

(98) $\lambda h, a, m \,.$
$$m.\text{PHENO} = h.\text{PHENO} \circ a.\text{SPEC.PHENO} \vee$$
$$m.\text{PHENO} = a.\text{SPEC.PHENO} \circ h.\text{PHENO}$$

both orders would be possible. Finally, the following predicate

(99) $\lambda h, a, m \,.\, \mathsf{true}$

means, that the determiner and the noun can be combined in an infinite number of ways. The resulting signs would always have $(h.\text{TECTO})(a.\text{TECTO})$ as their tecto, but their pheno would be any possible pheno object.

5.2.0.6 Format of constraints

The grammar uses several conventions to specify the set of constraints on combination corresponding to tecto application. As implied above, a constraint in the shape of (94) is uniquely determined by the following three properties:

1. the type of the tecto argument, $A = [\textsc{i}\colon A_i]_{\textsc{i} \in I}$
2. the type of the tecto mother, B
3. the boolean predicate $\varphi(h : \mathsf{Sign}(A \to B), a : \mathsf{Sign}(A), m : [\textsc{tecto}\ B, \textsc{pheno}\ \mathsf{Pheno}]) : \mathsf{Bool}$

We write such constraints in the form of

(100) $\vdash_{\mathsf{app}} A, B : \quad \varphi$

A and B are understood as polymorphic bounds, therefore this is a schema instantiated by all types A' and B' such that $A' \sqsubseteq A$ (A' is a subtype or equal to A) and $B' \sqsubseteq B$.

(101) $\vdash_{\mathsf{app}} \mathsf{Tecto}, \mathsf{Tecto} : \quad \varphi$

is written simply as

(102) $\vdash_{\mathsf{app}} \varphi$

We also conventionally use h, a and m for the three free variables in φ and omit the lambda binder. Therefore the constraint on combination of a determiner and a noun in (89) is written as

(103) $\vdash_{\mathsf{app}} [\textsc{spec}\ \mathsf{N}], \mathsf{NP} : \quad \mathsf{m}.\textsc{pheno} = \mathsf{h}.\textsc{pheno} \circ \mathsf{a}.\textsc{spec}.\textsc{pheno}$

and the constraint on a combination of a finite verb and a subject in (90) is written as:

(104) $\vdash_{\mathsf{app}} [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}], \mathsf{S} : \quad \mathsf{m}.\textsc{pheno} = \mathsf{a}.\textsc{subj}.\textsc{pheno} \circ \mathsf{h}.\textsc{pheno}$

Moreover, a conjunctive constraint may be written as several constraints. Therefore writing

(105) $\vdash_{\mathsf{app}} \mathsf{A}, \mathsf{B} : \quad \varphi_1 \ \&\ \varphi_2$

and

(106) $\vdash_{\mathsf{app}} \mathsf{A}, \mathsf{B} : \quad \varphi_1$
$\vdash_{\mathsf{app}} \mathsf{A}, \mathsf{B} : \quad \varphi_2$

is equivalent. Note that this interacts with the polymorphic bounds, therefore if a grammar contains the following two schemata:

(107) $\vdash_{\mathsf{app}} [\textsc{spec}\ \mathsf{N}], \mathsf{NP}: \quad \varphi_1$

$\vdash_{\mathsf{app}} \mathsf{Tecto}, \mathsf{Tecto}: \quad \varphi_2$

the combination resulting into a noun phrase is constrained by both φ_1 and φ_2.

App functions. More complex constraints use functions, both to capture linguistic generalization and to make the grammar modular and transparent. Many of such functions need access to $\mathsf{m}, \mathsf{h}, \mathsf{a}$, the three signs related by the constraint. To simplify the notation, we assume the three signs (and similarly the objects introduced below) are passed as implicit parameters. Such functions are called *application functions* or more generally *linearization functions* and we mark them by a subscript a:

(108) $\vdash_{\mathsf{app}} A, B: \quad \mathsf{m}.\textsc{pheno} = \mathsf{some\text{-}fnc}_a$

$\mathsf{some\text{-}fnc}_a : \mathsf{Pheno} := \mathsf{some\text{-}other\text{-}fnc}_a$

$\mathsf{some\text{-}other\text{-}fnc}_a : \mathsf{Pheno} := \mathsf{h}.\textsc{pheno} \circ \mathsf{a}.\textsc{pheno}$

5.2.1 Individual arguments

Because (i) many of the constraints take the argument tuple apart and then refer to its individual components, (ii) some of the constraints work with the arguments as with collection of signs, we provide slightly more comfort to avoid repetitive processing of the h and a signs. In addition to providing the three basic signs, referred to as m, h and a in the constraints, we also provide additional "preprocessed" information:

- nhDtrs : Set(Sign) – the non-head daughters, i.e., the individual components of the a tuple. Members of featured lists (individual adjuncts, attributes) are in nhDtrs individually.
- $\mathsf{dtrs} = \{\mathsf{h}\} \cup \mathsf{nhDtrs}$ – all daughters.
- fnc : Sign → Fnc – function returning for every sign in nhDtrs and dtrs a term that is isomorphic with the tuple indexes (subj for SUBJ, comps for COMPS, . . .) or head for the head.

5.2.2 Subject-Predicate Agreement Revisited

The purpose of the above constraints is to constrain the combination of whole signs. However, we can also use them to impose constraints on tecto terms, or more precisely, say that only some tecto terms can be used in signs. Below, we show how this can be used to avoid the problems we encountered when formalizing subject-verb agreement in §5.1.9.

Subject-Finite Verb Agreement. Using the constraints on combination of signs, we can capture subject-verb agreement in the following way:

```
(109) (Subject Predicate Agreement)
      $\vdash_{\mathsf{app}} [\textsc{subj}\ \mathsf{Tecto}]^{\oplus}, \mathsf{S_{fin}}$ :
          if $(s :: \mathsf{NP_{nom}})$
              $v.\mathsf{person} = s.\mathsf{person}\ \&\ v.\mathsf{nr} = s.\mathsf{nr}$
          else
              $v.\mathsf{person} = 3\ \&\ v.\mathsf{nr} = \mathsf{sg}$
          where
              $v = \mathsf{h}.\textsc{tecto}$
              $s = \mathsf{a}.\textsc{tecto}.\textsc{subj}$
```

The constraint applies to all combinations of sings when the argument sign has a tuple with a subject in its tecto,[78] and the result is a finite sentence sign. The rest states directly what we said informally: a nominative noun phrase agrees with the finite verb in number and person, any other subject requires the verb to be in the default form, which means 3rd person singular.

This means the type of finite verbs only specifies the subcategorization requirements and leaves enforcing agreement to the constraint above. It is possible to derive a tecto term by combining a subject in singular with a verb in plural. However, because of the agreement constraint we cannot pronounce such sentence, there is no sign that would contain it as its tecto component. We could handle the gender-number agreement of participles in a similar way. The simplification of the complex tecto type would be even more significant.

Neither of the two suggested choices is completely satisfactory. The treatment in §5.1.9 mixes subcategorization and agreement into a single type. If a verb had some idiosyncratic requirements on the type of its subjects, we would need to handle its agreement as special as well, even if it were perfectly regular. The other option, presented in this section, is that agreement is handled by constraints over combinations of whole signs. It separates subcategorization and agreement but it also means that agreement is handled outside of tectogrammar – tectogrammar overgenerates and terms with incorrect agreement are just not used in signs. It is possible to imagine arguments for

[78] $[\textsc{subj}\ \mathsf{Tecto}]^{\oplus}$ is a supertype of all record types containing SUBJ Tecto.

either of these options.[79] The problem is that here our choices are not results of answering such linguistic questions but are forced upon us by the formal properties of the formalism.

In §5.1.9, we suggested that at least some of the problems would be alleviated by using a more powerful polymorphic type system. Another solution would be to use the same strategy for objects in the grammar components as we use for whole signs. Similarly as not all possible tecto-pheno tuples are considered signs, we could also introduce a notion of "correct" tectogrammatical expressions. This approach would be very similar to the approach used by HPSG, where set of possible objects is determined by constraints over a type signature. We leave this problem for further research.

5.3 Inspiration from HPSG

In this section, we summarize some of the important points of the major HPSG framework for dealing with discontinuities and linearization. The framework was introduced by Reape (1994) and later extended by Kathol (Kathol 1995, 2000*a*; Kathol and Pollard 1995) and Penn (1999*a*,*b*). In the next sections, we use it as a basis for linearization in HOG.

5.3.1 Reapes's framework

Liberating word-order domains. In a standard HPSG, ordering constraints for a phrase are stated over the set of its immediate daughters (with non-local phenomena handled via the SLASH featur). Reape's theory (Reape 1994, 1996) detaches the syntactic hierarchy from the word-order domains. He encodes the syntactic hierarchy via the standard DTRS feature containing the list of immediate daughters and the domains via an additional feature DOM containing a list of signs that can be ordered in that particular phrase. The mother's domain is created on the basis of daughters' domains by one of the following two operations:

- *domain-insertion*: The daughter's domain is a member of her mother's domain. It means it is a single compact unbreakable unit relative to its mother's domain.

  This ensures that the daughter's domain will be realized as a continuous string. Also, the LP constraints of the mother cannot constrain the internal structure of that domain. This corresponds to the locality of context-free grammar.

[79] For example, in Czech a realistically modeled agreement must refer to word order in some way, because agreement with a coordinated subject depends on the relative order of the subject and predicate.

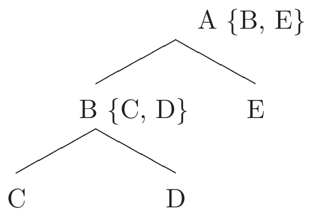
- *domain-union*: The daughter's domain extends her mother's domain. Therefore, the daughter's domain does not have to be realized as a continuous string. Also, the LP constraints of the mother can constrain the internal structure of that domain.

  The order of members must respect the order in an embedded domain. Informally, once the domain objects are ordered, they can be separated by objects from other domains, but the relative order cannot change.

In a context free grammar, all domains are constructed by insertion of daughters' domains. This means the domain list and the list of daughters coincide. In the following text, we also say that the members of the domain in the insertion case are compacted into the higher domain and in the domain-union case are liberated into it.

Example. Consider the tree in (110) (the word-order domains are written as sets of phrases following the label of a node.) In this case, all domains were inserted, thus the domains at the level of A and B correspond to the set of their immediate daughters. The linear order in the tree is the result of the constraints in (111).

(110)



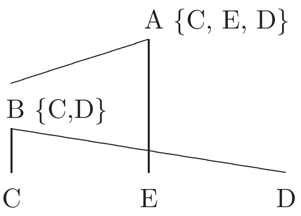
(111) $\mathsf{C} \prec \mathsf{D} \,\&\, \mathsf{B} \prec \mathsf{E}$

Adding a constraint in (112), has no effect on the tree because there is no domain where C could be ordered relative to E and and E relative to D.

(112) $\mathrm{C} \prec \mathrm{E} \prec \mathrm{D}$

To be able to constrain the order of E relative to C or D, we need to extend the domain of A by the domain of B, which is done via domain-union. Now, when we replace the local constraints in (111) by the constraint in (112), we get the tangled tree in (113), where the string corresponding to B is not continuous.

(113)



5.3.2 Kathols's framework

Kathol and Pollard (Kathol 1995, 2000*a*; Kathol and Pollard 1995) modify and extends Reape's formalism in several ways. The most important are the following three modifications:

1. Integration with the traditional idea of topological fields (see, e.g., Höhle 1986) – the LP constraints are not specified in terms of syntactic categories but in terms of abstract word-order categories,

2. Simpler domain objects. While Reape's domains contain full HPSG signs including recursively their syntactic daughters and domains, Kathol's domain objects essentially contain just the phonological string and the topological field.

3. Partial compaction. To account for certain word order phenomena (e.g., German extraposition), they introduces partial compaction. Partial compaction, increases the flexibility of domains. It allows to liberate only a portion of a daughter's domain, while the rest is inserted into mother's domain. For example, in a noun phrase, the determiner, an adjective and the noun can be compacted into a single unit, while a dependent relative clause may remain free to be ordered independently in mother's domain. In the following fragment the NP *einen Hund der Hunger hat* is partially compacted. *einen Hund* is inserted into the clausal domain while the rest (in this case only a single item, *der Hunger hat*) is liberated:

   (114) . . . dass Karl [einen Hund] füttert [der Hunger hat].
   . . . that Karl a dog feeds that hunger has.
   . . . that Karl feeds a dog that is hungry.

Partial compaction Formally, partial compaction is a relation between three objects: a sign ($\boxed{d}$, a daughter), a domain object ($\boxed{c}$) and a list of objects ($\boxed{ls}$). The domain list of the daughter is split (not necessarily continuously) into two sublists – the liberated portion ($\boxed{ls}$) and the portion ($\boxed{cs}$) that is compacted into ($\boxed{c}$). In HPSG this can be written as (the notation is slightly modified so that it is more similar to the HOG notation used here):

(115) $\textsf{p-compaction}(\boxed{\text{d}},\boxed{\text{c}},\boxed{\text{ls}}) \Leftrightarrow$

$$\boxed{\text{d}} : \begin{bmatrix} \textsf{Sign} & \\ \text{synsem} & \boxed{\text{ss}} \\ \text{dom} & \boxed{\text{os}} \end{bmatrix} \& \boxed{\text{c}} : \begin{bmatrix} \textsf{Dom} & \\ \text{synsem} & \boxed{\text{ss}} \\ \text{phon} & \textsf{join}_{\textsf{phon}}(\boxed{\text{cs}}, \boxed{\text{phons}}) \end{bmatrix} \& \textsf{shuffle}(\{\boxed{\text{cs}},\boxed{\text{ls}}\},\boxed{\text{os}})$$

where $\textsf{join}_{\textsf{phon}}$ is relation between a list of domain objects and the concatenation of their phonologies. It is a schematic relation which takes a list $\boxed{\text{xs}}$, maps it by a function f and concatenates the result, in HPSG, this can be defined recursively as follows:

(116) $\textsf{join}_f(\boxed{\text{xs}},\boxed{\text{ys}}) :=$

$(\boxed{\text{xs}} : \langle\rangle \;\&\; \boxed{\text{ys}} : \langle\rangle) \vee$

$(\textsf{cons}([f\ \boxed{\text{y}}], \boxed{\text{xtail}}, \boxed{\text{xs}}) \;\&\; \textsf{join}_f(\boxed{\text{xtail}},\boxed{\text{ytail}}) \;\&\; \boxed{\text{ys}} = \boxed{\text{y}} \circ \boxed{\text{ytail}})$

Reape (1994)'s full compaction is then a special case of partial compaction – the list of liberated domain objects ($\boxed{\text{ls}}$) is empty:

(117) $\textsf{compaction}(\boxed{\text{s}},\boxed{\text{c}}) \Leftrightarrow \textsf{p-compaction}(\boxed{\text{s}},\boxed{\text{c}},\langle\rangle)$

Note that (Kathol and Pollard 1995) do not require the compacted sublist (cs) to be continuous in s's domain. We do not see any linguistic motivation for such unconstrained compaction, moreover it results in many spurious ambiguities. This is significantly restricted by Yatabe (1996) who assumes the compacted objects to be a continuous prefix (for head-first languages) or a continuous suffix (for head-last languages) of the daughters domain list ($\boxed{\text{s}}$.DOM). We can make $\boxed{\text{cs}}$ continuous by replacing the $\textsf{shuffle}$ relation with concatenation:

(118) $\boxed{\text{s}}\textsf{.dom} = \boxed{\text{ls1}} \circ \boxed{\text{cs}} \circ \boxed{\text{ls2}} \quad \& \quad \boxed{\text{ls}} = \boxed{\text{ls1}} \circ \boxed{\text{ls2}}$

To require the compacted object to correspond to a prefix of the original domain list, we can set $\boxed{\text{ls1}}$ to be empty:

(119) $\textsf{prefix-compaction}(\boxed{\text{s}},\boxed{\text{c}},\boxed{\text{ls}}) \Leftrightarrow$

$$\boxed{\text{s}} : \begin{bmatrix} \textsf{Sign} & \\ \text{synsem} & \boxed{\text{ss}} \\ \text{dom} & \boxed{\text{os}} \end{bmatrix} \& \boxed{\text{c}} : \begin{bmatrix} \textsf{Dom} & \\ \text{synsem} & \boxed{\text{ss}} \\ \text{phon} & \textsf{join}_{\textsf{phon}}(\boxed{\text{cs}}) \end{bmatrix} \& \boxed{\text{os}} = \boxed{\text{cs}} \circ \boxed{\text{ls}}$$

5.4 Linearization in HOG

In this section, we introduce the basics of a framework for handling word order in HOG. In the simple phenogrammar presented in Chapter 2, phenogrammatical objects are simply strings of phonological words and they are always continuous. We extend the objects so that phonology is just one part of a more structured and informative object and that they can be discontinuous.

5.4.1 Basic functions

Before introducing the actual framework, we briefly summarize some of the general-purpose functions we use in this and the following sections. The definitions can be found in Appendix D.

- $\mathsf{filter}(s : \mathsf{Set}(A), \varphi : A \to \mathsf{Bool}) : \mathsf{Set}(A)$ written as $\{s \,|\, \varphi\}$

  A function (written in the usual set-theoretic notation) filtering a set s with a predicate φ. A similar function is used for lists, written as $\langle l|\varphi\rangle$ or $l[\varphi]$.

  For example,

  $\{\,\{1,2,3,4\} \,|\, \lambda x \,.\, x < 3\,\} = \{1,2\}$;

  $\langle 1,2,3,4,1\rangle[\lambda x \,.\, x < 3] = \langle 1,2,1\rangle$

- $\mathsf{set}(l : A^*) : \mathsf{Set}(A)$

  Set corresponding to members of a list. Usually implicit, thus we write e.g., $x \in \mathit{list}$ or $\mathit{set} \subseteq \mathit{list}$.

  For example, $\mathsf{set}(\langle 1,2,3,1,2\rangle) = \{1,2,3\}$

- $\mathsf{list}(s : \mathsf{Set}(A), \rho : \mathsf{Rel}(A)) : A^*$

  List corresponding to a set s ordered by a linear order ρ.

  For example, $\mathsf{list}(\{1,2,3\},\ \lambda a, b \,.\, a > b) = \langle 3,2,1\rangle$

- $\mathsf{orderOf}(l : A^*) : \mathsf{Rel}(A)$ written as $<_l$ operator

  A function returning a linear order corresponding to the ordering within a list. Obviously, this function is undefined for lists with repeating members.

  For example, $5 <_{\langle 7,5,2,1,4\rangle} 1$

- $\mathsf{map}(s : \mathsf{Set}(A), f : A \to B) : \mathsf{Set}(B)$

  A function mapping elements of a set s (and similarly of a list) using a function f.

  For example, $\mathsf{map}(\{1,2,3\}, \lambda x \,.\, 2x) = \{2,4,6\}$; $\mathsf{map}(\{1,2,3\}, \lambda x > 5) = \{\mathsf{false}\}$

- $\mathsf{concatenate}(ls : \mathsf{A}^{**}) : \mathsf{A}^*$

  Takes a list of lists ls and concatenates them all into a single list.

  For example, $\mathsf{concatenate}(\langle\langle 1,2\rangle, \langle\rangle, \langle 3,4\rangle, \langle 4\rangle\rangle) = \langle 1,2,3,4,4\rangle$

5.4.2 Domain Objects

Similarly to the HPSG linearization frameworks discussed in the previous section, we introduce domain objects. Domain objects serve two, closely related purposes. First, they represent potential discontinuity. A single pheno object may be represented in the domain of its mother by more than one object and these domain objects may be separated by domain objects from other pheno objects. Second, they allow to free the ordering constraints from the tecto hierarchy. The ordering constraints order domain objects and not the whole pheno objects.

Therefore we extend pheno objects with a list of domain objects in the following way:

(120)
$$\begin{bmatrix} \mathsf{Pheno} & \\ \mathsf{phon} & \mathsf{Phon} \\ \mathsf{objs} & \begin{bmatrix} \mathsf{Dom} & \\ \mathsf{phon} & \mathsf{Phon} \\ \mathsf{tecto} & \mathsf{Tecto} \end{bmatrix}^* \end{bmatrix}$$

The primitive type Dom is the type of domain objects. There is also a function objs giving a list of domain objects for every pheno object, i.e., a function having the type Pheno as its domain and the type Dom^*, i.e., list of domain objects, as its range.

(121) $\mathsf{objs} : \mathsf{Pheno} \rightarrow \mathsf{Dom}^*$

In addition, there are two functions defined on domain objects:

(122) $\mathsf{phon} : \mathsf{Dom} \rightarrow \mathsf{Phon}$

$\mathsf{tecto} : \mathsf{Dom} \rightarrow \mathsf{Tecto}$

The function phon is the phonology corresponding to the domain object, and tecto gives the corresponding tecto, roughly the syntactic part of HPSG's SYNSEM.

phon. Similarly to the HPSG framework, the phonology of the whole pheno object is simply the concatenation of phonologies of the domain objects. This is ensured by the following constraint, which is an analogue of Reape's Constituent Ordering Principle (Reape 1996, p. 225):[80]

(124) $\vdash_{\textsc{pheno}} \mathsf{phon} = \mathsf{join}(\mathsf{objs}, \mathsf{phon})$

where the function

(125) $\mathsf{join}(src : \mathsf{A}^*, f : \mathsf{A} \to \mathsf{B}) : \mathsf{A}^* := \mathsf{concatenate}(\mathsf{map}(src, f))$

is equivalent to (Kathol and Pollard 1995)'s $\mathsf{join}_f(list)$ in (116).

tecto. It is desirable to allow linear precedence and compaction constraints to refer not only to phenogrammatical properties of domain objects (segmental phonology in our setup, prosody, etc. in a more complex grammar) but also to properties of tecto objects like their syntactic category, case, etc. For example, expressing a constraint that auxiliary clitics precede reflexive clitics cannot in general be done purely in terms of phonology because many of the clitics are homophonous with non-clitics. In theory, it would be possible to introduce phenogrammatical categories that would capture just the information needed for ordering the domain objects, for example by using topological fields as in (Kathol 1995). This might be better in some cases, however using it exclusively would result in an unnecessary duplication of information without providing any clear benefit. Therefore, we assume that every domain object contains information about the corresponding tecto term.

5.4.3 Managing discontinuity

As mentioned above, Reape (1994) distinguishes two possibilities: a domain list is inserted into mother's domain list as a single unbreakable unit (the objects are compacted) or as individual items (domain union, the objects are liberated). Compaction ensures two things: (i) the compacted list of domain objects is continuous; (ii) constraint locality – constraints that apply higher in the derivation tree cannot refer to the individual members of the compacted list because they are inaccessible. Kathol (1995) extends this by allowing the compaction to be partial.

[80] The PHENO subscript on ⊢ means that the constraint on the type Sign is stated only in terms of its pheno. Thus it is a shorthand for the following constraint:

(123) $\vdash_{\mathsf{Sign}} \textsc{pheno}.\mathsf{phon} = \mathsf{join}(\textsc{pheno}.\mathsf{objs}, \mathsf{phon})$

We further generalize this setup for two main reasons. First, we need a way to allow several partial compactions. For example, consider multiple long-fronting from §3.4.4. Both the long-fronted expression and the rest of the clause must be compacted, but they must stay separate. Second, it is desirable to allow compaction to be the result of several independent constraints. For example, a noun may be compacted with its article for phonological reasons and with a post-modifier for syntactic reasons, the result is however a single object. (Penn 1999*a*) argues that compaction in Serbo-Croatian is best stated in terms of several constraints from different grammar levels. This is also desirable for more general reasons of modularity of grammar description.

5.4.3.1 corr function

Because of these requirements, we express the compaction in terms of a function corr:

(126) $\mathsf{corr} : \mathsf{Sign}^3 \times \mathsf{Dom} \rightarrow \mathsf{Dom}$,

It is a function relating domain objects of daughters with domain objects of the mother. Two domain objects are considered compacted if the function assigns them a single domain object. On liberated objects the function is just identity. In the following, the mother-head-argument triple, i.e., the sign combination that this corr is relative to, is left implicit.

The important difference from the approach in (Kathol and Pollard 1995) is that we can constrain two objects a and b to compact ($\mathsf{corr}(a) = \mathsf{corr}(b)$), without preventing another object c to compact with them as well ($\mathsf{corr}(a) = \mathsf{corr}(c)$). Thus a single domain object to be result of several independent compacting constraints. Also there is no restriction into how many objects a domain compacts. (Kathol and Pollard 1995) splits a daugter's domain into two sets (each may be empty), one compacts and the other is liberated. In these respects, our setup provides similar flexibility as Penn's (Penn 1999*a*,*b*) linearization framework. However, Penn's setup does not enforce constraint locality. All domain objects, including their phonologies and topological fields, are accessible at all levels of the syntactic hierarchy, whether compacted or not. Therefore there is no way to disallow constraints that would impose constraints on arbitrary embedded domain object. In our case, the corr function is parameterized by the three sings participating in the combination and therefore such problem does not arise. One could mimic Penn's setup by using a single global corr function, which would then form a tree over all the domain objects.

Sets and lists. We can generalize the function to sets and list in an obvious way:

(127) $\mathsf{corr}(dos : \mathsf{Set}(\mathsf{Dom})) : \mathsf{Set}(\mathsf{Dom}) := \mathsf{map}(dos, \mathsf{corr})$

in other words:

(128) $\mathsf{corr}(set) = \{\mathsf{corr}(do) \mid do \in set\}$

Finally, the constraints and functions below make use of the inverse of the corr function, which we name src. For a domain object of the mother, it returns the list of domain objects of the daughters that compacted into that object.

5.4.3.2 Framework constraints

In addition to the language specific constraints, discussed in the next section, there are language independent constraints capturing the basic properties of compaction. Below, we ensure that (i) domains respect the order of their sub-domains, (ii) compactions are truly continuous, (iii) compaction does not go across signs and (iv) define the phonology of the resulting object of a compaction in terms of the phonology of the sources.

Shuffle. We need to ensure that domains agree on order on the domain objects. This means domain objects from the daughters must be shuffled into domain objects of their mother:

(129) (Shuffle Constraint)
$\vdash_{\mathsf{app}} \forall d \in \mathsf{dtrs}\ \forall x, y \in d.\mathsf{objs}\,.\, x \prec_d y \Rightarrow \mathsf{corr}(x) \preceq_{\mathsf{m}} \mathsf{corr}(y)$

The order of domain objects of each daughter has to be preserved ($\prec$) or the objects are compacted ($=$). $\prec_p$ expresses the order in p's domain and is defined in §5.4.4 below. Note that preserving order does not imply preserving adjacency: adjacency for x and y does not imply adjacency for $\mathsf{corr}(x)$ and $\mathsf{corr}(y)$.

Continuous Compaction. If two objects compact, all objects between them must compact with them as well:

(130) (Continuous Compaction)
$\vdash_{\mathsf{app}} \forall d \in \mathsf{dtrs}\ \forall i, j\,.\, \mathsf{corr}(\mathsf{objs}[i]) = \mathsf{corr}(\mathsf{objs}[j]) \Rightarrow$
$\forall i < k < j\,.\, \mathsf{corr}(\mathsf{objs}[i]) = \mathsf{corr}(\mathsf{objs}[k])$

This constraint is analogous to the Planarity Constraint in (Penn 1999*a*). As discussed in §5.3.2 above, (Kathol and Pollard 1995)'s partial compaction does not require this constraint to hold.

Compaction does not cross daughters. Because only domain objects, and not phenos, are ordered, we need to restrict compaction to objects from a single domain:

(131) $\vdash_{\mathsf{app}} \forall o \in \mathsf{m}.\textsc{pheno}.\mathsf{objs}\ \exists d \in \mathsf{dtrs}\,.\,\mathsf{src}(o) \subseteq d.\mathsf{objs}$

Phonology. This means we can define that the phonology of the result of a compaction is equal to the joined phonologies of the source domain objects:

(132) $\vdash_{\mathsf{app}} \forall do \in \mathsf{m}.\textsc{pheno}.\mathsf{objs}\,.\, do.\mathsf{phon} = \mathsf{join}(\mathsf{src}(o), \mathsf{phon})$

This is trivially true for a liberated domain object.

5.4.3.3 Liberation, Insertion

We can now define functions liberalizing or compacting a set of domain objects into mother's domain.

(133) $\mathsf{liberate}_a(dos : \mathsf{Set}(\mathsf{Dom})) : \mathsf{Bool} := \forall do \in dos\,.\,\mathsf{corr}(do) = do$

The predicate $\mathsf{liberate}_a$ ensures that the domain objects dos are liberated in mother's domain, i.e., they are inserted individually, and we are thus extending mother's domain to the portion of daughter's domain containing these objects (the objects may, but need not originate in a single sign).

(134) $\mathsf{compact}_a(dos : \mathsf{Set}(\mathsf{Dom}), o : \mathsf{Dom}) : \mathsf{Bool} := \forall do \in dos\,.\,\mathsf{corr}(do) = o$

The predicate $\mathsf{compact}_a$ ensures that the domain objects dos are compacted in mother's domain as the domain object o. Note that the set dos does not have to be exhaustive. All objects in dos compact into o, but there can be additional objects that compact into o as well. This means that a single compaction may be result of several independent constraints using this function. For example, part of it may be conditioned prosodically and part syntactically.

We can also define a more restricted version of this function, where the set of compacted objects is exhaustive:

(135) $\mathsf{insert}_a(dos : \mathsf{Set}(\mathsf{Dom}), o : \mathsf{Dom}) : \mathsf{Bool} := \mathsf{src}(o) = dos$

Because in some constraints, it is only important whether a set of domain objects is compacted or not without a need to refer to the corresponding domain objects in mother's domain, we define the following predicates:

(136) $\mathsf{compact}_a(dos : \mathsf{Set(Dom)}) : \mathsf{Bool} := \exists o : \mathsf{Dom} \,.\, \mathsf{compact}_a(dos, o)$

(137) $\mathsf{insert}_a(dos : \mathsf{Set(Dom)}) : \mathsf{Bool} := \exists o : \mathsf{Dom} \,.\, \mathsf{insert}_a(dos, o)$

The predicates are defined in terms of sets of domain objects. We can define their variants for whole signs, i.e., for all domain objects of a sign, in an obvious way. Then $\mathsf{insert}_a(s : \mathsf{Sign}) : \mathsf{Bool}$ is the exact equivalent of Reape's domain-insertion operation (full compaction) and $\mathsf{liberate}_a(s : \mathsf{Sign}) : \mathsf{Bool}$ of his domain union.

5.4.3.4 Partial compaction

Using the above relations, we can define a relation equivalent to (Kathol and Pollard 1995)'s partial compaction discussed in §5.3.2:

(138) $\mathsf{p\text{-}compact}_a(s : \mathsf{Sign}, c : \mathsf{Dom}, ls : \mathsf{Dom}^*) : \mathsf{Bool} :=$
$\mathsf{insert}_a(s.\textsc{pheno}.\mathsf{objs} - ls, c)\ \&$
$c.\textsc{tecto} = s.\textsc{tecto}\ \&$
$\mathsf{liberate}_a(ls)$

It is a relation between a sign s, a domain object c that corresponds to the sing's compacted portion and a list of the remaining objects (cs) that are liberated into the higher domain, i.e., they are inserted individually. The first line:

(139) $\mathsf{insert}_a(s.\textsc{pheno}.\mathsf{objs} - ls)$

compacts the non-liberated domain object into mother's domain as c. Note that if we indeed wanted to replicate (Kathol and Pollard 1995)'s partial compaction including possibility of compaction of non-continuous subdomains, we would need to drop the Continuous Compaction constraint in (130). The next line

(140) $c.\textsc{tecto} = s.\textsc{tecto}$

simply ensures that the compacted portion has the same tecto as the whole sign s.[81] The last line then liberates the remaining objects of s into the domain of s's mother:

[81] Here we follow Kathol's analysis, such approach is not without problems, especially when the compacted portion does not contain the head of the whole phrase.

(141) $\mathsf{liberate}_a(ls)$

A more specialized version, Yatabe (1996)'s prefix compaction (suffix compaction is analogous), can be defined as:

(142) $\mathsf{prefix\text{-}compact}_a(s : \mathsf{Sign}, c : \mathsf{Dom}, ls : \mathsf{Dom}^*) : \mathsf{Bool} :=$
$\qquad \mathsf{p\text{-}compact}_a(s, c, ls)\ \&\ \mathsf{suffix}(s.\textsc{pheno}.\mathsf{objs}, ls)$

It requires that the compacted objects form the prefix of the daughter's domain (or the liberated form a suffix). The predicate suffix can be defined in terms of concatenation of two list, when the later is the suffix:

(143) $\mathsf{suffix}(list : \mathsf{A}^*, suff : \mathsf{A}^*) : \mathsf{Bool} := \exists pref : \mathsf{A}^* \, . \, list = pref \circ suff$

5.4.4 Ordering

For each pheno object p, we can define a linear order abstracted from the list of domain objects:

(144) $\prec \ : \mathsf{Pheno} \to \mathsf{Rel}(\mathsf{Dom})$
$\prec \ := \lambda p : \mathsf{Pheno} \, . \, \mathsf{orderOf}(p.\mathsf{objs}) \qquad (= <_{p.\mathsf{objs}})$

We write $a \prec_p b$ for $\prec(p)(a, b)$. However, in constraints on a particular pheno object or sign, we omit the subscript. The ordering constraints can be stated in terms of this order.

The order can be generalized to sets, with the meaning that it holds for all members of that set (we do not distinguish between the two orders notationally, since they are uniquely determined by the type of arguments):

(145) $\prec : \mathsf{Pheno} \to \mathsf{Rel}(\mathsf{Set}(\mathsf{Dom}))$
$\vdash \forall a, b : \mathsf{Set}(\mathsf{Dom}) \, . \, a \prec b \Leftrightarrow (\forall x \in a, y \in b \, . \, x \prec y)$

5.4.5 Meadows

Many of the constraints in the following sections are expressed over particular collections of expressions. Meadows, as we call these collections, might be viewed as similar to topological fields used traditionally in description of German syntax (see for example, Drach 1937; Erdmann 1886; Herling 1821; Höhle 1986). However, there are many differences:

1. A single object might be assigned to several meadows. It means we can classify domain objects according to various criteria. For example, we partition domain of a finite clause relative to information structure into a part of fronted objects, a part ordered by IS, and a part inert to IS. We also partition it relative to clitics into the first-position, the second-position and the rest. There might be constraints relating such partitioning, e.g., relating the first-position and the fronted expressions.

2. A meadow need not be in general continuous, for example fronted expressions can be split by the second-position clitics.

3. There can be a hierarchy of meadows, for example the second-position clitics are further divided into auxiliary clitics, complement dative clitics, complement adjunct clitics, etc. The meadows are then ordered relative to each other within the second-position. This is similar to Penn's topological tree used for constraining Serbo-Croatian clitics (Penn 1999*a*).

We may model this in two ways, either generalize the HPSG's approach to fields which are labels assigned to individual domain objects. Instead of a single label, we would assign a set of them. The other possibility is to model meadows as predicates on the type Dom. We choose the latter possibility, but nothing hinges on that choice.

(146) $\mathsf{Meadow} : \mathsf{Dom} \rightarrow \mathsf{Bool}$

If p is a pheno object and m is a meadow, we can refer to the objects of the pheno p belonging to the meadow m by filtering the list of its domain objects by the meadows as a predicate:

(147) $p.\mathsf{objs}[m]$ (equivalent to $\mathsf{filter}(p.\mathsf{objs}, m)$)

We define the following notation for expressing ordering constraints in terms of the meadows

(148) a. Linear order restricted to a meadow:

$\prec_{p,m}$ stands for $\prec_{p.\mathsf{objs}[m]}$.

b. Order of two meadows:

$m_1 \prec_p m_2$ stands for $\mathsf{objs}[m_1] \prec_p \mathsf{objs}[m_2]$

In both cases, p : Pheno and is usually left implicit. Moreover, relation between meadows may be expressed as any relations between sets and lists, such as subset, partitions, etc. Finally, we often need to specify that a certain meadow is appropriate or required for phenos of certain signs.[82]

- appropriateness: the meadow can be non-empty only for a pheno associated with a particular tecto:

  $\mathsf{appropriate}(mead : \mathsf{Meadow}, \mathsf{T} \sqsubset \mathsf{Tecto}) : \mathsf{Bool}$

  $\forall s : \mathsf{Sign} \,.\, s.\textsc{pheno}.\mathsf{objs}[mead] \neq \emptyset \Rightarrow s.\textsc{tecto} :: \mathsf{T}$

- requirement: every pheno associated with a particular tecto has the field nonempty:

  $\mathsf{required}(mead : \mathsf{Meadow}, \mathsf{T} \sqsubseteq \mathsf{Tecto}) : \mathsf{Bool}$

  $\forall s : \mathsf{Sign}(\mathsf{T}) \,.\, s.\textsc{pheno}.\mathsf{objs}[mead] \neq \emptyset$

5.5 Czech word order in HOG

In this section, we specify the relationship between the simple tectogrammar from §5.1 and a phenogrammar defined along the lines of the previous section.

5.5.1 Information Structure in tecto

As discussed in Chapter 3, the major element in determining word order in a Czech sentence is its Information Structure (IS). In our opinion, the most appropriate way to model Information Structure in HOG would be to treat it as a separate component of signs, parallel to phenogrammar, tectogrammar and semantics. However, here we assume that IS is simply accessible from tectogrammar: there is a function defined on all tecto terms that returns the IS associated with them. This is equivalent to saying that there are syntactic entities or features corresponding to IS.

In §3.3, we concluded that the following three distinctions must be made:

1. theme – rheme
2. contrast – background. In Czech, distinguished only in theme, in some languages (Catalan, Finish) in both theme and rheme.

[82] We assume every field is defined for every pheno object but for some of them it is necessarily empty. It would be possible to define fields only for particular pheno objects, but the typing would be rather complex.

3. proper – not-proper: Theme proper is the most thematic, most salient part of the theme. Rheme proper is the most rhematic part of the rheme.

This information may be captured by a triple of boolean values:

(149) $$\mathsf{IS} := \begin{bmatrix} \textsc{rheme} & \mathsf{Bool} \\ \textsc{proper} & \mathsf{Bool} \\ \textsc{contrast} & \mathsf{Bool} \end{bmatrix}$$

There are eight terms of this record type (records are extensional, i.e., records with equal components are equal, §C.2(2b)). If we wanted to limit the possibility of contrast only to theme, we would define the type IS as a subtype of the tuple:

(150) $$\mathsf{IS} := [\, x : \begin{bmatrix} \textsc{rheme} & \mathsf{Bool} \\ \textsc{proper} & \mathsf{Bool} \\ \textsc{contrast} & \mathsf{Bool} \end{bmatrix} \mid \textsc{contrast} \Rightarrow \textsc{theme}\,]$$

Now, we can define a function assigning the IS terms to every tecto expression:

(151) $\mathsf{is} : \mathsf{Tecto} \rightarrow \mathsf{IS}$

We can define the following convenience predicates, indicating whether an expression is thematic (i.e., not rhematic) or in the background (i.e., not contrasted):

(152) $\mathsf{theme} := \textsc{rheme}.\mathsf{neg}$

$\mathsf{bckg} := \textsc{contr}.\mathsf{neg}$

Technically, the function RHEME is a projection (record attribute), while the function theme is a regular function. However, we do not see any reason to consider one basic and the other derived; hence we slightly abuse the notation and write them both as function: rheme and not RHEME; and similarly for the other two projections.

We *define* the IS of a constituent to be the same as the IS of its head:

(153) (passIS)

$\vdash_{\mathsf{app}} \mathsf{m}.\textsc{tecto}.\mathsf{is} = \mathsf{h}.\textsc{tecto}.\mathsf{is}$

Note, that this does not imply that expressions with heterogenous IS (e.g., in an NP, the adjective is in rheme, while the head noun is in theme) must be continuous. The parts with different IS can be represented via separate domain objects. The tecto expressions corresponding to each domain object has then different IS.

5.5.2 Information Packaging

5.5.2.1 Basic setup

To capture the view of Czech Information Packaging presented in §3.3.5, we introduce the following three meadows partitioning finite clauses :

(154) a. front – fronted domain objects (both short and long fronted, sFront, lFront). In objective ordering, the objects must be in theme proper, in subjective ordering in rheme proper.

b. inert – expressions whose ordering is not determined directly by Information Structure (this includes clitics which are handled in §5.6)

c. isRest – the rest of the clause, where the word order (roughly) corresponds to increasing communicative dynamism, which is the following order on IS:

(155) Theme Proper < other Theme < other Rheme < Rheme Proper

None of these meadows are necessarily continuous. For example, fronted elements may be split by clitics, long fronted elements may climb to higher clauses, and clitics split isRest in a fully rhematic sentence.

5.5.2.2 Ordering

There are two ordering constraints that we can state at this level of detail. First, fronted expressions must be indeed fronted:

(156) $\vdash_{\text{PHENO}} \mathsf{front} \prec \mathsf{isRest}$

Second, the isRest is ordered according to the order in (155). To achieve this we define a partial order on the 8 objects of the type IS corresponding to (155):

(157) $<_{IS}$: $\mathsf{Rel(IS)}$

and require that the linear order of domain objects in isRest respects it:

(158) $\vdash_{\text{PHENO}} \mathsf{respects}(\prec_{\mathsf{isRest}},\ \mathsf{getOrder}(<_{IS}, \text{TECTO}.\mathsf{is}))$

where the function

(159) $\mathsf{getOrder}(\rho : \mathsf{Rel}(B), f : A \to B) : \mathsf{Rel}(A)$

returns a partial order on domain objects according to their information structure. And the predicate

(160) $\mathsf{respects}(\rho : \mathsf{Rel}(B), \sigma : \mathsf{Rel}(B)) : \mathsf{Bool}$

ensures that the linear order on the domain objects respects such a relation. It is also possible to state the constraint in (158) in the following form:

(161) $\vdash_{\mathsf{Sign}(\mathsf{S}_{\mathsf{fin}}).\textsc{pheno}} \forall a, b \in \mathsf{objs}[\mathsf{isRest}] \,.\, a.\textsc{tecto}.\mathsf{is} <_{IS} b.\textsc{tecto}.\mathsf{is} \Rightarrow a \prec b$

5.5.2.3 Fronting

Theme Proper or Rheme Proper. As we discussed in §3.3, which expression is fronted depends on which ordering the speaker uses. In the so-called objective ordering it is the theme proper and and in the so-called subjective ordering, it is the rheme proper. Objective ordering can be considered the default choice, while subjective ordering is used in certain specific contexts, especially in emphatic or excited speech. We assume that there is an external parameter determining whether objective or subjective ordering is to be used:

(162) $\mathsf{objectiveIS} : \mathsf{Bool}$

Depending on the value of the parameter, either theme proper or rheme proper may be fronted:

(163) $\mathsf{frontable}(do : \mathsf{Dom}) : \mathsf{Bool} := \mathsf{is}.(\mathsf{proper} \,\&\, \varphi)$
where $\varphi = \mathsf{if\ objectiveIS\ then\ theme\ else\ rheme}$

Frontable objects that are in a finite clause domain (either because they are clausal or because they climbed from more embedded phrases) are fronted:

(164) $\vdash_{\mathsf{Sign}(\mathsf{S}_{\mathsf{fin}}).\textsc{pheno}} \forall do \in \mathsf{objs} \,.\, \mathsf{front}(do) \Leftrightarrow \mathsf{frontable}(do)$

One consequence of this constraint is that only sentences with all clausal constituents rhematic and only in objective ordering have no fronted expressions. A constituent can be split-fronted if it is a clausal constituent (of both finite or infinitive clauses):

(165) $\vdash_{\mathsf{app}} \mathsf{Tecto}, \mathsf{S}_{\mathsf{fin} \vee \mathsf{inf}} : \quad \forall d \in \mathsf{dtrs} \,.\, \mathsf{insert}_a(d.\textsc{pheno}.\mathsf{objs}[\neg(\mathsf{front} \mathbin{\&} \mathsf{climbCs})])$

The constraint inserts the part of the constituent that is neither fronted nor it is climbing clitics (see below). We leave the nature of the fronted part in the higher domain unspecified. Independent constraints need to ensure that it is either inserted as a single object or liberated as several independent objects to capture difference between sentences in (3.4.4.2 27), repeated below:

(166) a. [Hlídat$_2$ děti Novákům] *si*$_1$ teda netroufnu$_1$.
watch$_{\mathsf{inf}}$ children Nováks$_D$ refl$_D$ so not-dare
'I DO NOT DARE$_R$ to babysit for the Nováks$_C$.'

b. [Hlídat děti | Novákům] *si*$_1$ teda netroufnu$_1$.
watch$_{\mathsf{inf}}$ children Nováks$_D$ refl$_D$ so not-dare
'I DO NOT DARE$_R$ to babysit$_C$ for the Nováks$_C$.'

c. [[Hlídat děti] a [Novákům]] *si*$_1$ teda netroufnu$_1$.
watch$_{\mathsf{inf}}$ children and Nováks$_D$ refl$_D$ so not-dare
'I DO NOT DARE$_R$ to babysit$_C$ for the Nováks$_C$.'

From the point of fronting, NPs in PPs are behaving in the same way as clausal NPs, therefore we liberate the NP's domain into the domain of the PP, which means all the NP's domain objects are also PP's domain objects, and thus are accessible to the clausal domain:

(167) $\vdash_{\mathsf{app}} \mathsf{m}.\textsc{tecto} :: \mathsf{PP} \Rightarrow \mathsf{liberated}_a(\mathsf{nhDtrs})$

Prepositions. However, the preposition cannot be split-fronted alone. It always forms a single unit with *at least* the first member of the NP's domain list, whether it is an attribute of the noun or the noun itself:

(168) $\vdash_{\mathsf{app}} \forall pp : \mathsf{Sign}(\mathsf{PP}) \,.\, pp \in \mathsf{dtrs} \Rightarrow$
$\mathsf{compacted}_a(\{pp.\textsc{pheno}.\mathsf{objs}[1], pp.\textsc{pheno}.\mathsf{objs}[2]\})$

Note that nothing prevents other constraints from requiring other parts of the PP to compact into the same domain object as the preposition.

Multiple fronting Multiple fronting is possible, but all fronted expressions must be either contrasted, or be spatio-temporal/path/period adverbials (we assume there are predicates determining it, we would need semantics):

(169) $\vdash_{\mathsf{Sign}(\mathsf{S}_{\mathsf{fin}}).\textsc{pheno}}$ $\mathsf{card}(\mathit{fronted}) > 1 \Rightarrow (\forall do \in \mathit{fronted}\,.\, do.\textsc{tecto}.\mathsf{is.contr}) \vee$
$\mathsf{spatio\text{-}temporal}(\mathit{fronted}) \vee \mathsf{pathPeriod}(\mathit{fronted})$
$\mathsf{where}\ \mathit{fronted} = \mathsf{objs}[\mathsf{front}]$

When such multiple expressions are long-fronted, they cannot be split (by clitics), i.e., lFront is inserted into the higher domain where it must be fronted.

(170) $\vdash_{\mathsf{app}} \forall s : \mathsf{Sign}(\mathsf{S}_{\mathsf{fin}})\,.\, s \in \mathsf{dtrs} \Rightarrow$
$\exists o : \mathsf{Dom}\,.\,\mathsf{inserted}_a(s.\textsc{pheno}.\mathsf{objs}[\mathsf{lFront}], o)\ \&\ \mathsf{front}(o)$

5.5.3 Additional constraints

P in PPs. Czech prepositional phrases are without exception head initial. We can enforce it by the following constraint:

(171) $\vdash_{\mathsf{app}} \mathsf{headFirst}_a(\mathsf{PP})$

The polymorphic predicate headFirst$_a$ ensures that the domain objects corresponding to the head of PP are initial in PP's pheno. It accepts a type of the phrase and tests whether the pheno of the head is initial in the pheno of the whole phrase:

(172) $\mathsf{headFirst}_a(\mathsf{T} \sqsubseteq \mathsf{Tecto}) : \mathsf{Bool} :=$
$\forall s : \mathsf{Sign}(\mathsf{T})\,.\,\mathsf{prefixOf}(s.\textsc{pheno}.\mathsf{head}, s.\textsc{pheno}.\mathsf{objs})$

where prefixOf is a simple predicate testing if a list is a prefix of another list. head is the list of domain object(s) corresponding to the head (it is a feature of Pheno, therefore, precisely it is a function from

(173) $\vdash_{\mathsf{app}} \mathsf{m}.\textsc{pheno}.\mathsf{head} := \mathsf{map}(\mathsf{h}.\textsc{pheno}.\mathsf{objs}, \mathsf{corr})$

C in $\bar{\mathsf{S}}$ $\bar{\mathsf{S}}$ clauses are head first in the sense, that only the long fronted-expressions can precede the complementizer:

(174) $\vdash_{\mathsf{Sign}(\bar{\mathsf{S}}).\textsc{pheno}} \mathsf{prefixOf}(\mathsf{objs}[\mathsf{lFront}] \circ \langle \mathsf{head} \rangle, \mathsf{objs})$

5.6 Clitics

5.6.1 Representation of clitics

We treat clitics as regular words. Tectogrammar specifies the term $\mathsf{clitics}$, the set of tecto terms that correspond to clitics, i.e., to the words enumerated in §4.3.

(175) $\mathsf{clitics} : \mathsf{Set(Tecto)}$

This set is in fact a Tecto-predicate (written as clitic; $\mathsf{clitic} = \mathsf{clitics}$), which can be used in subtyping, for example pronominal clitics have the type:

(176) $\mathsf{PPron}_{\mathsf{clitic}}$

The clitic pronoun *mu* 'him$_{dat}$' is of type

(177) $\mathsf{him}_{dat,c} : \mathsf{PPron}_{\mathsf{3,m,sg,dat,clitic}}$ (*mu*)

and its non-clitic counterpart *jemu* 'him$_{dat}$' is of type

(178) $\mathsf{him}_{dat,nc} : \mathsf{PPron}_{\mathsf{3,m,sg,dat,\neg clitic}}$ (*jemu*)

For convenience, we define analogous function on domain objects:

(179) $\mathsf{clitic} : \mathsf{Dom} \rightarrow \mathsf{Bool} := \textsc{tecto}.\mathsf{clitic}$

Inconstant clitics Inconstant clitics are represented as two different tecto terms. For example *jí* 'her$_D$' would have a distinct tecto term for its clitic (weak) and non-clitic (strong) variant.

(180) $\mathsf{her}_{dat,c} : \mathsf{PPron}_{\mathsf{3,f,sg,dat,clitic}}$ (*jí*)
$\mathsf{her}_{dat,nc} : \mathsf{PPron}_{\mathsf{3,f,sg,dat,\neg clitic}}$ (*jí*)

Information Structure Clitics must be in a non-contrastive theme (except the rare cases of contrasted conditional auxiliary):

(181) $\vdash_{\textsc{tecto}} \mathsf{clitic}\ \&\ (\mathsf{this} \neq \mathsf{would}) \Rightarrow \mathsf{is.theme}\ \&\ \neg\mathsf{is.contr}$

where $\mathsf{would} = \coprod_{g \in \mathsf{Gender}, p \in \mathsf{Person}, n \in \mathsf{Nr}} \mathsf{would}_{g,p,n}$

5.6.2 Climbing

5.6.2.1 Climbing Queues

To express some of the climbing constraints, we need some limited structural information about the climbing clitics. Constraints on monotonicity (§4.6.3.1), GDEC (§4.6.3.3) and potentially also the morphological constraints (haplology and identity, §4.6.1) refer to clitics with more and with less embedded governors. For every two of clitics, we need to know whether they have the same governor or which of them has a more deeply embedded governor. On the other hand, we do not need to know which governor it is, its syntactic categories or even its depth of embedding. To capture the needed nonlocal information but not more, we assume that the collections modeling clitic climbing are partially ordered queues.[83]

(182) $\mathsf{EQueue}(\mathsf{A})$

Two clitics are at the same place in such a queue if they have the same governor and a clitic precedes another clitic when the former is less embedded. Therefore the structure induces two relations: a partial order and an equivalence. We construct the queues in such a way that the following holds:

(183) for q : $\mathsf{EQueue}(\mathsf{Dom})$:
$\ll_q$: $\mathsf{Rel}(\mathsf{Dom})$ – $a \ll_q b$ – a has a less embedded governor than b
$\equiv_q$: $\mathsf{Rel}(\mathsf{Dom})$ – $a \equiv_q b$ – a and b have the same governors the queue

The queue can be viewed as a list of sets. In such view, if two objects are in the same set, $\equiv$ holds between them, if they are in different set, they can be related by $\ll$.

In addition there are the following constructors:

(184) a. $\langle\rangle$: $\mathsf{EQueue}(\mathsf{Dom})$
constructs an empty queue

b. $\langle s\rangle$: $\mathsf{EQueue}(\mathsf{Dom})$ for s : $\mathsf{Set}(\mathsf{Dom})$
constructs a queue where all members have the same governor

c. $a \oplus b$: $\mathsf{EQueue}(\mathsf{Dom})$ for a, b : $\mathsf{EQueue}(\mathsf{Dom})$
joins queues; all members of a have a less embedded governor than those in b

[83] This treatment is more constrained than our analysis in (Hana 2004), where the constraints related the tectogrammatical structure to the whole pheno-structure.

We also use the standard set operations like $\in$, $\subset$, etc.

Now, we can define the following collections. For every domain, we partition all the clitics in that domain in two different ways:

(185) a. By their origin:

i. $\mathsf{locCs} : \mathsf{Pheno} \rightarrow \mathsf{Set}(\mathsf{Dom})$ – clitics originating in this domain. Unlike the other collections, this is a simple set, because all clitics in locCs have the same governor so there is no benefit in making it a queue.

ii. $\mathsf{lowCs} : \mathsf{Pheno} \rightarrow \mathsf{EQueue}(\mathsf{Dom})$ – clitics climbing from an embedded domain

b. By their destination:

i. $\mathsf{stayCs} : \mathsf{Pheno} \rightarrow \mathsf{EQueue}(\mathsf{Dom})$ – clitics staying in this domain

ii. $\mathsf{climbCs} : \mathsf{Pheno} \rightarrow \mathsf{EQueue}(\mathsf{Dom})$ – clitics climbing to a higher domain.

5.6.2.2 Origin fields

locCs contains all the local clitics, i.e., domain objects of all clitic daughters:

(186) (Local Clitics)

$\vdash_{\mathsf{app}} \textsc{pheno}.\mathsf{locCs} = \mathsf{map}(\mathsf{filter}(\mathsf{dtrs}, \textsc{tecto}.\mathsf{clitic}), \textsc{pheno}.\mathsf{objs})$

There can be maximally one source of climbing, i.e., clitics cannot climb from two daughters at the same time.[84] Mother's lowCs is then identical to that daughter's climbCs:

(187) $\vdash_{\mathsf{app}} \mathsf{card}(\mathit{climbCss}) \leq 1 \,\&$

$\mathsf{m}.\textsc{pheno}.\mathsf{lowCs} = \mathsf{if}\ \mathit{climbCss} = \emptyset\ \mathsf{then}\ \langle\rangle\ \mathsf{else}\ \mathit{climbCss}.\mathsf{sing}$

$\mathsf{where}\ \mathit{climbCss} = \{f \in \mathsf{map}(\mathsf{dtrs}, \textsc{pheno}.\mathsf{climbCs}) \mid f \neq \langle\rangle\}$

5.6.2.3 Basic climbing constraints

Clitics can climb only from phrases headed by infinitives and predicate adjectives or quantified phrases. Similarly as (Pollard 2006, (191)), we assume that adjectival predicates are small clauses (type $\mathsf{S}_{\mathsf{adj}}$). This means climbCs is appropriate (i.e., can be nonempty) only for these three phrases:

(188) $\vdash_{\mathsf{Sign}} \mathsf{appropriate}(\mathsf{climbCs}, \mathsf{S}_{\mathsf{inf}} + \mathsf{S}_{\mathsf{adj}} + \mathsf{QP})$

[84] We are ignoring the rare cases of genitives climbing at the same time from both the subject and an complement.

Clitics can climb only through infinitival clauses:

(189) $\vdash_{\mathsf{Sign}}$ $\textsc{pheno}.\mathsf{lowCs} \neq \emptyset$ & $\textsc{pheno}.\mathsf{climbCs} \neq \emptyset \Rightarrow \textsc{tecto} :: \mathsf{S}_{\mathsf{inf}}$

A single domain can contain maximally one cluster and the cluster must compact with the head of the phrase:

(190) $\vdash_{\textsc{pheno}}$ $\mathsf{compact}(\mathsf{stayCs} \cup \mathsf{head})$

This implies for example, that when an infinitival phrase is partially split, the clitics must appear with the infinitive.

5.6.2.4 Monotonicity

The following constraint requiring climbing to be monotonic was stated in §4.6.3.1:

(191) A clitic can climb to a particular cluster only if also all clitics with a less embedded governor climbed to that or a higher cluster.

This constraint can be enforced simply by 'splitting' the complete queue of clitics (note that any of the sub-queues can be empty):

(192) (Monotonicity)
$\vdash_{\textsc{pheno}}$ $\langle\mathsf{locCs}\rangle \oplus \mathsf{lowCs} = \mathsf{climbCs} \oplus \mathsf{stayCs}$

In the $\langle\mathsf{locCs}\rangle \oplus \mathsf{lowCs}$ queue, the local clitics ($\langle\mathsf{locCs}\rangle$) stand before the clitics climbing from an embedded domain (lowCs). This queue is then split and the first part ($\mathsf{climbCs}$) climbs up and the lower (stayCs) stays. Note that this implies that if there is any climbing, the local clitics must climb.

For example, if A are local clitics, and B and C are clitics of more embedded governors, then

(193) $\langle\mathsf{locCs}\rangle \oplus \mathsf{lowCs} = \langle A\rangle \oplus \langle B, C\rangle = \langle A, B, C\rangle$

Then there are the following possibilities:

(194) a. $\mathsf{climbCs} = \langle A, B, C\rangle, \mathsf{stayCs} = \langle\rangle$ – All clitics climb

b. $\mathsf{climbCs} = \langle A, B\rangle, \mathsf{stayCs} = \langle C\rangle$ – Local and the least embedded clitics climb

c. $\mathsf{climbCs} = \langle A\rangle, \mathsf{stayCs} = \langle B, C\rangle$ – Local clitics climb

d. $\mathsf{climbCs} = \langle\rangle, \mathsf{stayCs} = \langle A, B, C\rangle$ – No climbing

Or for a real sentence:

1. All clitics climb:

climbCs	
⟨locCs⟩	lowCs

This is the case in the following example, with derivation in Figure 5.3. In the domain of the verb *pomoct* ‘help$_{\mathsf{inf}}$ both the local clitic *mu* ‘him$_D$’ and the clitic of the lower verb *ho* ‘him$_A$’.

(195) Všichni *jsme*$_0$ *se*$_1$ *mu*$_2$ *ho*$_3$ snažili$_1$ pomoci$_2$ najít$_3$.
all aux$_{1pl}$ refl$_A$ him$_D$ him$_A$ tried help$_{\mathsf{inf}}$ find$_{\mathsf{inf}}$
‘All of us tried to help him to find it.’

2. All clitics stay:

stayCs	
⟨locCs⟩	lowCs

On the other hand, in the following sentence, with a derivation in Figure 5.4, both clitics stay in the domain of the verb *pomoct* (therefore none of the clitics in the main clitic cluster climbed there).

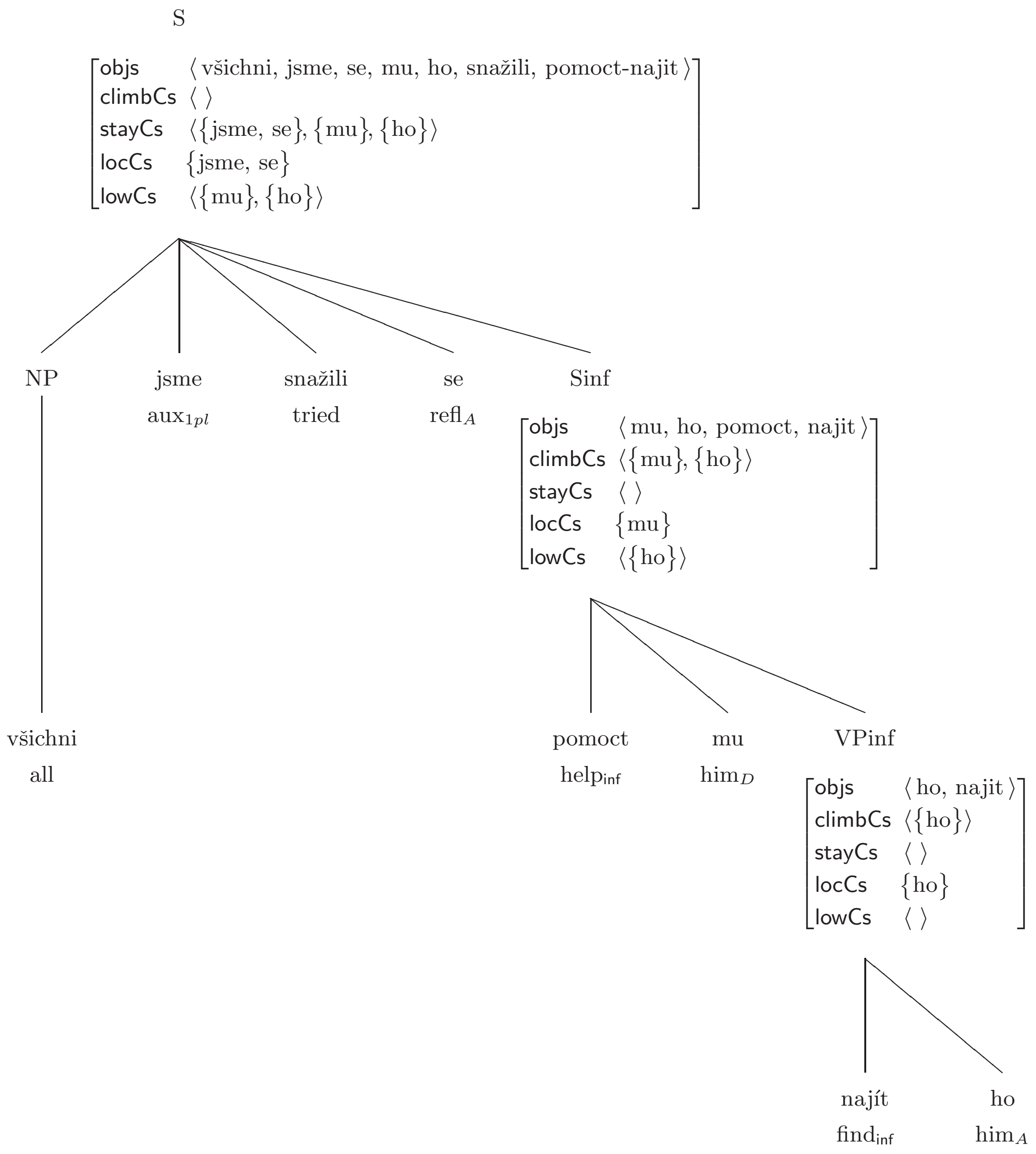
(196) Všichni *jsme*$_0$ *se*$_1$ snažili$_1$ *mu*$_2$ *ho*$_3$ pomoct$_2$ najít$_3$.
all aux$_{1pl}$ refl$_A$ tried him$_D$ him$_A$ help$_{\mathsf{inf}}$ find$_{\mathsf{inf}}$
‘All of us tried to help him to find it.’

3. Some clitics (less embedded) climb, some clitics (more embedded) stay:

climbCs		stayCs
⟨locCs⟩	lowCs	

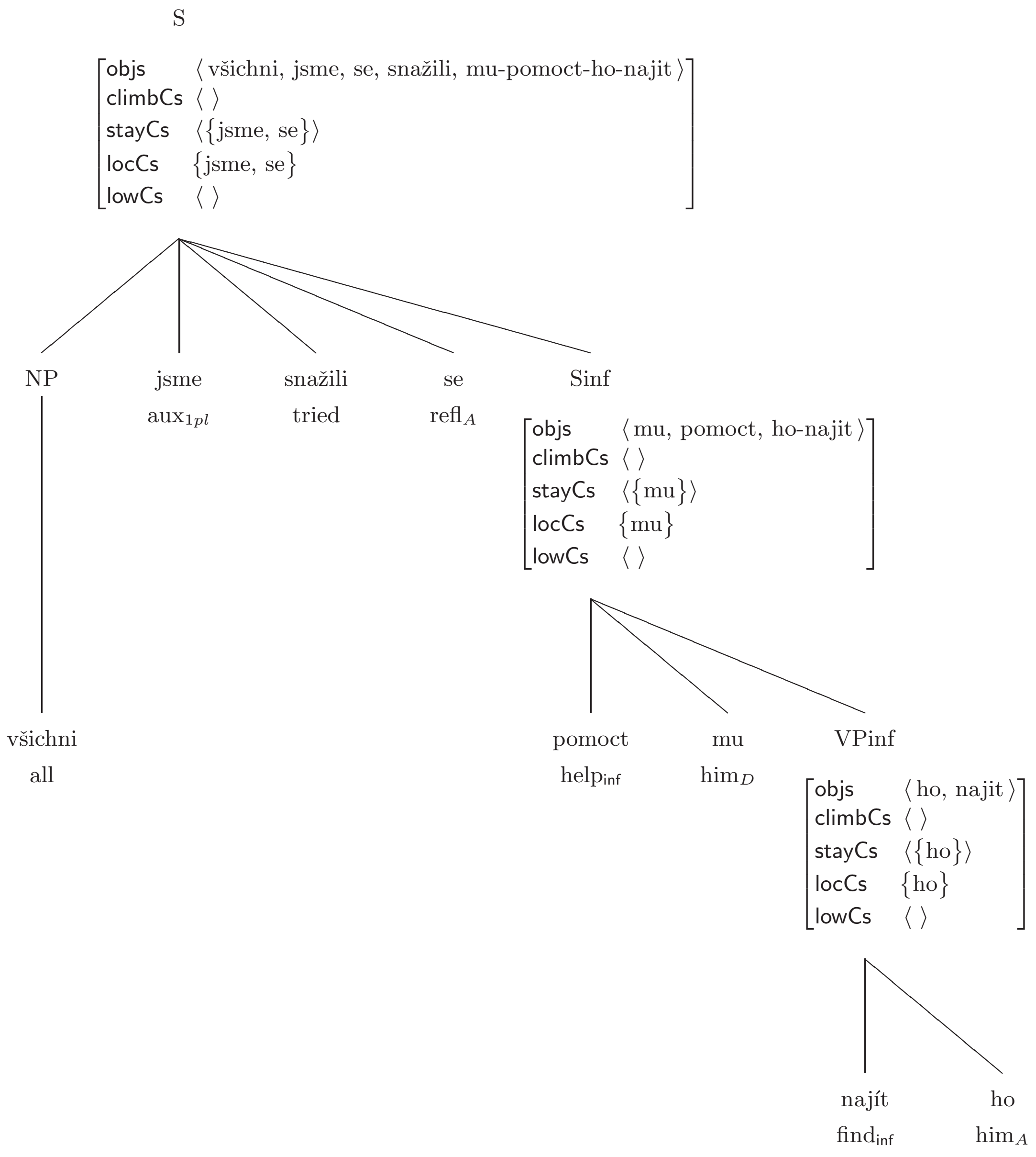
Finally, in the following sentence, with derivation in Figure 5.4, the local clitic of the verb *pomoct* ‘help$_{\mathsf{inf}}$’ climbs up, while the clitic of the lower verb stays.

(197) Všichni *jsme*$_0$ *se*$_1$ *mu*$_2$ snažili$_1$ *ho*$_3$ pomoct$_2$ najít$_3$.
all aux$_{1pl}$ refl$_A$ him$_D$ tried him$_A$ help$_{\mathsf{inf}}$ find$_{\mathsf{inf}}$
‘All of us tried to help him to find it.’



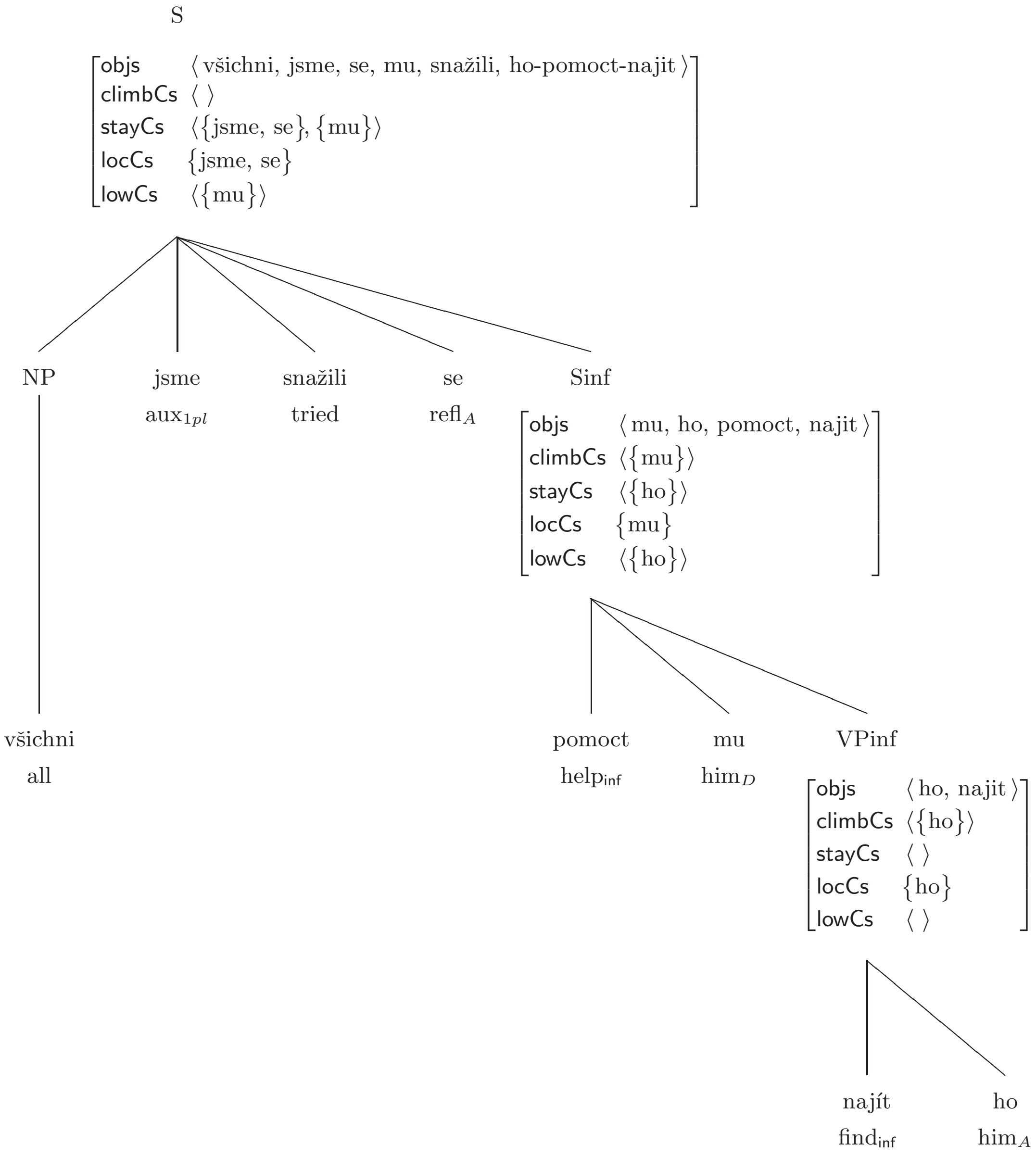
‘All of us tried to help him to find it.’

Figure 5.3: Monotonic climbing: All climb



‘All of us tried to help him to find it.’

Figure 5.4: Monotonic climbing: None climbs



‘All of us tried to help him to find it.’

Figure 5.5: Monotonic climbing: Some climb

However, there is no way how to derive the incorrect case when the local clitic stays while the lower clitics continue climbing.

(198)

* Všichni	jsme_0	se_1	ho_3	snažili_1	[mu_2	pomoci_2	najít_3.]
all	aux_{1pl}	refl_A	him_A	tried	him_D	$\text{help}_{\mathsf{inf}}$	$\text{find}_{\mathsf{inf}}$

‘All of us tried to help him to find it.’

5.6.2.5 GDEC

In §4.6.3.3 we stated the following constraint:

(199) **Ordering by Governors’ Degree of Embeddedness Constraint (GDEC)**

All (nonreflexive) dative clitics in the same cluster with the same case are ordered by the degree of embedding of their governors: namely, a clitic governed by a less deeply embedded verb precedes a clitic governed by a more deeply embedded verb. The surface order of the governors is irrelevant. The same probably holds also for personal accusative and possibly genitive clitics.

In other words, for datives personal pronouns in a single cluster, the linear order on domain objects $\prec$ respects the order induced by the queue stayCs:

(200) (GDEC)

$$\vdash_{\textsc{pheno}} \forall a, b : \mathsf{PPron}_{\mathsf{dat}} \,.\, a \prec_{\mathsf{stayCs}} b \Leftrightarrow a \ll_{\mathsf{stayCs}} b$$

If in fact the constraint holds for any case, we could write the following polymorphic constraint:[85]

(202) (GDEC’)

$$\vdash_{\textsc{pheno}} \forall c : \mathsf{Case}\ \forall a, b : \mathsf{PPron}_c \,.\, a \prec_{\mathsf{stayCs}} b \Leftrightarrow a \ll_{\mathsf{stayCs}} b$$

[85] Non-schematically, this would be:

(201) (GDEC”)

$$\vdash_{\textsc{pheno}} \forall a, b : \mathsf{PPron} \,.\, (a.\mathsf{case} = b.\mathsf{case} \,\&\, a \prec_{\mathsf{stayCs}} b) \Leftrightarrow a \ll_{\mathsf{stayCs}} b$$

5.6.2.6 Bonet's constraint

In §4.6.3.4, we rejected Bonet's (1991; 1994) constraint that disallows co-occurence of 1st and 2nd person accusative pronominal clitics with dative pronominal arguments of the same verb. However, as an example we show how such constraint would be formalized:

(203) (Bonet's constraints)

$$\vdash_{\text{PHENO}} \forall a : \mathsf{PPron}_{\mathsf{acc}} \in \mathsf{stayCs}\ \forall d \in \mathsf{compDatCs}\,.\, a.\textsc{tecto}.\mathsf{person} \in \{1,2\} \Rightarrow a \not\equiv_{\mathsf{stayCs}} d$$

The condition $a \not\equiv_{\mathsf{stayCs}} d$ ensures that they do not have the same governor.

5.6.3 Main cluster position

In §4.4, we came to the conclusion that the main clitic cluster occurs after one of the following anchors:

1. the first constituent – this may be the first fronted constituent, the first constituent in rheme-only sentences without fronting, or the complementizer;
2. the first fronted constituent (which may be preceded by a complementizer)
3. all fronted constituents

The constituents are partial in case of split-fronting, otherwise they are full constituents. We also noted, that in most cases all these three possibilities come to one. The reason is that (i) usually one and only one constituent is fronted; (ii) except embedded clauses, fronted expressions are initial.

To formalize this view, we define two meadows corresponding to the informal notions of 1P (first position) and 2P (second/Wackernagel position) used in the Chapter §4:

(204) a. $\mathsf{1P} : \mathsf{Meadow}$ – expressions immediately preceding $\mathsf{2P}$

b. $\mathsf{2P} : \mathsf{Meadow}$ – the main clitic cluster

First, we enforce that the main clitic cluster immediately follows the first position:

(205) (First then Second)

$$\vdash_{\text{PHENO}} \mathsf{1P} \prec^{1} \mathsf{2P}$$

and now we can directly translate the three possibilities for clitic anchors into the following constraint:[86]

(206) (Second position)

$\vdash_{\mathsf{Sign}(\bar{\mathsf{S}}+\mathsf{S}_{\mathsf{fin}}).\textsc{pheno}}\ \mathsf{objs}[\mathsf{1P}] \in \{\langle \mathsf{objs}[1]\rangle, \langle \mathsf{objs}[\mathsf{front}][1]\rangle, \mathsf{objs}[\mathsf{front}]\}$

The domain-object corresponds to partial-constituents in case of split-fronting, otherwise they are full constituents. Finally we need to enforce that 1P is not empty when there are no fronted constituents:

(207) (Clitics not first)

$\vdash_{\mathsf{Sign}}\ \mathsf{required}(\mathsf{1P}, \bar{\mathsf{S}} + \mathsf{S}_{\mathsf{fin}})$

In some colloquial registers, this constraint is not present.

5.6.4 Ordering within a cluster

In §4.5 we analyzed a rather rigid order of clitics in a clitic cluster with the following conclusions (leaving out preferences):

1. auxiliaries < reflexives < adjunct dative < compl. dative < accusative/genitive < *to*
2. ethical dative occurs anywhere after the position of auxiliaries and before the position of complement datives (or accusatives for some speakers),
3. Fringe clitics (e.g., *tu*, *však*, *prý*/*prej*, *už*, *ale*) follow the position of *to*. They can also precede the position of auxiliaries and for some speakers they can even be freely positioned anywhere within the clitic cluster (usually before *to*).

This can be enforced by creating meadows for clitics with relevant properties and then ordering them within stayCs. This approach was used for example by Penn (1999*a*) for Serbo-Croatian clitics in HPSG, or Rosen (2001, §7.4) for Czech clitics in FGD (both formalized in RSRL; Richter 2000). Below, we list the predicates associated with the obvious clitics:

[86] For this constraint to work, we need to assume that a complementizer rises the valency of its object the same way the auxiliaries do (§5.1.8). This means in an embedded clause, the main clitic cluster occurs directly in $\bar{\mathsf{S}}$ without being first in $\mathsf{S}_{\mathsf{fin}}$. The other possibility would be to formulate this constraint as a constraint on sign application – and enforce it "at the last moment", i.e., for every clause ($\mathsf{S}_{\mathsf{fin}}$ or $\bar{\mathsf{S}}$) that is not a daughter of an $\bar{\mathsf{S}}$.

(208) $\mathsf{auxCs}, \mathsf{reflCs}, \mathsf{adjDatCs}, \mathsf{compDatCs}, \mathsf{ethDatCs}, \mathsf{accGenCs}, \mathsf{toCs}, \mathsf{fringeCs} : \mathsf{Meadow}$

(209) (Morpholexical order – basic)
$\vdash_{\text{PHENO}} \mathsf{auxCs} \prec_{\mathsf{stayCs}} \mathsf{reflCs} \prec_{\mathsf{stayCs}} \mathsf{adjDatCs} \prec_{\mathsf{stayCs}} \mathsf{compDatCs} \prec_{\mathsf{stayCs}} \mathsf{accGenCs} \prec_{\mathsf{stayCs}} \mathsf{toCs}$

(210) (Morpholexical order – ethical dative)
$\vdash_{\text{PHENO}} \mathsf{auxCs} \prec_{\mathsf{stayCs}} \mathsf{ethDatCs} \prec_{\mathsf{stayCs}} \mathsf{compDatCs}$

(for some speakers: $\mathsf{auxCs} \prec_{\mathsf{stayCs}} \mathsf{ethDatCs} \prec_{\mathsf{stayCs}} \mathsf{accGenCs}$)

For those speakers that position the fringe clitics freely, we can simply leave them unconstrained, for those which put them either before auxiliaries or after *to*, we can simply define two meadows and require clitics satisfying one of them at the beginning of the cluster and the other at the end:

(211) $\vdash_{\text{PHENO}} \mathsf{fringeCs} = \mathsf{fringeBeforeCs} \cup \mathsf{fringeAfterCs}$

(212) (Morpholexical order – Fringe clitics before/after)
$\vdash_{\text{PHENO}} \mathsf{fringeBeforeCs} \prec_{\mathsf{stayCs}} \mathsf{auxCs} \;\&\; \mathsf{toCs} \prec_{\mathsf{stayCs}} \mathsf{fringeAfterCs}$

While the occupancy of the meadows mostly follows from other constraints (e.g., there cannot be two auxiliaries), it is necessary to enforce that a single clitic cluster can contain only one reflexive:

(213) (Max One Reflexive)
$\vdash_{\text{PHENO}} \mathsf{card}(\mathsf{stayCs}[\mathsf{reflCs}]) \leq 1$

5.6.5 Open Issues

We did not formalize certain properties of clitics described in Chapter 4. We omitted haplology and clitic contractions because, in our view, for proper handling of these we need some morphology. Obviously, for haplology we could have used the same trick as Rosen (2001, p. 231), which in our situation would correspond two domain lists, one with all clitics and the other without haplologized clitics.

We also did not handle the constraints relating climbing and subject control from §4.6.3.2 Because of the unconclusive results, it is not clear to us whether the constraints should refer to syntactic, semantic or simply lexical properties of the verbs. In the current setup, it would be possible to define a tecto terms for verbs with various type of control and then constrain the possibilities of climbing to their domains.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

As stated in the introduction, the thesis has three interrelated goals: (i) description of Czech clitics, (ii) development of a HOG framework allowing transparent and modular treatment of word order, and (iii) testing (ii) by applying it on (i). To a large extent, we have succeeded in all three of these, but at the same time there are many tasks left for future work.

Clitics. In Chapter 4, we have analyzed the phenomenon of Czech special clitics. We have examined the actual set of such elements, their rather rigid placement both relative to a clause and relative to each other, and finally the properties of clitic climbing. The analysis builds on previous research by others, but it also provides new insights, especially in the position of the clitic cluster and in the constraints on clitic climbing.

We have argued that clitics can be positioned either relative to the first constituent or to the fronted expressions, which in most cases results into the same placement. Clitics usually follow the first clausal constituent in a phrase. However, there are many exceptions to this placement. The main cluster can be preceded by a partial constituent on the one hand or by several constituents on the other. We have argued that these are not unusual clitic positions but instead, unusual frontings. Constraints determining which partial or multiple constituents can or cannot precede clitics are better analyzed as constraints on fronting than constraints on clitic placement. In addition we have shown that placement of clitics in embedded clauses is far more regular than has been claimed by various researchers.

While probably not every linguist will share our view that higher-order formalism is the right basis for analyzing Czech clitics, it is worth noting that many of the conclusions in Chapter 4 result from a better understanding of the problem we gained while working on the formalization.

Word order in HOG. In Chapter 2, we have generalized the standard HOG setup where phenogrammar is related to tectogrammar by a homomorphism to a setup where it is a general relation. The grammar then specifies the constraints on this relation both in terms of constraints on individual signs and in terms of constraints on their combinations.

In Chapter 5, we have added a support for capturing and constraining discontinuity of constituents. Our approach was inspired by the HPSG linearization framework (Kathol 2000*a*; Kathol and Pollard 1995; Penn 1999*a*; Reape 1994, 1996), which was successfully applied to various word-order phenomena, including Slavic word order (e.g., Penn 1999*a*; Rosen 2001). The separation of grammar into phenogrammar and tectogrammar in HOG, and the ability to use higher-order functions lead to an increase in the modularity and transparency of grammars.

Caveats. However, there were also areas where, from this point of view, we clearly failed. One such example would be the treatment of subject-predicate agreement (§5.1.9). We had two choices, neither of them completely satisfactory. One possibility was to capture agreement together with sub-categorization requirements of the verb into a single type. Verbs with any idiosyncratic agreement requirements would be assigned another unrelated type. Also in case of auxiliaries combining with participles, the auxiliaries have to capture not only their agreement requirements, but also the requirements of their participles. The other option is to capture only sub-categorization requirements in tectogrammar, and express agreement as a constraint over combinations of whole signs. It is possible to imagine a linguistic argument for either of the two presented solutions of handling agreement. Unfortunately, our choices were driven by the formalism and not by our linguistic insights.

6.2 Future work

Apart from further refining the linguistic analysis of Czech clitics, we would like to continue in elaborating the framework of Higher-Order Grammar. Probably unsurprisingly, we see the further development especially in three areas: linguistic, formal and computational.

Beyond Syntax. First, as suggested throughout the thesis, it is necessary to incorporate layers of language beyond syntax and rudimentary phonology. We would like to focus on determining the proper place of morphology and information structure in the system. Morphology is needed for a more realistic lexicon of inflective languages (recall that in the present setup, a lexicon is simply a list of primitive signs). Higher-order systems have been successfully used in computational models of lexicons for various languages, including Swedish (Forsberg 2004), Arabic (Otakar Smrž p.c.), or Urdu (Humayoun 2006). However, morphology should also be an integral part of the grammar system. In the last section, we suggested that it is desirable for an adequate treatment of phenomena like haplology or contractions. It remains to be seen whether this would imply phenogrammar being essentially morphological with phonology being a separate layer, or whether they would somehow share their place in a more structured phenogrammar.

The same applies to Information Structure. In the previous chapter, we assumed that Information Structure is reflected in features of tecto-terms and made some simplifying assumptions about it. We suggested that it should probably be treated as a separate part of the sign (possibly with other pragmatic information). The constraints over signs would need to relate it to the other parts of the signs, including phonology and semantics.

Increasing expressiveness. The second area is a logical consequence of the caveats described in the conclusion above. We would like to explore more expressive type-systems, especially in the area of subtyping, bounds on polymorphic variables and most-likely some limited version of dependent types. While some linguists are worried about working in an expressive formalism, we do not share this concern. There are usually two arguments against using expressive formalisms: their psycholinguistic in-adequacy and their poor performance in NLP applications. Let us have a closer look at these two arguments, which we do not believe to be conclusive.

If one wishes to write psycholinguistically motivated grammars (no such attempt has been made here), one has two choices: either to choose a formalism where the desired constraints are partially captured by the expressiveness of the formalism, or to choose an expressive formalism and to formulate the constraints as part of the theory within the formalism.

We believe the expressive formalism approach is more flexible and preferable for practical reasons. First, people tend to disagree on what the psycholinguistic constraints are. In the restricted formalism approach, this means they are forced to work in different frameworks, making it harder to communicate and compare the different analyses. In the expressive formalism approach, they may just differ in a couple of constraints within a large grammar. Second, the restricted formalism approach gives rise to a plethora of formalisms that are not only unknown outside of linguistics but even to other linguists in many cases. Formulating one's theory as constraints within a standard formalism has the advantage that the theory is communicable to scientists and engineers in other disciplines.

The other part of the concern relates to the efficiency of NLP applications using such description, where the idea is that less expressive formalism allow for more efficient implementations. The first answer is that HOG is not a programming language. The formalism is intended for linguists to describe and understand the problem, not for actually doing parsing, generation or assist in writing sms on a cell phone. For computational processing, a grammar written in HOG would need to be compiled, optimized and possibly simplified into a form suitable for that particular task, be it a higher-order logic with a simpler type system or maybe even finite state automata. The ideal state is that all this compilation would be done automatically. However, it is possible that some optimizations would need to be done manually. In such a case, it is much easier for a computational linguist or engineer to do optimizations, or even trade-offs between language coverage and efficiency, when it is obvious what the grammar actually states.

The second answer is that in fact, using a standard formalism such as Higher Order Logic, over a specialized formalism such as HPSG/RSRL is advantageous exactly because it is linking this work to a generally established framework. It means that one's parser can build on existing research in implementation techniques for such a formalism, research on compilations, various optimizations (whether beforehand or at the time when actual data are encountered) and heuristics. It is unlikely that the relatively small group of scientists involved in symbolic computational linguistic would be able to replicate the decades of research in this area. Instead, they can focus on problems specific to linguistics.

Finally, note that many of the major linguistic frameworks are maximally expressive in a formal sense, i.e., they are able to simulate a Turing Machine; see for example, (Kepser 2004) for results on RSRL, the logic behind HPSG, (Carpenter 1999) for multimodal categorial grammars, (Peters and Ritchie 1973) for Transformational Grammar of Chomsky's *Aspects of the Theory of Syntax* (1965).[87]

In sum, from our point of view, the question does not seem to be how much theoretical expressivity, but how much practical expressivity a formalism should have in order to express linguistic generalizations. Two formalism may have the same theoretical power, but writing actual linguistic theories may be much easier in one of them.

[87] On the other hand, Combinatory Categorial Grammar (Steedman 2000*b*) or Tree-Adjoining Grammar (Joshi et al. 1975) are more restricted.

Computational model. The third area of further research will be concerned with computation. Naturally, the ultimate goal is to have a parser and generator. In the near future, we would like to explore the feasibility of implementing HOG within an existing higher-order theorem prover, such as HOL (hol.sourceforge.net, Gordon 1989), or MetaPRL (metaprl.org, Hickey 2001; Hickey et al. 2003).

However, even a partial implementation in the form of a type checking algorithm would be highly beneficial. It would enable computer assisted grammar writing and it would also allow automatic discovery of many errors and inconsistencies in HOG grammars, in a similar way as many errors in programs written in typed programming languages can be discovered by compilers.

APPENDIX A

CZECH

The Czech language is one of the West Slavic languages. It is spoken by slightly more than 10 million speakers, mostly in Czechia. In this section, we discuss properties of morphology and syntax of the language relevant to our work. For a more detailed discussion, see for example (Karlík et al. 1996; Petr 1987). Alas, there is no detailed grammar of Czech in English, but basic overviews can be found in (Fronek 1999; Harkins 1953; Janda and Townsend 2002; Naughton 2005; Short 1993).

For historical reasons, there are two variants of Czech: Official (Literary, Standard) Czech and Common (Colloquial) Czech. The official variant is a 19th-century resurrection of 16th-century Czech. Sometimes it is claimed, with some exaggeration, that it is the first foreign language Czechs learn. The differences are mainly in morphology and lexicon. The two variants are influencing each other, resulting in a significant amount of irregularity, especially in morphology.

form	lemma	gloss	category
měst-a	město	town	noun neut sg gen
			noun neut pl nom (voc)
			noun neut pl acc
tém-a	téma	theme	noun neut sg nom (voc)
			noun neut sg acc
žen-a	žena	woman	noun fem sg nom
pán-a	pán	man	noun masc-anim sg gen
			noun masc-anim sg acc
ostrov-a	ostrov	island	noun masc-inanim sg gen
předsed-a	předseda	president	noun masc-anim sg nom
vidě-l-a	vidět	see	verb past participle fem sg
			verb past participle neut pl
vidě-n-a			verb passive participle fem sg
			verb passive participle neut pl
vid-a			verb transgressive masc sg
dv-a	dv-a	two	numeral masc sg nom
			numeral masc sg acc

Table A.1: Homonymy of the *a* ending.

A.1 Morphology

Like other Slavic languages, Czech is a richly inflected language. The morphology is important in determining the grammatical functions of phrases. The inflectional morphemes are highly ambiguous, as Table A.1 shows. There are three genders: neuter, feminine and masculine. The masculine gender further distinguishes the subcategory of animacy. Sometimes, it is assumed that there are four genders: neuter (neut/n), feminine (fem/f), masc. animate (masc/m) and masc. inanimate (inam/i); we follow that practice. In addition to singular and plural, some dual number forms survive in body parts nouns and modifiers agreeing with them.[88] There are seven cases: nominative, genitive, dative, accusative, vocative, locative, instrumental. Only nouns, only in singular, and only about half of the paradigms have a special form for vocative, otherwise the vocative form is the same as nominative.

[88] In Common Czech, there is no dual. The colloquial plural forms are the same as the official dual forms. For example, official: *velkýma rukama* 'big$_{fem.dl.ins}$ hands$_{fem.dl.ins}$' vs. *velkými lžícemi* 'big$_{fem.pl.ins}$ spoons$_{fem.pl.ins}$' (there is no 'hands$_{fem.pl.ins}$' or spoons$_{fem.dl.ins}$); colloquial: *velkejma rukama* 'big$_{fem.pl.ins}$ hands$_{fem.pl.ins}$' vs. *velkejma lžícema* 'big$_{fem.pl.ins}$ spoons$_{fem.pl.ins}$' (according to Oral2006, *ejma ending is the most frequent accounting for 82% of 263 tokens, *ými for 8% *ýma for 10% and *ejmi has no occurrences).

	N	F	F	M	M	I
	Monday	song	fly	Jirka	brother	castle
nom. sg.	pondělí	píseň	moucha	Jirka	bratr	hrad
gen. sg.	pondělí	písně	mouchy	Jirky	bratra	hradu
dat. sg.	pondělí	písni	mouše	Jirkovi	bratru/ovi	hradu
acc. sg.	pondělí	píseň	mouchu	Jirku	bratra	hrad
voc. sg.	pondělí	písni	moucho	Jirko	bratře	hrade
loc. sg.	pondělí	písni	mouše	Jirkovi	bratru/ovi	hradu
ins. sg.	pondělím	písní	mouchou	Jirkou	bratrem	hradem
nom. pl.	pondělí	písně	mouchy	Jirkové	bratři/ové	hrady
gen. pl.	pondělí	písní	much	Jirků	bratrů	hradů
dat. pl.	pondělí	písním	mouchám	Jirkům	bratrům	hradům
acc. pl.	pondělí	písně	mouchy	Jirky	bratry	hrady
voc. pl.	pondělí	písně	mouchy	Jirkové	bratři	hrady
loc. pl.	pondělích	písních	mouchách	Jircích*	bratřích*	hradech
ins. pl.	pondělími*	písněmi*	mouchami*	Jirky*	bratry*	hrady*

* – Official Czech form

Table A.2: Examples of declined nouns.

A.1.1 Nouns

Traditionally, there are 13 basic noun paradigms distinguished – 4 neuter, 3 feminine, 4 animate and 2 inanimate; plus there are nouns with adjectival declension (other 2 paradigms). In addition, there many subparadigms, subsubparadigms. All of this involves a great amount of irregularity and variation. As an illustration, Table A.2 shows declension of few nouns.

A.1.2 Adjectives

Adjectives follow two paradigms: *hard* and *soft*. Both of them are highly ambiguous, filling the 60 (4 genders × (2 numbers × 7 cases + 1 dual form)) non-negated first grade categories with only 12 and 8 forms, respectively (10 and 8 in Common Czech). See Table A.3 for the hard paradigm and Table A.4 for the soft one.

Negation and comparison forms are expressed morphologically. Negation by the prefix *ne-*, comparative by the suffix *-(e)jší-* and superlative by adding the prefix *nej-* to the comparative. The comparative and superlative forms are declined as soft adjectives.

	Official Czech				Common Czech			
	M	I	N	F	M	I	N	F
nom. sg.	mladý		mladé	mladá	mladej		mladý	mladá
gen. sg.	mladého			mladé	mladýho			mladý
dat. sg.	mladému			mladé	mladýmu			mladý
acc. sg.	mladého	mladý	mladé	mladou	mladýho	mladej	mladý	mladou
voc. sg.	mladý		mladé	mladá	mladej		mladý	mladá
loc. sg.	mladém			mladé	mladým			mladý
ins. sg.	mladým			mladou	mladým			mladou
nom. pl.	mladí	mladé	mladá	mladé	mladý*			
gen. pl.	mladých				mladých			
dat. pl.	mladým				mladým			
acc. pl.	mladé		mladá	mladé	mladý*			
voc. pl.	mladí	mladý	mladá	mladé	mladý*			
loc. pl.	mladých				mladých			
ins. pl.	mladými				mladýma			
ins. dl.	mladýma							

* – for neuter, and to some extent for feminine, can be also *mladé*

Table A.3: Hard adjectival paradigm.

	M	I	N	F
nom. sg.	jarní			
gen. sg.	jarního			jarní
dat. sg.	jarnímu			jarní
acc. sg.	jarního	jarní		
voc. sg.	jarní			
loc. sg.	jarním			jarní
ins. sg.	jarním			jarní
nom. pl.	jarní			
gen. pl.	jarních			
dat. pl.	jarním			
acc. pl.	jarní			
voc. pl.	jarní			
loc. pl.	jarních			
ins. pl.	jarními			
ins. dl.	jarníma			

Table A.4: Soft adjectival paradigm.

A.1.3 Pronouns

Some pronouns have nominal declension, some have adjectival declension and some have their own (e.g., *já* 'I'). Some forms of personal pronouns are listed in Table 4.2.

A.1.4 Numerals

Only *jeden* '1', *dva* '2', *tři* '3', and *čtyři* '4' fully decline, all of them distinguishing case and *jeden* and *dva* also gender. The inflection of the other cardinal numerals is limited to distinguishing oblique and non-oblique forms. Numerals expressing hundreds and thousands have in certain categories a choice between an undeclined numeral form or a declined noun form (*sto dvaceti*, *sta dvaceti* '120.genitive'). Ordinal complex numerals have all parts in the ordinal form and fully declining (*dvacátý pátý* '25th')[89]. Similarly as in German, two-digit numerals may have an inverted one-word form (*pětadvacet* '25', lit: five-and-twenty, *pětadvacátý* '25th').

A.1.5 Verbs

As in all Slavic languages, verbs distinguish aspect – perfective and imperfective. Aspect is usually marked by prefixes, sometimes suffixes or by suppletion. Change of aspect is usually accompanied by a change, often subtle, in lexical meaning. For example, *psát* 'write$_{imp}$', *napsat* 'write$_{perf}$', *dopsat* 'finish writing$_{perf}$', *sepsat* 'write up$_{perf}$', *sepisovat* 'write up$_{imp}$', etc. For more information on Czech aspect see (Filip 1999).

There are three tenses – present, past and future. Present tense of imperfective verbs and future tense of perfective verbs is marked inflectionally, distinguishing number and person. Perfective verbs do not have a present tense. The conjugations of perfective future and imperfective present are the same; sometimes they are both called morphological present tense. Past tense and imperfective future is expressed periphrastically.[90] Sample conjugations are in Table A.5. In Common Czech, gender distinction in plural past participles is lost, all being pronounced as the official feminine plural form. Also Common Czech uses adjectives instead of passive participles. Modern Czech does not have a pluperfect or an aorist tense.

[89] Again, this is the case of the official language, complex numerals in Common Czech usually have only their tens and units in ordinal forms.

[90] Note however, that there is no auxiliary in 3rd person past tense. Although some linguists (Veselovská 1995), assume phonologically null auxiliary.

	‘to be’	‘lubricate$_{impf}$’	‘say please$_{impf}$’	‘do/make$_{impf}$’	‘do/make$_{perf}$’
inf	být	mazat	prosit	dělat	udělat
present					
1.sg.	jsem	mažu	prosím	dělám	udělám
2.sg.	jsi	mažeš	prosíš	děláš	uděláš
3.sg.	je	maže	prosí	dělám	udělám
1.pl.	jsme	mažeme	prosíme	děláme	uděláme
2.pl.	jste	mažete	prosíte	děláte	uděláte
3.pl.	jsou	mažou	prosí	dělají	udělají
past prtcp					
M/I sg.	byl	mazal	prosil	dělal	udělal
F sg.	byla	mazala	prosila	dělala	udělala
N sg.	bylo	mazalo	prosilo	dělalo	udělalo
M pl.	byli	mazali	prosili	dělali	udělali
F/I pl.	byly	mazaly	prosily	dělaly	udělaly
N pl.	byla	mazala	prosila	dělala	udělala
pass prtcp					
M/I sg.	-	mazán	prosen	dělán	udělán
F sg.	-	mazána	prosena	dělána	udělána
N sg.	-	mazáno	proseno	děláno	uděláno
M pl.	-	mazáni	proseni	děláni	uděláni
F/I pl.	-	mazány	proseny	dělány	udělány
N pl.	-	mazána	prosena	dělána	udělána
imperative					
2.sg.	buď	maž	pros	dělej	udělej
1.pl.	buďme	mažme	proste	dělejme	udělejme
2.pl.	buďte	mažte	prosme	dělejte	udělejte

Table A.5: Sample Verbal Paradigms (Official Czech).

Five main conjugational types are recognized. Each class has several, quite similar, paradigms (6, 3, 2, 3, 1; 15 in total). Certain categories are expressed analytically; various forms of the verb *být* serve as the auxiliary. Some of the auxiliary forms are constant or inconstant clitics – see §4.3.4.

A.2 Syntax

A.2.1 Agreement

In Czech, there is agreement between subject and predicate and agreement within the NP. Below, we provide a basic overview; for a detailed description of Czech agreement see (Avgustinova et al. 1995).

A.2.1.1 Subject-predicate agreement

One can distinguish two types of agreement with subject:

- Subject – finite verb agreement.

  The finite verb agrees with the subject in person and number.

  (1) Střední Evropa je/*jsem/*jsou ve vzduchoprázdnu.
  Central Europe is$_{3sg}$/am/are$_{3pl}$ in vacuum
  'Central Europe is in vacuum.' [syn6]

- Subject – participles/predicative adjectives agreement

  Predicative adjectives and participles in periphrastic constructions agree in number and gender with subject. In (2), the dropped 2nd person singular (and masculine since referring to *Oto*) subject agrees with the participle *byl* and adjective *zavřený* in number and gender. Similarly *služba* in (3) agrees with *povinná* in number and gender.

  (2) Oto, za co jsi byl/*byla/*byli zavřený/*zavřená/*zavření?
  Ota$_{m.sg}$, for what aux$_{2sg}$ was$_{m.sg}$/was$_{f.sg}$/was$_{m.pl}$ jailed$_{m.sg}$/jailed$_{f.sg}$/jailed$_{m.pl}$
  'Ota, what were you jailed for?' [syn6]

  (3) Vojenská služba je ve Švédsku povinná.
  Military$_{fem.sg}$ service$_{fem.sg}$ is in Sweden obligatory$_{fem.sg}$
  'Military service is obligatory in Sweden.' [syn6]

  Only Official Czech distinguishes gender for plural participles (see Table A.5). In spelling, there are three forms: *chrápali* [-lɪ] 'snored$_{m.pl}$', *chrápaly* [-lɪ] 'snored$_{f/i.pl}$', *chrápala* [-la] 'snored$_{n.pl}$' (note that *chrápali* and *chrápaly* have the same pronunciation). Common Czech uses the [-lɪ] form for all genders in plural (spelling is unclear). Plural adjectives pattern similarly (§A.1.2).

Non-nominative subjects In case of non-nominative subjects (certain numeric expression (4a),[91] (4b), etc.) and constructions that are traditionally analyzed as subject-less (4c or 4d), the predicate is in 3rd person singular neuter form.

(4) a. Pět/Mnoho lodí zmizelo.
five/many ships$_{fem.pl.gen}$ disappeared$_{n.sg}$
'Five/Many ships disappeared'

[91] In similar phrases, the noun in genitive is traditionally seen as the head. Obviously we could also assume the numeral to be the head. In such a case, it would be natural to assume the numeral is in the default form (neuter singular).

b. Otevřít soubor je jednoduché.
open$_{\mathsf{inf}}$ file$_{inam.sg}$ is$_{3sg}$ simple$_{n.sg}$
'To open a file is easy.'

c. Prší/Pršelo.
rains$_{3sg}$/rained$_{n.sg}$
'It is/was raining.'

d. Je mi příjemně.
is$_{3sg}$ me$_{D}$ fine$_{adverb}$
'I am feeling fine.'

Coordinated subjects Agreement with coordinated subjects is rather complex. The gender of the predicate is the minimal gender of participants of coordination, computed under the following order: $m < \{i, f\} < n$. This covers also the trivial case when the gender of all participants is the same. However there is an exception: if all participants have neuter gender and at least one is in singular then the gender of the predicate is feminine. This complexity is absent in Common Czech because colloquial plural participles and adjectives do not distinguish gender. There is a similar hierarchy for determining person of subject with heterogenous persons. Under certain conditions (especially when the predicate precedes the subject, or the subject consists of abstract nouns), the predicate can agree only with the member of the coordinated subject it is closest – as (5c) and (5d) show.

(5) a. *Two concrete nouns:*

Byl jsem rád, že máma s tátou byli/byly v pořádku.
was$_{m.sg}$ aux$_{1sg}$ happy, that mom with dad were$_{m.pl}$/were$_{f.pl}$ fine
'I was happy that mom and dad were fine.' [syn6]

b. Hitler a Německo už měli hotové plány na znovuzískání Horního Slezska ..
Hitler$_{m.sg}$ and Germany$_{n.sg}$ already had$_{m.pl}$ finished$_{A}$ plans$_{A}$ for reclaiming Upper Silezia ..
'Hitler and Germany already had finished plans for reclaiming Upper Silesia ...' [syn5]

c. *Two abstract nouns:*

Přesnost a srozumitelnost je příznačná / jsou příznačné pro jeho výklady.
Accuracy$_{fem.sg}$ and comprehensibility$_{fem.sg}$ is typical$_{fem.sg}$ / are typical$_{fem.pl}$ for his explanations.
'Accuracy and comprehensibility are typical for his explanations.' [Karlík et al. 1996]

d. *Verb preceding subject:*

Včera přišla / přišli máma a táta dumů brzo.
Yesterday came$_{fem.sg}$ / came$_{fem.pl}$ mom and dad home early.
'Yesterday cama mom and dad home early.'

A.2.1.2 Agreement within the NP

So called *agreeing attributes* agree with the noun in gender, number and case. This includes

- normal adjective as *starý* 'old'. For example, in (3) the adjective *vojenská* 'military$_{fem.sg}$' agrees with the noun *služba* 'service$_{fem.sg}$'.
- possessive adjective as *otcův* 'father's'.[92]
- relative clauses. However the relative pronoun agrees with the modified noun only in gender and number; its case is dependent on its function in the relative clause. In Common Czech, relative clauses are often introduced by a universal nondeclined relative pronoun *co. jenž* 'that' is also often not declined.
- ordinal numerals
- possessive pronouns and various determiners

Note that there are some limited exceptions. For historical reasons, attributes modifying accusative or nominative pronouns like *nic* 'nothing' or *něco* iq'something' are in genitive as in (6a) In nominative or vocative, the gender can be feminine even when the noun is not, this gives the phrase an expressive flavor as in (6b).

(6) a. Nikdo z obou pánů nechtěl říci nic konkrétního.
Nobody from both men not-wanted say$_{\mathsf{inf}}$ nothing$_{neut.sg.acc}$ concrete$_{neut.sg.gen}$
'Neither gentleman wanted to say anything concrete' [syn5]

b. Kluku líná!
Boy$_{masc.sg.voc}$ lazy$_{fem.sg.voc}$
'You lazy boy!'

[92] However, in the dialects of Southern and Western Bohemia, including my native dialect, the possessive adjectives do not decline. The form ending in *-ovo* (for masculine possessors) or *-ino* for feminine possessors is used regardless of case, number and gender of the possessed noun. In other dialects this form is used only for accusative singular However, the dialects of Southern and Western Bohemia also often use prenominal genitive to express possession instead, especially when the possessive adjective would involve a phonological change: *s Hanky kolem* 'with Hanka$_{gen}$ bike$_{i.sg.ins}$' for *s Hančino kolem* 'with Hanka's bike$_{i.sg.ins}$' for official *s Hančiným kolem*$_{i.sg.ins}$ 'with Hanka's$_{i.sg.ins}$ bike$_{i.sg.ins}$'.

A.2.2 Numeral expressions

Numerals expressions with *jeden* ‘1’, *dva* ‘2’, *tři* ‘3’, *čtyři*, ‘4’, *oba* ‘both’ behave in a “normal” way: a numeral agrees with its noun in case; *jeden*, *dva* and *oba* also in gender. However, numerals *pět* and above in nominative or accusative positions are followed by nouns in genitive plural (see (4a)). Otherwise (other numerals or other cases), the noun is in the same case as the whole phrase.

A.2.3 Negation

Sentence negation in Czech is formed by the prefix *ne-* attached to the verb. As in the other Slavic languages, multiple negation is the rule, negative subject or object pronouns, adjectival pronouns and adverbs combine with negative verbs.

(7) Nikdy nikomu nic neslibuj.
never nobody$_D$ nothing$_A$ not-promise$_{imper.2sg}$
‘Never promise anything to anybody.’

APPENDIX B

DATA, EXAMPLES, GLOSSES

B.1 Sources of Examples, Corpora

Many of the examples in this thesis are real utterances, usually taken from the Czech National Corpus (CNC, http://ucnk.ff.cuni.cz/) or Prague Dependency Treebank (PDT, http://ufal.mff.cuni.cz/pdt2.0). Well known examples from the linguistic literature are quoted, often accompanied by an example drawn from a corpus. Any example that does not have a source listed is based on my own Czech native competence.

The CNC consists of various subcorpora. Those used as sources of examples in this dissertation are summarized in Table B.1. Some of the corpora contain sentence boundaries and some are tagged with morphological information. This annotation was automatically provided and it is not without errors; depending on the corpus, the accuracy is between 94% and 95.6% (M. Křen, p.c.), so about 1 in 16 to 22 words is tagged incorrectly.[93] The errors are obviously not evenly spread across all linguistic phenomena, and due to the nature of current tagging technology it is likely that there will be more errors in less frequent constructions and especially in constructions involving discontinuities, both of concern in this thesis.

syn2000 and syn2005 are large balanced corpora of contemporary written Czech. The fact that they are balanced means they strive to be representative of a broad range of genres and authors, with

[93] However, note that tags encode 13-14 morphological categories and a mismatch in a single category counts as error even if the other categories were correct.

Corpus	abbreviation	size (M)	balanced	source		period	annotation
syn2000	syn0	100	yes	written	15% fiction	20th cent	morph automatic
					25% non-fiction		
					60% news	1990-99	
syn2005	syn5	100	yes	written	40% fiction	20th cent	morph automatic
					27% non-fiction	1990-2004	
					33% news	2000-04	
syn2006pub	syn6	300	no	written	news	1989-2004	morph automatic
KSK		1	no	written	private letters	1990-2004	no
PMK		0.8	no	spoken	colloquial, Prague	1988-96	no
BMK		0.6	no	spoken	Brno	1994-99	no
Oral2006	oral	1	no	spoken	informal	2002-06	morph automatic
PDT		2	no	written	news, scientific journal		morph, syntax, manual

Table B.1: Corpora used as example sources

	Tokens	Sentences
m-layer (morphology)	1,957,247	115,844
a-layer (surface syntax)	1,503,739	87,913
t-layer (deep syntax, Inf. Struct.)	833,195	49,431

Table B.2: Prague Dependency Treebank: Size of data by layers

proportions reflecting their importance to the language development. Obviously, this last criterion is very subjective. Indeed, the view of importance of various genres changed between publication of the syn2000 and syn2005 corpora as can be seen from the change in proportions of individual genres. The corpora contain so-called "good authors", original and translated fiction, poetry and scripts, tabloids and more serious news, popular and scientific non-fiction, textbooks, etc. For some reason it does not contain private correspondence, e-mails, SMS and similar written media.

The Prague Dependency Treebank is the only corpus used that is not part of the CNC. Its sources are two daily newspapers, a business weekly and a popular scientific journal. The data is annotated on three layers: the morphological layer, analytical layer (surface dependency syntax), and tectogramatical layer (deep dependency syntax, includes Information Structure and coreference). Table B.2 shows the amount of data annotated on each layer (text with the t-layer annotation has the a-layer annotation, text with a-layer annotation has m-layer annotation). The theoretical basis for the annotation comes from Functional Generative Description (Sgall et al. 1986), a dependency grammar theory). The corpus can be searched via the Negraph tool (http://ufal.mff.cuni.cz/netgraph/).

Simplifying somewhat, the spoken corpora can be seen as capturing Common Czech, while the written corpora, especially the news texts, as capturing Official Czech with Common Czech expressions occasionally slipping through. Dialogs in fiction are usually in Common Czech. The KSK corpus of private correspondence is a mixture of both; sometimes even within a single sentence.

The examples that are not mine are accompanied by their source. Examples that come from a corpus but were analyzed by some researcher are annotated as [reference to paper/corpus], e.g., [Rosen 2001 (p.c.) / syn2005]. If it is relevant to mention the original source of a corpus example (e.g., a particular book or newspaper) it is done as [syn2005/M.Viewegh: Účastníci zájezdu].

B.2 Examples – glosses and translations

In glosses, morphological categories that are not systematically expressed in English are marked by subscripts: N/nom = nominative, G/gen = genitive, D/dat = dative, A/acc = accusative, L/loc = locative, I/ins = instrumental; F/fem = feminine, M/(masc-)anim = masculine animate, I/(masc-)inam = masculine inanimate, masc = either masculine, N/neut = neuter; sg = singular, pl = plural, dl = dual. To make reading easier for English speakers, prepositions like *of* for genitive or *to* for dative are added to the glosses of NPs, in addition to the case subscripts. These categories are often not marked when not relevant to a particular problem.

Present auxiliary (present form of the verb *být* 'to be' used to form past tense and colloquially sometimes also conditional) is glossed as 'aux$_{Person.Number}$' (e.g., *jsem* – *aux*$_{1sg}$), other auxiliaries are glossed with the corresponding English auxiliary: future *budu* 'will$_{1sg}$', conditional *bych* 'would$_{1sg}$'. Copula is translated as forms of the verb 'to be'. Past participles are glossed as past tense forms even in periphrastic conditional. Also, the gender and number of past and passive participles is not marked, since they are usually not relevant to the discussion of clitics, but also are completely regular – see §A.1.5 and especially Table A.5.

All 2P clitics (but not prepositions and the like) are given in italics for easier orientation. Often, we use numerical subscripts to show the relation between clitics and the word governing them; the subscripts increase with the degree of embedding of the governors. Clitic auxiliaries have subscript zero.

The pronoun *to*, accusative singular form of the demonstrative pronoun *ten*, sometimes called *universal to* is an inconstant clitic. It has roughly the same meaning that is common to *this* and *that*, in other words it does not distinguish closeness/distance. We gloss it as 'it$_A$', because *it* is usually the closest translation. When used as a determiner, we gloss it as 'that'. The gloss 'that' is also used for the demonstrative pronoun *tamten* and the complementizer *že*, 'this' is used for *tento*.

B.2.0.1 Marking information structure

As discussed in §3.3, word order in Slavic languages, including Czech, correlates to a large degree with information structure. In linguistic literature about Czech, the information structure of the original sentences is usually translated to English by syntactic means, e.g., by passive, fronting, or clefts. Instead, we use mainly prosody. There are two reasons for this. First, my knowledge of English is not deep enough to be able to exactly relate the subtle differences of these English syntactic constructions to Czech information structure. Second, prosody is the primary means for expressing information structure in English. In English intonation, rheme (focus) is marked by so-called *A-accent* (Jackendoff 1972) and contrastive theme (contrastive topic) by fall-rise pitch accent or so-called *B-Accent*. We use capitals for rheme (focus) and sans-serif font for contrast in theme. In Czech sentences, we usually only mark rheme when it does not occur sentence finally, the typical place. When relevant, we mark the contrastive theme in the same way as in English. As an example, consider the sentence in (1) – *Pavel* and *Martin* are in the contrastive theme, while their destinations are rhemed.

(1) (The room with Martin and Pavel must be temporarily vacated for renovation. Their boss discusses with his colleague what to do about the two).

Víš co? Pavel$_C$ pomůže na čtyřce a Martin$_C$ půjde domů.
know$_{2sg}$ what Pavel$_N$ will-help at four and Martin$_N$ will-go home

‘You know what? Pavel$_C$ will HELP AT [THE DEPARTMENT] FOUR$_R$ and Martin$_C$ will GO HOME$_R$.

APPENDIX C

HOL FOR HOG

In this appendix, we present the formalism of Higher Order Grammar. We discuss the additional features of the logic of HOG in comparison with Ty2, the logic used Montague's grammar.[94] The reason is that most linguists are familiar with Montague's grammar and Ty2.

[94] To be precise, Ty2 is (Gallin 1975)'s formalisation of a higher-order logic equivalent to Montague's Intensional Logic.

The higher order logic of HOG is more powerful than Ty2. HOG has a larger set of basic types, including a type of natural numbers. The type system is polymorphic (although only schematically) and allows definition of separation subtypes and supertypes via coproducts. Moreover, while Ty2 has only one type constructor → (defining functional types), HOG has two additional primitive type constructors – products, and coproducts. We discuss each of these extensions in more detail below.

While the core of the thesis focuses mostly on the utility of the framework, this appendix is more formal. In the following, we systematically describe the constructs of the formalism. As is apparent from the text below, the exact configuration of some of the components (e.g., supertypes of infinite number of types §C.4.2) needs further research.

C.1 Ty2, Lambda calculus, Functional types, Basic Types

C.1.1 Ty2

Ty2 (Gallin 1975) is a higher-order logic (HOL) equivalent to Montague's Intensional Logic. It is an extension of Henkin's (1950) logic which in turn is based on Church's Simple Theory of Types. In more detail:

Simple Theory of Types (Church 1940) is a logic obtained from typed lambda calculus by moving term equivalence from meta-language to object-language (thus term equivalence can be stated within a theory instead of imposed externally) and by adding constants for logical connectives and quantifiers. There are two basic types: truth values (in HOG called Bool) and entities (Ent), and one type constructor for creating functional types (→)[95].

Henkin (1950) provided Church's theory with models (Henkin models). He also added the axiom of propositional extensionality, which says that for truth values, there is no difference between equality and bi-implication (equivalence). He showed that all the logical connectives, constants and quantifiers do not have to be assumed as primitives but are lambda definable.

Finally, Gallin (1975) added one more basic type for possible worlds and showed that the resulting system, Ty2, is equivalent to Montague's IL (Montague 1970, 1973).

[95] This constructor is called function space or exponential; $\mathsf{A} \to \mathsf{B}$ is the type of functions from A into B. In some formalisms, $\mathsf{A} \to \mathsf{B}$ is written as $\langle \mathsf{A}, \mathsf{B} \rangle$ (Montague's IL) or $(\mathsf{A}\ \mathsf{B})$.

C.1.2 Basic Types

The set of basic types in HOG is larger than just the three basic types of Ty2 (Bool,[96] Ent, World). The grammar writer specifies the set of basic type, e.g., syntactic categories NP, S, ..., types of feature values Case, Number, ...

C.2 Tuples (Products, Records)

HOG contains indexed tuples, or indexed cartesian products. Tuples (especially when using non-numerical indices) are similar to records in programming languages like Pascal. They are a generalization of the usual binary cartesian product $\mathsf{A} \times \mathsf{B}$ – the type of all pairs $[a, b]$, where $a : \mathsf{A}$ and $b : \mathsf{B}$. The indexed products are indexed by a finite set of indices (e.g., natural numbers or a set of suggestive labels like SUBJ, COMPS, etc.). The indexes are grammar specific. The usual binary product is just an indexed product with indices being 0 and 1.

There are also term constructors. The tupling constructors create tuples from terms and the *projection* constructors, allow accessing the members of such tuples. Two tuples are equal if they have they are equal on their components.

(1) **Definition (Products)**

Let J be a finite set of indexes used in the grammar and let $I \subseteq J$, $I = \{i_1, \ldots, i_n\}$. Let $(\mathsf{A}_i)_{i \in I}$ be a family of types, and $(a_i : \mathsf{A}_i)_{i \in I}$ a family of terms, then we can define:

1. An indexed product type: $\prod_{i \in I} \mathsf{A}_i$, equivalently $[i_1 : \mathsf{A}_1, \ldots, i_n : \mathsf{A}_n]$
2. An indexed tuple: $[i_1 : a_1, \ldots, i_n : a_n]_{(\mathsf{A}_i)_{i \in I}} : \prod_{i \in I} \mathsf{A}_i$
3. Projections: $\pi^j_{(\mathsf{A}_i)_{i \in I}} : \prod_{i \in I} \mathsf{A}_i \to \mathsf{A}_j$

The subscript $(\mathsf{A}_i)_{i \in I}$ is usually omitted.

When the set of indexes are natural numbers, we usually write:

1. $\mathsf{A}_0 \times \ldots \times \mathsf{A}_n$ instead of $[0 : \mathsf{A}_0, \ldots, n : \mathsf{A}_n]$.
2. $[a_1, \ldots, a_n]$ instead of $[0 : a_1, \ldots, n : a_n]$

[96] The type of truth values Bool is a definable type in HOG. $\mathsf{Bool} = \mathsf{Unit} + \mathsf{Unit}$, where Unit is the nullary cartesian product (§C.2) and $+$ is the cartesian coproduct (§C.4.1). The two injections ι_0 and ι_1 are then the constants true and false. However, for linguistics, it does not make a difference whether the type is defined or primitive. In Ty2 notation, Bool is called t, Ent e and World s.

When the set of indexes is empty:

1. The nullary product type: $\mathsf{Unit} = \prod_{i\in\emptyset}$
2. The nullary product term: $* : \mathsf{Unit}$

   Unit and $*$ are formally important because they can serve as a distinguished type and a distinguished term – we know there is a type Unit and there is a term $*$.

Both the types and terms can be written in AVMs in an obvious way:

(1a)
$$\begin{bmatrix} \mathrm{I}_1\text{:} & \mathsf{A}_1 \\ \vdots & \\ \mathrm{I}_n\text{:} & \mathsf{A}_n \end{bmatrix} = \prod_{i\in I} \mathsf{A}_i$$

(1b)
$$\begin{bmatrix} \mathrm{I}_1\text{:} & a_1 \\ \vdots & \\ \mathrm{I}_n\text{:} & a_n \end{bmatrix} = [i_1 : a_1, \ldots, i_n : a_n]$$

(2) **Equations (Products)**

For products, the following equations hold:

(2a) $\pi^j([\ldots, j : e, \ldots]) = e$

(π^j are projections, i.e., they pick the right element)

(2b) $[i_1 : \pi^{i_1}(p), \ldots, i_n : \pi^{i_n}(p)] = p$

(when we take a tuple apart and then put it back together we get the original tuple. Or: two tuples with identical components are identical tuples.)

(2c) $\forall a : \mathsf{Unit}\ .\ a = *$

(there is just one term of the type Unit)

(3) **Example (Valency)**

1. We can define TVv, a type of transitive verb valencies (ignoring agreement, cases, etc.), as:

   (3a) $\mathsf{TVv} = [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$

2. If john, and mary are terms of the type NP, we can use the product term constructor to form a term of the type TVv:

   (3b) $\mathsf{v} = [\textsc{subj}\ \mathsf{john}, \textsc{comps}\ \mathsf{mary}] : \mathsf{TVv}$

3. The projections can be used to access individual NP's of that tuple:[97]

   (3c) $\pi^{\textsc{subj}}(\mathsf{v}) = \textsc{subj}(\mathsf{v}) = \mathsf{john}$

   (3d) $\pi^{\textsc{comps}}(\mathsf{v}) = \textsc{comps}(\mathsf{v}) = \mathsf{mary}$

4. If we assume verbs to be functions receiving their arguments (in a linguistic way of speaking) as arguments (in a mathematical way of speaking) and constructing a sentence, then the type of transitive verbs TV can be

   (3e) $\mathsf{TV} = [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \rightarrow \mathsf{S}$

   that means the type of functions taking two NPs and returning a sentence. Thus $\mathsf{loves} : \mathsf{TV}$ can be applied to the pair v, obtaining a term of type S:

   (3f) $\mathsf{loves}(\textsc{subj}\ \mathsf{john}, \textsc{comps}\ \mathsf{mary}) : \mathsf{S}$

(4) **Example (HOG vs. Java)**

To make the matter clearer, we compare HOG with Java/C++. Constructs in many other programming languages are very similar. The closest thing in Java corresponding to HOG products are so-called classes. Below, we show how to simulate rational numbers in both formalisms. One can see that the two constructs are very similar.

[97] We use the usual convention of simplifying the names of projection functions. Instead of writing $\pi^{\textsc{subj}}(a)$ (or even $\pi^{\textsc{subj}}_{\mathsf{NP},\mathsf{NP}}(a)$), we simply write $\textsc{subj}(a)$ or $a.\textsc{subj}$.

HOG | Java

Type constructor. First it is necessary to define the proper types:

$\mathsf{Ratio} = \langle \mathsf{num} : \mathbb{N}, \mathsf{den} : \mathbb{N} \rangle$

```
class Ratio {
  int num;
  int den;
}
```

Term constructions. A term denoting $\frac{1}{2}$ can be created:[98]

$\langle \mathsf{num} : 1, \mathsf{den} : 2 \rangle_{\mathbb{N},\mathbb{N}}$

```
    Ratio(1, 2)
```

Projections. Multiplication of ratios can be easily defined:

$$\begin{aligned}&\mathsf{mult} = \lambda a, b : \mathsf{Ratio} \ . \ \langle \\ &\quad \mathsf{num} : \mathsf{num}(a) * \mathsf{num}(b), \\ &\quad \mathsf{den} : \mathsf{den}(a) * \mathsf{den}(b) \rangle_{\mathbb{N},\mathbb{N}}\end{aligned}$$

```
Ratio mult(Ratio a, Ratio b) {
    return Ratio(
        a.num * b.num,
        a.den * b.den);
}
```

C.2.1 Currying

Currying is the transformation of a function with multiple parameters into a function that takes a single argument (the first of the arguments of the original function) and returns a new function which takes the remainder of the arguments. For example the curried version of a function of two arguments, is a function of one argument returning another function of one argument.

[98] For various reasons, Java does not provide the term constructors automatically in the way HOG does. So the following term constructor with the obvious meaning must be defined:

```
Ratio(int aNum, int aDen) {
  num = aNum;
  den = aDen;
}
```

Although HOG provides the corresponding term constructor automatically, nothing prevents us from defining the following function:

$\mathsf{ratio} = \lambda num, den : \mathbb{N} \ . \ \langle \mathsf{num} : num, \mathsf{den} : den \rangle_{\mathbb{N},\mathbb{N}}$

It can then be used to construct terms of the type Ratio; $\mathsf{ratio}(1, 2)$ being equivalent to $\langle \mathsf{num} : 1, \mathsf{den} : 2 \rangle_{\mathbb{N},\mathbb{N}}$.

Formalisms without products, like Ty2, require all functions to be curried. Curried functions, especially with multiple arguments and within complex expressions, are usually harder to read than uncurried functions. However, there are cases when working with curried functions can be more convenient. In HOL used by HOG, functions can be freely transformed between their curried and uncurried versions.

(5) **Example (Curried and Uncurried Functions)**

The addition operation in the usual uncurried form is a function of two arguments:

(5a) $\mathsf{plus\text{-}uncur} : (\mathbb{N} \times \mathbb{N}) \to \mathbb{N}$

Using this functions, $3 + 4$ can be calculated as

(5b) $\mathsf{plus\text{-}uncur}(3, 4)$

The curried version of the same function, is a function of one argument returning another function of one argument:

(5c) $\mathsf{plus\text{-}cur} : \mathbb{N} \to (\mathbb{N} \to \mathbb{N})$

and $3 + 4$ is calculated in two steps: first the number 3 is supplied and a function of one parameter adding 3 is returned, then this function is applied to the number 4.

(5d) $\mathsf{plus\text{-}cur}(3)(4)$

The definition of the increment function is easier using the curried version:

(5e) $\mathsf{inc} = \mathsf{plus\text{-}cur}(1)$

while slightly more complicated using the uncurried one:

(5f) $\mathsf{inc} = \lambda x : \mathbb{N} \,.\, \mathsf{plus\text{-}uncur}(1, x)$

Currying and uncurrying are lambda definable, both are polymorphic functions (the type subscripts are usually omitted):

(6) **Definition (Basic Currying and Uncurrying)**

- $\mathsf{curry}_{A,B,C} := \lambda f : (\mathsf{A} \times \mathsf{B}) \rightarrow \mathsf{C}\ \lambda x : \mathsf{A}\ \lambda y : \mathsf{B}\ .\ f(x, y)$
- $\mathsf{uncurry}_{A,B,C} := \lambda f : \mathsf{A} \rightarrow (\mathsf{B} \rightarrow \mathsf{C})\ \lambda(x, y) : \mathsf{A} \times \mathsf{B}\ .\ f(x)(y)$

The currying above works for the usual binary products. It is possible to generalize it to indexed products. Such curry splits the indexes of the product into two sets, returning a function taking a product with the first set of indexes and returning a function taking the second set of indexes. We skip a formal definition and just show the type of such curry function:

$$(7)\qquad \begin{aligned}\mathsf{curry}_{\{\mathsf{f}_1,\ldots,\mathsf{f}_n\},\{\mathsf{G}_1,\ldots,\mathsf{G}_m\},C} :\ & ([\mathsf{f}_1 : A_1, \ldots, \mathsf{f}_n : A_n, \mathsf{g}_1 : B_1, \ldots, \mathsf{g}_m : B_m] \rightarrow C) \rightarrow \\ & ([\mathsf{f}_1 : A_1, \ldots, \mathsf{f}_n : A_n] \rightarrow ([\mathsf{g}_1 : B_1, \ldots, \mathsf{g}_m : B_m] \rightarrow C))\end{aligned}$$

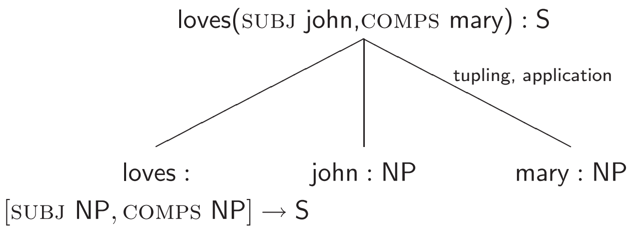
Since $\{\mathsf{G}_1, \ldots, \mathsf{G}_m\}, C$ are clear from context, we omit them. Also, we write $\mathsf{curry}_{\{\mathsf{f}_1,\ldots,\mathsf{f}_n\}}(f)$ as $f^{\mathsf{c}(\mathsf{f}_1,\ldots,\mathsf{f}_n)}$.

(8) **Example (Curried and Uncurried verbs)**

Below, we show three terms corresponding to the sentence *John loves Mary* – with the verb uncurried, object-curried and subject-curried. Each of the terms is also accompanied by a tree structure that shows the structure of the proof denoted by the term – the only proof rules used are the curry function, and implication elimination (modus ponens), corresponding to functional application.

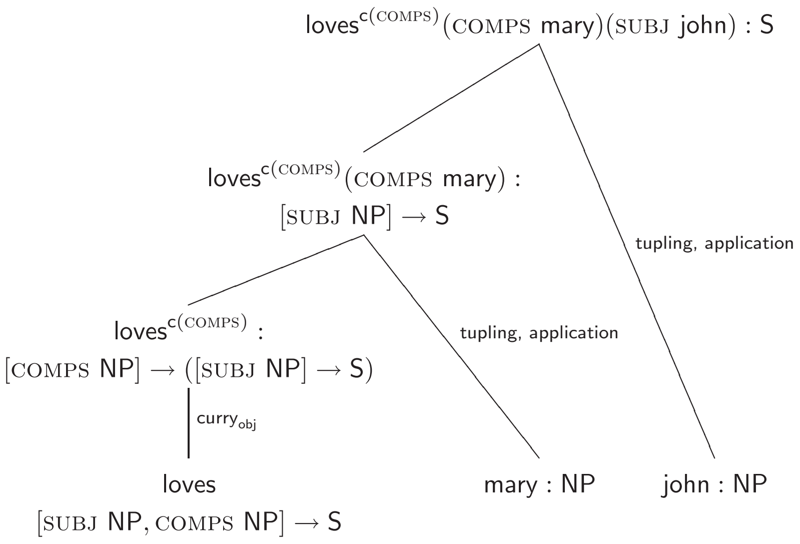
a. Uncurried verb: $\mathsf{loves} : \mathsf{TV} = [\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \rightarrow \mathsf{S}$

The verb combines with its subject and object in one step.



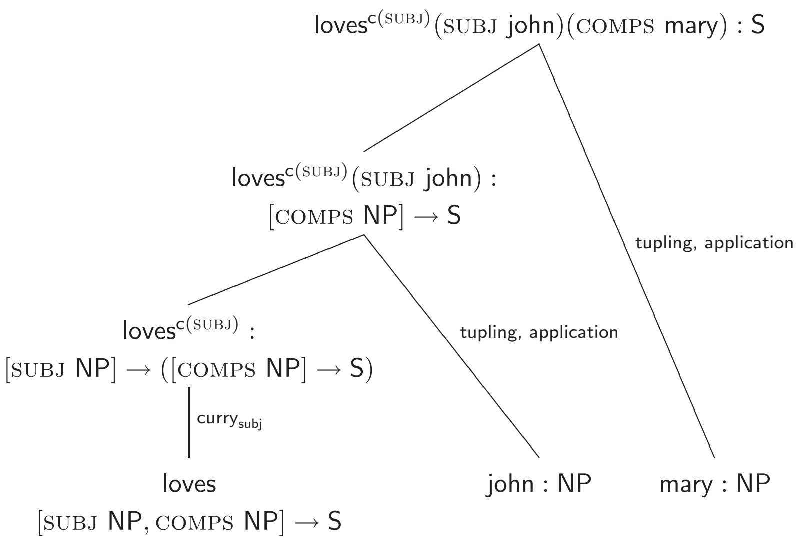
b. Object-curried verb: $\mathsf{loves}^{\mathsf{c}(\textsc{comps})} : [\textsc{comps}\ \mathsf{NP}] \rightarrow ([\textsc{subj}\ \mathsf{NP}] \rightarrow \mathsf{S})$

This corresponds to the usual phrase structure analysis VP → V NP and S → NP VP; the verb combines first with its object and the result combines with the subject.


$\mathsf{loves}^{\mathsf{c}(\textsc{comps})}(\textsc{comps}\ \mathsf{mary})(\textsc{subj}\ \mathsf{john}) : \mathsf{S}$
$\mathsf{loves}^{\mathsf{c}(\textsc{comps})}(\textsc{comps}\ \mathsf{mary}) :$
$[\textsc{subj}\ \mathsf{NP}] \to \mathsf{S}$
tupling, application
$\mathsf{loves}^{\mathsf{c}(\textsc{comps})} :$
$[\textsc{comps}\ \mathsf{NP}] \to ([\textsc{subj}\ \mathsf{NP}] \to \mathsf{S})$
tupling, application
$\mathsf{curry}_{\mathsf{obj}}$
loves
$[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$
mary : NP
john : NP

c. Subject-curried verb: $\mathsf{loves}^{\mathsf{c}(\textsc{subj})} : [\textsc{subj}\ \mathsf{NP}] \to ([\textsc{comps}\ \mathsf{NP}] \to \mathsf{S})$

The verb combines first with its subject and the result combines with the object.


$\mathsf{loves}^{\mathsf{c}(\textsc{subj})}(\textsc{subj}\ \mathsf{john})(\textsc{comps}\ \mathsf{mary}) : \mathsf{S}$
$\mathsf{loves}^{\mathsf{c}(\textsc{subj})}(\textsc{subj}\ \mathsf{john}) :$
$[\textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$
tupling, application
$\mathsf{loves}^{\mathsf{c}(\textsc{subj})} :$
$[\textsc{subj}\ \mathsf{NP}] \to ([\textsc{comps}\ \mathsf{NP}] \to \mathsf{S})$
tupling, application
$\mathsf{curry}_{\mathsf{subj}}$
loves
$[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$
john : NP
mary : NP

The terms and trees *look* different. However, they are all equivalent – the terms denote equivalent proofs of the type/formula S from the same axioms. Currying and uncurrying are part of HOG, therefore if one decides to handle subject and objects via indexed products, as we do in Chapter 5, there is no difference between n-ary branching and binary branching, moreover, the flavor of binary branching would also make no difference.

C.3 Polymorphism

Underlyingly, the type system used in HOG is a first-order type system. A second-order type system (also known as *parametric polymorphism*) allows abstraction and quantification over types. It is possible to define functions accepting types as parameters and passing them as results. The second-order types, i.e., the “types” of types or sets of types are called *kinds*. This means

1. It is possible to freely extend the set of type constructors; for example, to define a function (a type operator) returning multisets of type A.

2. Functions can be generalized over types; for example to define a function min that would return the minimum of any ordered set of the type A.

Many programming languages support various degrees of such polymorphism, e.g., C++ (templates), Java (generics), ML or Haskell.[99] For example, Java allows defining parametric type constructors like `List<A>` (list of some type). This can be used to provide types like `List<Integer>` (list of integers) or `List<Set<Boolean>>` (list of sets of truth values). HPSG also has some sort of parametric polymorphism (e.g., $\mathit{list}(\sigma)$ or $\mathit{set}(\sigma)$ in (Pollard and Sag 1994, :396)), although the details have never been spelled out precisely.

Because the requirements for HOG are different from the requirements on programming languages, HOG replaces full polymorphism with schematic polymorphism.[100] That means a grammar written in a polymorphic formalism can be translated into a formalism without polymorphism. This is possible because for every HOG grammar the number of expressions where a type is passed as a parameter is finite. This approach gives most of the benefits of a polymorphic type-system while avoiding the complexity of models of polymorphically typed logics (hyperdoctrines, see Crole 1993).

The kind of all types is called TYPE. Other kinds can be defined in two ways:

1. By simply listing the types or kinds of types belonging to the kind:

   For example: $\mathsf{NOMINALS} = \{\mathsf{NP}, \mathsf{AP}\}$

   For example: $\mathsf{TWONOMINALS} = \{\mathsf{NOMINALS} \times \mathsf{NOMINALS}\}$

[99] For a short overview of polymorphism in programming languages, see (Cardelli and Wegner 1985).

[100] For example, one disadvantage of a programming language using schematic polymorphism (e.g., templates in C++) is that a compiler has to compile the program as a whole – polymorphic modules cannot be fully precompiled. Although this is a serious practical inconvenience for a programming language, for HOG, as a mathematical formalism, this is irrelevant.

2. By closing a kind with all or some the available type constructors.[101]

   For example: $\mathsf{TECTO} = \mathsf{closeKind}(\{\mathsf{NP}, \mathsf{N}, \mathsf{S}\})$

   The kind TECTO contains NP, N and S, but also $[\textsc{subj}\ \mathsf{NP}]$, $[\textsc{subj}\ \mathsf{NP}] \to \mathsf{S}$, $\mathsf{NP}_{\mathsf{acc}}$, etc. (assuming there is a product feature SUBJ and there is a predicate acc on NP).

Informally, the set of types is as follows:

1. type variables: $\mathsf{X}, \mathsf{Y}, \ldots$
2. basic types: $\mathsf{NP}, \mathsf{S}, \mathsf{Case}, \mathsf{Prop}, \mathbb{N}, \ldots$
3. $\mathsf{F}(\mathsf{T}_1, \ldots, \mathsf{T}_n)$ where T_i are types and F is a type constructor.

   $\to$, $\prod_{i \in I}$, $\coprod_{i \in I}$, and List are among type constructors. Thus $\mathsf{List}(\mathsf{NP})$ is a type of lists of NPs ($\mathsf{F} = \mathsf{List}, \mathsf{T}_1 = \mathsf{NP}$).
4. $\forall \mathsf{X} : \mathsf{K}\ .\ \mathsf{T}$, where K is a kind and T is a type. (K is omitted if it is TYPE)

The following example illustrates the polymorphic type expressions and the usually adopted simplified notation.

(9) **Example (Reversing lists.)**

Consider a function $\mathsf{reverse}$ taking a list of any type as its argument and returning a reversed list. The result is of course of the same type as the argument. The properly written type of such a function would be:

(9a) $\mathsf{reverse} : \forall \mathsf{A}\ .\ \mathsf{List}(\mathsf{A}) \to \mathsf{List}(\mathsf{A})$

but we usually write just:

(9b) $\mathsf{reverse} : \mathsf{List}(\mathsf{A}) \to \mathsf{List}(\mathsf{A})$

The definition of that function involves two abstractions – one over types of list elements (marked by the Λ operator) and one usual abstraction over lists to reverse (marked by the λ operator):

(9c) $\mathsf{reverse} = \Lambda \mathsf{A}\ \lambda x : \mathsf{List}(\mathsf{A})\ .\ \ldots$

[101] Except infinite coproducts – see §C.4.2.

but we usually write type variable as subscripts:

(9d) $\mathsf{reverse}_{A} = \lambda x : \mathsf{List}(A)\,.\,\ldots$

Moreover, the subscripts are usually omitted:

(9e) $\mathsf{reverse} = \lambda x : \mathsf{List}(A)\,.\,\ldots$

All of the above is purely schematic. A polymorphic definition must be regarded in a similar way as a macro – it is an abbreviatory notation. If a grammar used lists of integers and list of noun phrases, formally the grammar contains two unrelated non-polymorphic types.

C.4 Subtypes and supertypes

There are two ways how to express a type-subtype relationship in HOG. One is tow define supertypes via coproducts, the other is to define subtypes via predicates:

1. For any countable set of types we can define a supertype of those types:

   (10) $\mathsf{NominalP}$ is the type of all noun phrases and adjectival phrases:
   $\mathsf{NominalP} = \mathsf{NP} + \mathsf{AP}$

   (11) Tecto is the type of all tecto phrases:
   $\mathsf{Tecto} = \mathsf{ClosingSupertype}(\{\mathsf{NP}, \mathsf{N}, \mathsf{S}\}) = \coprod_{A:\mathsf{closeKind}(\{\mathsf{NP},\mathsf{N},\mathsf{S}\})} A$

2. For any type and a predicate on that type, we can define a subtype determined by the predicate:

   (12) The type of all accusative noun phrases:
   $\mathsf{NP}_{\lambda x:\mathsf{NP}.\mathsf{case}(x)\mathsf{acc}}$ or $[\,x : \mathsf{NP} \,|\, \mathsf{case}(x) = \mathsf{acc}\,]$, usually written as $\mathsf{NP}_{\mathsf{acc}}$

Below, we discuss both of these possibilities in more formal detail.

C.4.1 Coproducts (disjoint unions)

Intuitively, a coproduct $A+B$ is a type that can contain terms of type A or of type B. This is similar to partitioning types in HPSG signature, for example $\mathsf{Head} = \mathsf{Substantive} + \mathsf{Functional}$ is equivalent to saying the type Head is the supertype of the types $\mathsf{Substantive}$ and $\mathsf{Functional}$.

The formal machinery is slightly more complicated – the coproduct requires the proper injections: if $a : \mathsf{A}$ then $\iota^0_{\mathsf{A}+\mathsf{B}}(a) : \mathsf{A}+\mathsf{B}$ and if $b : \mathsf{B}$ then $\iota^1_{\mathsf{A}+\mathsf{B}}(b) : \mathsf{A}+\mathsf{B}$. Or when viewed from the other side: if $x : \mathsf{A}+\mathsf{B}$ then either $\exists a : \mathsf{A} \,.\, x = \iota^0_{\mathsf{A}+\mathsf{B}}(a)$ or $\exists b : \mathsf{B} \,.\, x = \iota^1_{\mathsf{A}+\mathsf{B}}(b)$. The injections can be easily omitted because they retrievable from the context.[102]

The notion of coproducts, or disjoint unions, is dual to the notion of products. In programming languages, they have various, usually less universal or abstract, counterparts. For example switch/case statements in Java/C++ are very similar. But so are, in other point of view, unions in C++ or variant records in Pascal. In case of products, one can see tuples as datastructures and projections as programs (although trivial) for accessing the data. In the case of coproducts, it is the other way round – the injections represent (tagged) data structures and co-tuples are programs for accessing and manipulating the data.

(13) **Definition (Coproducts)**

Let J be a finite set of indexes used in the grammar and let $I \subseteq J$, $I = \{i_1, \ldots, i_n\}$. Let $(\mathsf{A}_i)_{i\in I}$ be a family of types, and $(a_i : \mathsf{A}_i)_{i\in I}$ a family of terms, then we can define:

- An indexed coproduct type: $\coprod_{i\in I} \mathsf{A}_i$, equivalently $(i_1 : \mathsf{A}_1, \ldots, i_n : \mathsf{A}_n)$
- An indexed co-tuple: $(i_1 : a_1, \ldots, i_n : a_n)_{(\mathsf{A}_i)_{i\in I}} : \coprod_{i\in I} \mathsf{A}_i$
- Injections: $\iota^j_{(\mathsf{A}_i)_{i\in I}} : \mathsf{A}_j \to \coprod_{i\in I} \mathsf{A}_i$

The subscript $(\mathsf{A}_i)_{i\in I}$ is usually omitted.

When the set of indexes are natural numbers (which is the case of all coproducts in this thesis), we usually write:

- $\mathsf{A}_0 + \ldots + A_n$ instead of $(0 : \mathsf{A}_0, \ldots, n : \mathsf{A}_n)$.
- $(a_1, \ldots, a_n)$ instead of $(0 : a_1, \ldots, n : a_n)$

When the set of indexes is empty, we get:

- The nullary coproduct type: $\mathsf{Zero} = \coprod_{i\in\emptyset}$
- There is no term of the type Zero

[102] They are not uniquely retrievable for "nonlinear" coproducts of the form $\mathsf{A}+\mathsf{A}$ (more than once the same type). However, we cannot think of any linguistic motivation for such types.

The only type of such form is $\mathsf{Bool} = \mathsf{Unit} + \mathsf{Unit}$. But that's more a consequence of the formal game of trying to assume as few primitive types as possible. For linguistics, it would not make any difference if Bool would be a primitive type and the "nonlinear" products were prohibited.

(14) **Equations (Coproducts)**

(14a) $(\langle \ldots, j : e, \ldots \rangle)\iota^j = e$

ι^j are injections.

C.4.2 Arbitrary coproducts

We assume that the formalism allows infinite coproducts and we can therefore define a supertype for every countable set of types (a kind). Thus, for example if TECTO is the kind of all tecto phrases, defined as ($\mathsf{closeKind}$ is introduced in §C.3):

(15) $\mathsf{TECTO} = \mathsf{closeKind}(\{\mathsf{NP}, \mathsf{N}, \mathsf{S}\})$

We can define the supertypes of all types in this kind as:

(16) $\mathsf{Tecto} = \coprod_{A:\mathsf{TECTO}} A$

Tecto is a supertype of all the basic types (NP, N, S), but also of all the types constructed on top of them by any possible type constructor (except infinite coproduct, thus $\mathsf{Tecto} \notin \mathsf{TECTO}$), for example

(17) $[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$
$[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}] \to \mathsf{S}$
$[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ [\textsc{subj}\ \mathsf{NP}] \to \mathsf{S}] \to \mathsf{S}$
$[\textsc{spec}\ \mathsf{N}] \to \mathsf{NP}$
$\vdots$

This is written simply as

(18) $\mathsf{Tecto} = \mathsf{ClosingSupertype}(\{\mathsf{NP}, \mathsf{N}, \mathsf{S}\})$

It is also possible to specify the closing type constructors (particular products/records can be specified via the indices). For example, the following specifies all record types over the the type NP with indices SUBJ and COMPS.

Therefore,

(19) $\mathsf{ClosingSupertype}(\{\mathsf{NP}\}, \{\textsc{subj}, \textsc{comps}\})$

is a supertype of the terms of the following types:

(20) $[\textsc{subj}\ \mathsf{NP}]$
$[\textsc{comps}\ \mathsf{NP}]$
$[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ \mathsf{NP}]$
$[\textsc{subj}\ \mathsf{NP}, \textsc{comps}\ [\textsc{subj}\ \mathsf{NP}]]$
$\vdots$

but not, for example, of

(21) NP
$\mathsf{S}_{\mathsf{fin}}$
$[\textsc{comps}\ \mathsf{NP}] \to \mathsf{NP}$
$\vdots$

The consequences of adding infinite coproducts to the formalism are far from obvious and further research is needed in this area of HOG. Arbitrary coproducts can be formed, for example, in the so-called copowered toposes (Bell 1982, p. 319).

C.4.3 Separation Subtyping

HOG supports a very powerful kind of subtyping, so called separation types (Lambek and Scott 1986). For every type A and every predicate $\varphi : \mathsf{A} \to \mathsf{Bool}$ there is a type A_φ whose members are exactly those members of A that have the property φ. In other words, φ is the characteristic function of A_φ. The subtype A_φ is also written as

(22) $[\, x : \mathsf{A} \,|\, \varphi(x) \,]$

For example, if NP is the type of noun phrases, $\mathsf{Case} = \{\mathsf{nom}, \mathsf{acc}\}$ is a type, and $\mathsf{case} : \mathsf{NP} \to \mathsf{Case}$ is a function assigning a case to every NP, then there is, for example, the type of accusative NP's:

(23) $\mathsf{NP}_{\lambda x:\mathsf{NP}.\mathsf{case}(x)=\mathsf{acc}} = [\, x : \mathsf{NP} \,|\, \mathsf{case}(x) = \mathsf{acc} \,]$

Often, for predicates of the form $\lambda x\,.\,f(x) = c$ when it is clear which function f is used, we simply write only the constant c:

(24) $\mathsf{NP}_{\mathsf{acc}} = \mathsf{NP}_{\lambda x:\mathsf{NP}\,.\,\mathsf{case}(x)=\mathsf{acc}}$

C.4.3.1 Boolean algebra

The subtypes of a given type form a boolean algebra, it is therefore possible to define intersection and union of subtypes:

$$\text{(25)}\quad \begin{array}{lllll} \text{top:} & \mathsf{A} & & & \\ \text{meet (union):} & \mathsf{A}_\varphi \cap \mathsf{A}_\psi & = \mathsf{A}_{\varphi,\psi} & = \mathsf{A}_{\varphi\&\psi} & = [\, x : \mathsf{A} \,|\, \varphi \,\&\, \psi \,] \\ \text{join (intersection):} & \mathsf{A}_\varphi \cup \mathsf{A}_\psi & & = \mathsf{A}_{\varphi\vee\psi} & = [\, x : \mathsf{A} \,|\, \varphi \vee \psi \,] \\ \text{bottom:} & \mathsf{A}_{\mathsf{false}} & & & = [\, x : \mathsf{A} \,|\, \mathsf{false} \,] \\ \text{difference} & \mathsf{A}_\varphi - \mathsf{A}_\psi & & = \mathsf{A}_{\varphi\&\neg\psi} & = [\, x : \mathsf{A} \,|\, \varphi \,\&\, \neg\psi \,] \end{array}$$

C.4.4 Problem with Monotyping

The formalism of HOL requires that every term belongs to exactly one type (so-called *monotyping* property). Therefore functions defined for a particular type cannot be applied to objects of subtypes of that type (e.g., function defined for $\mathbb{N}$ cannot be applied to the type of even natural numbers). It is possible to weaken such property , but there is a formally simpler solution with the same practical consequences.

For every two types A, B, where A is a subtype of B ($\mathsf{A} \sqsubset \mathsf{B}$), there is an obvious function $\mathsf{ker}_{\mathsf{A},\mathsf{B}}$ mapping the elements of tvA on the corresponding elements of B:

- In case of predicate subtyping, this function is determined by the predicate.

  $\mathsf{ker}_{\mathsf{B}_a,\mathsf{B}} = \mathsf{ker}_a : \mathsf{B}_a \to \mathsf{B}$ such that:

  1. $\forall x, y : \mathsf{B}_a\,.\,\mathsf{ker}_a(x) = \mathsf{ker}_a(y) \Rightarrow x = y$
  2. $\forall x : \mathsf{B}\,.\,a(x) \Leftrightarrow \exists y : \mathsf{B}_a\,.\,x = \mathsf{ker}_a(y)$

  Note that ker must be a primitive, not a term, because it is impossible to assign a type to such function in the formalism. We would need dependent types to be able to do it – the ‘type’ of the function is dependent on the predicate a.

- In case of coproducts, this function is the appropriate injection. For example, if $\mathsf{B} = \mathsf{A} + \mathsf{A}_1$ then $\mathsf{ker}_{\mathsf{A},\mathsf{B}} = \iota_0$.

Now, let $g : \mathsf{B} \rightarrow \mathsf{C}$, and let $\mathsf{A} \sqsubset \mathsf{B}$, then g can be

- directly applied to objects of type B: $g(x)$ for $x : \mathsf{B}$.
- indirectly applied to objects of type A via $\mathsf{ker}_{\mathsf{A},\mathsf{B}}$: $g(\mathsf{ker}_{\mathsf{A},\mathsf{B}}(y))$ for $y : \mathsf{A}$.

Since for a particular grammar, the inclusion functions ker can be derived from context (except for types of the form $\mathsf{A}+\mathsf{A}$), they can be omitted, so it is possible to write $g(y)$ instead of $g(\mathsf{ker}_{\mathsf{A},\mathsf{B}}(y))$.

C.4.5 Term-of-type (::) predicate

For subtypes of a type we can define a predicate testing whether a term is in that particular subtype.

If A is a subtype of B, then $a :: \mathsf{A}$ is sugar for $\exists b : \mathsf{A} \,.\, \mathsf{ker}_{\mathsf{A},B}(a) = b$.

C.5 Natural numbers and induction

The logic contains a type of the natural numbers, written as $\mathbb{N}$, as a primitive type. Unlike the other primitive types, this is an infinite type. Thus adding the type is equivalent to adding the axiom of infinity.

The type goes with several terms: numbers (zero and the successor function) and a primitive recursor for induction:

(26) **Definition (Terms on $\mathbb{N}$)**

$\mathsf{0} : \mathbb{N}$

$\mathsf{succ} : \mathbb{N} \rightarrow \mathbb{N}$

$\mathsf{ind} : \mathsf{A} \times (\mathbb{N} \times \mathsf{A} \rightarrow \mathsf{A}) \times \mathbb{N} \rightarrow \mathsf{A}$

(27) **Equations (Induction)**

$\mathsf{ind}(x, f, 0) = x$

$\mathsf{ind}(x, f, \mathsf{succ}(n)) = f(n, \mathsf{ind}(x, f, n))$

There are also the usual Peano axioms.

(28) **Example (Induction)**

Using the induction function, it is possible to define many recursive functions. As an example, we show how to define addition and factorial.

- Addition. Addition can be recursively defined in the following way:

  (28a) $\mathsf{add}(i, 0) = i$

  (28b) $\mathsf{add}(i, j) = \mathsf{succ}(\mathsf{add}(i, j-1))$

  Thus in HOG, we can define the function as:

  (28c) $\mathsf{add} = \lambda i, j : \mathbb{N} \,.\, \mathsf{ind}(i, \lambda n, k : \mathbb{N} \,.\, \mathsf{succ}(k), n)$

  The variable k of the induction steps corresponds to $\mathsf{add}(i, j-1)$. Note that the induction step ignores the depth of recursion (n): 1 is always added regardless whether it is the first or 25th addition.

  Multiplication is analogous.

- Factorial.

  (28d) $\mathsf{factorial}(0) = 1$

  (28e) $\mathsf{factorial}(j) = j * \mathsf{factorial}(j-1)$

  Thus the function can be defined as:

  (28f) $\mathsf{factorial} = \lambda j : \mathbb{N} \,.\, \mathsf{ind}(1, \lambda n, k : \mathbb{N} \,.\, n * k, j)$

  The variable k of the induction steps corresponds to $\mathsf{factorial}(j-1)$. This time we cannot ignore the depth of recursion – when called for the 1st time it multiplies by 1, when for the 25th time, by 25:

C.6 Collections

C.6.1 Sets

The sets are modeled via their characteristic function. That means a set of objects of type A is a function $\mathsf{A} \to \mathsf{Bool}$, we write this type as $\mathsf{Set}(\mathsf{A})$. In addition, we assume there is a singularizer, a function that returns the only member of a singleton set and is undefined for other sets:

(29) **singularizer**

$\mathsf{sing} : \mathsf{Set}(\mathsf{A}) \to \mathsf{A}$

$\mathsf{sing}(\{\mathsf{sg}\}) = \mathsf{sg}$

It is then possible to define all the usual operations (type subscripts are usually omitted):

(30) **set terms**

empty set: $\emptyset_{\mathsf{A}} = \lambda x : \mathsf{A}\,.\,\mathsf{false}$

membership: $x \in s \Leftrightarrow s(x)$

set terms: $s = \{a, b, c\} \Leftrightarrow s(a)\ \&\ s(b)\ \&\ s(c)\ \&\ (\forall x : \mathsf{A}\,.\,x \neq a\ \&\ x \neq b\ \&\ x \neq c \Rightarrow \neg s(x))$

cardinality: $\mathsf{card} : \mathsf{Set}(\mathsf{A}) \to \mathbb{N}$

subset: $s_1 \subseteq s_2 \Leftrightarrow \forall x. x \in s_1 \Rightarrow x \in s_2$

C.6.2 Lists (Kleene star)

Lists are defined as functions from natural numbers (indexes) to the type of the elements. However, since HOG allows only total functions and does not have dependent types, such a function could be directly used only for infinite lists. The finite lists are then defined as equivalence classes on those infinite lists where members are considered only up to certain point.[103] This means that lists are lambda-definable within HOG. Of course, users of the formalism do not have to care about the way in which the type constructor List() is defined and they can use it as if it were a primitive type. Any computational implementation of HOG would also implement lists directly as primitives.

[103] The polymorphic type List(A) is defined in two steps. First, the auxiliary (polymorphic) type Prelist:

(31a) $\mathsf{Prelist}(\mathsf{A}) = [\textsc{iList}\colon \mathbb{N} \to \mathsf{A}, \textsc{len}\colon \mathbb{N}]$

Members of this type are pairs where iList is an infinite list and len is the length of the modeled list. The real lists are defined as equivalence classes on Prelists, where the irrelevant elements (i.e., anything beyond len) are ignored. Below is he corresponding equivalence relation – it considers prelists equivalent if they have the same relevant elements:

(31b)
$$\begin{array}{l}\mathsf{same}(l_1 : \mathsf{Prelist}(\mathsf{A}), l_2 : \mathsf{Prelist}(\mathsf{A})) : \mathsf{Bool} = \\ \qquad \textsc{len}\ (l_1) = \textsc{len}\ (l_2) \quad \& \quad \forall i \in \textsc{len}\ (l_2)\,.\,\textsc{iList}\ (l_1)(i) = \textsc{iList}\ (l_2)(i)\end{array}$$

Now it is possible to define the real type constructor:

(31c) $\mathsf{List}(\mathsf{A}) = [\,x : \mathsf{Set}(\mathsf{Prelist}(\mathsf{A}))\,|\,\exists q \in x\ \forall p : \mathsf{Prelist}(\mathsf{A})\,.\,p \in x \Rightarrow \mathsf{same}(q, p)\,]$

and basic functions for working with lists:

(31d) $\mathsf{length}(l : \mathsf{List}(\mathsf{A})) : \mathbb{N} = \mathsf{sing}\{\textsc{len}\ (k) \mid k \in l\}$

(31e) $\mathsf{itemAt}(l : \mathsf{List}(\mathsf{A}), i : \mathbb{N}) : \mathsf{A} = \mathsf{sing}\{\textsc{iList}\ (k)(i) \mid k \in l\}$

(32) **list types**

List	(type constructor, arity = 1)
$*$	(equivalent notation, Kleene-star)

(33) **list terms**

$\langle e_1, \dots e_n\rangle_{\mathsf{A}}$	$: \mathsf{List}(\mathsf{A})$	(list of length n)
len	$: \mathsf{List}(\mathsf{A}) \to \mathbb{N}$	(list length)
itemAt	$: \mathsf{List}(\mathsf{A}) \times \mathbb{N} \to \mathsf{A}$	(indexed list access)
$\_[\_]$	$: \mathsf{List}(\mathsf{A}) \times \mathbb{N} \to \mathsf{A}$	(equivalent notation, $l[i] = \mathsf{itemAt}(l, i)$)

C.7 Models and Proofs in HOG

C.7.1 Logic of types, Curry-Howard isomorphism

The Curry-Howard isomorphism (Curry and Feys 1958; Howard 1980) states that the type system forms a logic. Type expressions, like NP, $\mathsf{NP} \times \mathsf{NP} \to \mathsf{S}$ are propositions in that logic. In fact, the Curry-Howard isomorphism is not really a correspondence between a type system and a logic but simply two different views at the same thing.

The type system of the logic presented in this appendix (with functional types, products and coproducts) can be viewed as a full intuitionistic propositional logic[104] with the following correspondence between the names usually used in logic and those used when talking about types:

(34)

$\Rightarrow$	implication	$\to$	function space (exponential)
$\&$	conjunction	$\times$	product
$\vee$	disjunction	$+$	coproduct (disjoint union)
true	true	Unit	nullary product
false	false	Zero	nullary coproduct
$\neg$	negation		defined as $\neg \mathsf{A} = \mathsf{A} \to \mathsf{Zero}$
	atomic formulas		basic types

and finally some syntactic sugar for specifying lists:

(31f) $[e_1, \dots e_n]_{\mathsf{A}}$ means $\{[\text{ILIST}\colon k, \text{LEN}\colon m] : [\mathbb{N} \to \mathsf{A}, \mathbb{N}] \mid m = n \;\&\; k(i) = e_i\}$

[104] Exactly implicative intuitionistic propositional logic – a propositional logic with implication and without the rule of excluded middle Thus one cannot prove $\neg A \vee A$, but also double negation $\neg\neg A = A$, or Peirce's Law $((A \Rightarrow B) \Rightarrow A) \Rightarrow A$. It can be proven that $A \Rightarrow \neg\neg A$, but the inverse $A \Rightarrow \neg\neg A$ cannot be proven. So double negation can be introduced but not eliminated. Such logic allows constructive proofs.

C.7.1.1 Proofs and type inhabitance

However, the view of type expressions as formulas is only half of the Curry-Howard isomorphism. The other part is identifying the closed lambda-terms with proofs. A type expression i.e., a formula in the logic of types, is a theorem if there is a term of that type. In other words, a type is "true" if it is inhabited, which means that in the model of the logic, there are objects which are members of the interpretation of the type. The table in (35) relates the term usually used when speaking about proofs with terms usually used when speaking about the objects as lambda-terms.

(35)

proofs of formula A	terms of type A
nonlogical axioms	constants
undischarged hypotheses	free variables
proof reduction (cut)	beta-reduction

C.7.1.2 A Matter of presentation – Introduction and elimination rules

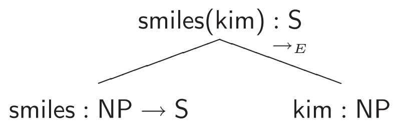
The fact that the type system forms a logic becomes more apparent, when the definitions and equations used to introduce various type expressions are presented in the forms of introduction and elimination rules – the more usual style of introducing logical constants. For example, one can rewrite the definitions and equations for product types (for simplicity, only for binary products) in such form. Then pairing is conjunction introduction and projections are left and right conjunction elimination. When this is done as in (36), it is clear that products *is* conjunction. Similarly, functional types correspond to implication, with abstraction corresponding to implication introduction (hypothetical proof) and application corresponding to implication elimination (modus-ponens).

$$(36) \quad \frac{a : A \quad b : B}{[a,b] : A \times B}\times\text{intro} \qquad \frac{[a,b] : A \times B}{\pi^0[a,b] : A}\times\text{left elim} \qquad \frac{[a,b] : A \times B}{\pi^1[a,b] : B}\times\text{right elim}$$

C.7.1.3 Proof trees

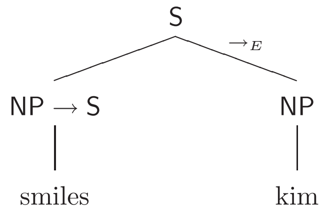
The proofs can be represented as trees. For example, assume $\mathsf{kim} : \mathsf{NP}$ and $\mathsf{smiles} : \mathsf{NP} \to \mathsf{S}$. This means two things. First, the terms kim and smiles have the types NP and $\mathsf{NP} \to \mathsf{S}$. Second, under Curry-Howard isomorphism, the term kim denotes a proof of the formula/type NP in the type logic. The term smiles denotes a proof of the formula/type $\mathsf{NP} \to \mathsf{S}$. The term $\mathsf{smiles}(\mathsf{kim})$ then denotes a proof of the formula/type S. The whole proof can be depicted in a tree:

(37)



The leaves correspond to the axioms kim : NP and smiles : NP → S, only one inference rule was used – implication elimination $\to_E$ or modus ponens. This rule in type logic corresponds to function application in the calculus of terms. The tree is just a notation, the same proof can be depicted by a tree that looks more like a phrase-structure:

(38)



C.7.2 Semantics

The models of Ty2 are Henkin models (Henkin 1950). The models of HOG are also Henkin models[105] but enhanced with interpretations for products, coproducts and subtypes:

1. Products are interpreted as cartesian products.
   (a) Unit as $\{0\} = 1$
   (b) $*$ as identity on 1 (id_1).
2. Coproducts as disjoint unions
   (a) Zero as the empty set
   (b) Bool $=$ Unit $+$ Unit as $\{0, 1\} = 2$
3. Subtypes are interpreted as subsets: If type A is interpreted as X, then a subtype A_φ is interpreted as a set Y, where $Y \subseteq X$ and the characteristic function of Y relative to X is the function interpreting φ (restricted to X). ker_φ is then interpreted as an injection from Y to X.

[105] This is not exactly true because set-theoretical Henkin models do not allow noninhabited types other than Zero (the nullary coproduct). For that, categorical generalizations of sets, toposes, are needed (see Crole 1993; Lambek and Scott 1986).

When modeling syntax of a natural language in HOG, (some) types denote syntactic categories (sets of expressions), terms of those types denote syntactic expressions and constants of those types denote forms in a lexicon. Similarly, phenogramatical terms denote actual expressions and semantic terms denote their meaning.

This means that types and lambda terms have a model. On the other hand, because of Curry-Howard isomorphism (§C.7.1), the type expressions form a propositional logic, with lambda terms being the proofs in such logic. Thus in HOG, the model-theoretic aspect (à la HPSG (Pollard and Sag 1994)) is automatically connected with its proof-theoretic aspect (à la Type Logical Grammar (Morrill 1994)).

APPENDIX D

GENERAL FUNCTIONS

This appendix summarizes general purpose functions, types and other constructs that are not specific to grammars.

D.0.3 List

Lists are defined in §C.6.2, together with the basic functions $\_[\_]$ (element at a particular index), len or length (length), $\langle\rangle$ (empty list).

- $\mathsf{NonemptyList}$ written as A^+

  Polymorphic type of non-empty lists.

  $\mathsf{NonemptyList} := \Lambda \mathsf{A} \,.\, [\, x : \mathsf{A}^* \,|\, \mathsf{len}(x) > 0 \,]$

- $\mathsf{idxs}(l : \mathsf{A}^*) : \mathsf{Set}(\mathbb{N})$

  Set of all natural numbers that can serve as indexes to this list. The set is empty for an empty list.

  $\mathsf{idxs}(l : \mathsf{A}^*) : \mathsf{Set}(\mathbb{N}) :=$
  $\quad \{i : \mathbb{N} \,|\, 0 \leq i < \mathsf{length}(l)\}$

- $\mathsf{tail}(l : \mathsf{A}^+) : \mathsf{A}^*$

  Returns the tail of a list, i.e., the list without the first element.

- $\odot\,(h : \mathsf{A}, t : \mathsf{A}^*) : \mathsf{A}^*$

  This is the equivalent of the [_|_] operator in Prolog, the : operator in Haskell, or the [FIRST h, REST t] structure in HPSG. It creates a list from an element (the head) and another list (the tail).

- $\mathsf{set}(l : \mathsf{A}^*) : \mathsf{Set}(\mathsf{A})$

  Set corresponding to members of a list. Usually implicit, thus we write e.g., $x \in \mathit{list}$ or $\mathit{set} \subseteq \mathit{list}$.

  For example, $\mathsf{set}(\langle 1, 2, 3, 1, 2\rangle) = \{1, 2, 3\}$

  $\mathsf{set}(l : \mathsf{A}^*) : \mathsf{Set}(\mathsf{A}) :=$
  $\quad \lambda x : \mathsf{A} \,.\, \exists i : \mathbb{N} \,.\, l[i] = x$

- $\mathsf{list}(s : \mathsf{Set}(\mathsf{A}), \rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{A}^*$

  List corresponding to a set s ordered by a linear order ρ.

  For example, $\mathsf{list}(\{1, 2, 3\},\ \lambda a, b \,.\, a > b) = \langle 3, 2, 1\rangle$

  $\mathsf{list}(s : \mathsf{Set}(\mathsf{A}), \rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{A}^* :=$
  $\quad \mathsf{sing}\{l : \mathsf{A}^* \mid \mathsf{set}(l) = s \,\&\, \forall i, j \in \mathsf{idxs}(l) \,.\, i < j \Rightarrow \rho(l[i], l[j])$

- $\mathsf{isSuffix}(\mathit{list} : \mathsf{A}^*, \mathit{suf} : \mathsf{A}^*) : \mathsf{Bool}$

  Test whether the list suf is a suffix of the list list.

  $\mathsf{isSuffix}(\mathit{list} : \mathsf{A}^*, \mathit{suff} : \mathsf{A}^*) : \mathsf{Bool} :=$
  $\quad \exists \mathit{pref} : \mathsf{A}^* \,.\, \mathit{pref} \circ \mathit{suff} = \mathit{list}$

- $\mathsf{isSublist}(\mathit{list} : \mathsf{A}^*, \mathit{sub} : \mathsf{A}^*) : \mathsf{Bool}$

  Test whether the list sub is a continuous sublist of the list list.

  $\mathsf{isSublist}(\mathit{list} : \mathsf{A}^*, \mathit{sub} : \mathsf{A}^*) : \mathsf{Bool} :=$
  $\quad \exists \mathit{pref}, \mathit{suff} : \mathsf{A}^* \,.\, \mathit{pref} \circ \mathit{suff} \circ \mathit{suff} = \mathit{list}$

- $\mathsf{filter}(\mathit{list} : \mathsf{A}^*, \varphi : \mathsf{A} \to \mathsf{Bool}) : \mathsf{A}^*$ written as $\langle l | \varphi\rangle$ or $l[\varphi]$

  A function removing all elements of a list list not satisfying predicate φ.

  For example,

  $\langle 1, 2, 3, 4, 1\rangle[\lambda x \,.\, x < 3] = \langle 1, 2, 1\rangle$

  $\mathsf{filter}(\mathit{list} : \mathsf{A}^*, \varphi : \mathsf{A} \to \mathsf{Bool}) : \mathsf{A}^* :=$
  $\quad \mathsf{fold}(\mathit{list}, f, \langle\rangle)$
  $\quad \mathsf{where}\ f(l' : \mathsf{A}^*, x : A) = \mathsf{if}\ \varphi(x)\ \mathsf{then}\ \langle x\rangle \circ l'\ \mathsf{else}\ l'$

- $\mathsf{map}(l : \mathsf{A}^*, f : \mathsf{A} \rightarrow \mathsf{B}) : \mathsf{A}^*$

  A function mapping elements of a list l using a function f.

  For example, $\mathsf{map}(1, 2, 3^*, \lambda x\,.\,2x) = 2, 4, 6^*$; $\mathsf{map}(1, 2, 3^*, \lambda x > 2) = \{\mathsf{false}, \mathsf{false}, \mathsf{true}\}$

  ```
  $\mathsf{map}(\mathit{list} : \mathsf{A}^*, f : \mathsf{A} \rightarrow \mathsf{B}) : \mathsf{A}^* :=$
      $\mathsf{fold}(\mathit{list}, f.\odot, \langle\rangle)$
  ```

- $\mathsf{fold}(\mathit{list} : \mathsf{A}^*, \mathit{plus} : \mathsf{A} \times \mathsf{B} \rightarrow \mathsf{B}, \mathit{zero} : \mathsf{B}) : \mathsf{B}$

  A quite general function for expressing certain recursive operations over lists. In a list recursively viewed as $\mathit{head} \odot \mathit{tail}$, fold replaces $\odot$ by *plus*, and the final empty list by *zero*. Many other functions can be viewed as a special case of fold, including:

  1. sum of a list of numbers:

     $\mathsf{sum} = \lambda \mathit{nrs} : \mathbb{N}^*\,.\,\mathsf{fold}(\mathit{nrs}, +, 0)$

  2. minimum of numbers in a list:

     $\mathsf{minOfAList} = \lambda \mathit{nrs} : \mathbb{N}^*\,.\,\mathsf{fold}(\mathit{nrs}, \min, \infty)$

  It is usually defined recursively in this manner:

  ```
  $\mathsf{fold}(\mathit{list} : \mathsf{A}^*, \mathit{plus} : \mathsf{B} \times \mathsf{A} \rightarrow \mathsf{B}, \mathit{zero} : \mathsf{B}) : \mathsf{B} :=$
      if  $(\mathit{list} = \langle\rangle)$
          then $\mathit{zero}$
          else $\mathit{plus}(\mathit{list}[0], \mathsf{fold}(\mathsf{tail}(\mathit{list}), \mathit{plus}, \mathit{zero}))$
  ```

  Because we did not define general recursion, it must be done using induction instead:

  ```
  $\mathsf{fold}(\mathit{list} : \mathsf{A}^*, \mathit{plus} : \mathsf{A} \times \mathsf{B} \rightarrow \mathsf{B}, \mathit{zero} : \mathsf{B}) : \mathsf{B} :=$
      $\mathsf{ind}(\mathit{zero}, g, r(0))$
      where
          $g(n : \mathbb{N}, t : \mathsf{B}) = \mathit{plus}(\mathit{list}[r(n)], t)$
          $r(n : \mathbb{N}) = \mathsf{len}(\mathit{list}) - n - 1$
  ```

  The primitive recursor function ind is introduced in §C.5. The locally defined function r gives for a normal list index the corresponding index when counting from the end of the list.

- $\mathsf{concatenate}(\mathit{ls} : \mathsf{A}^{**}) : \mathsf{A}^*$

  Takes a list of lists ls and concatenates them all into a single list.

  For example,

  $\mathsf{concatenate}(\langle\langle 1, 2\rangle, \langle\rangle, \langle 3, 4\rangle, \langle 4\rangle\rangle) = \langle 1, 2, 3, 4, 4\rangle$

$\mathsf{concatenate}(\langle\, \langle our, Adam\rangle, \langle feeds\rangle, \langle a, goat\rangle \,\rangle) = \langle our, Adam, feeds, a, goat\rangle$

$\mathsf{concatenate}(ls : \mathsf{A}^{**}) : \mathsf{A}^{*} :=$
$\quad \mathsf{fold}(ls, \circ, \langle\rangle)$

- $\mathsf{orderOf}(l : \mathsf{A}^{*}) : \mathsf{Rel}(\mathsf{A})$ written as $<_l$ operator

  A function returning the linear order corresponding to the ordering within a list. Obviously, this function is undefined for lists with repeating members.

  For example, $5 <_{\langle 7,5,2,1,4\rangle} 1$

  $\mathsf{orderOf}(l : \mathsf{A}^{*}) : \mathsf{Rel}(\mathsf{A}) :=$
  $\quad \lambda x, y : \mathsf{A}\ \exists i, k : \mathbb{N} \,.\, l[i] = x \,\&\, l[i + k + 1] = y$

- $\mathsf{isOrderedBy}(l : \mathsf{A}^{*}, \rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{Bool}$

  Predicate testing if a list is ordered by an order (not necessarily linear):

  $\mathsf{isOrderedBy}(l : \mathsf{A}^{*}, \rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{Bool} := \forall x, y : \mathsf{A}\ \forall i, j \in \mathsf{idxs}(l) \,.\, \rho(x, y) \,\&\, l[i] = x \,\&\, l[j] = y \Rightarrow i < j$

D.0.4 Sets

Sets are modeled as characteristic functions. All the standard functions such as $\emptyset$, $\subseteq$, card (cardinality), ..., are definable in the usual way, see §C.6.1.

- $\mathsf{filter}(s : \mathsf{Set}(\mathsf{A}), \varphi : \mathsf{A} \to \mathsf{Bool}) : \mathsf{Set}(\mathsf{A})$ written as $\{s \,|\, \varphi\}$

- $\mathsf{filter}(s : \mathsf{Set}(\mathsf{A}), \varphi : \mathsf{A} \to \mathsf{Bool}) : \mathsf{Set}(\mathsf{A})$ written as $\{s \,|\, \varphi\}$

  A function (written in the usual set-theoretic notation) filtering a set s with a predicate φ. This is in fact a different notation for set intersection.

  For example,

  $\{\, \{1, 2, 3, 4\} \,|\, \lambda x \,.\, x < 3 \,\} = \{1, 2\};$

  $\langle 1, 2, 3, 4, 1\rangle[\lambda x \,.\, x < 3] = \langle 1, 2, 1\rangle$

  $\mathsf{filter}(s : \mathsf{Set}(\mathsf{A}), \varphi : \mathsf{A} \to \mathsf{Bool}) : \mathsf{Set}(\mathsf{A}) :=$
  $\quad \lambda x : \mathsf{A} \,.\, s(x) \,\&\, \varphi(x)$

- $\mathsf{map}(s : \mathsf{Set}(\mathsf{A}), f : \mathsf{A} \to \mathsf{B}) : \mathsf{Set}(\mathsf{B})$

  A function mapping elements of a set s using a function f.

  For example, $\mathsf{map}(\{1, 2, 3\}, \lambda x \,.\, 2x) = \{2, 4, 6\}$; $\mathsf{map}(\{1, 2, 3\}, \lambda x > 5) = \{\mathsf{false}\}$

$\mathsf{map}(s : \mathsf{Set}(\mathsf{A}), f : \mathsf{A} \to \mathsf{B}) : \mathsf{Set}(\mathsf{B}) :=$

$\quad \lambda y : \mathsf{B} \,.\, \exists x : \mathsf{A} \,.\, y = f(x)$

D.0.5 Relations

Relations are modeled as sets of pairs, or equivalently as predicates on pairs.

- $\mathsf{reflexiveClosure}(\rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{Rel}(\mathsf{A})$

  Reflexive closure of a relation.

  $\mathsf{reflexiveClosure}(\rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{Rel}(\mathsf{A}) :=$

  $\quad \lambda x, y : \mathsf{A} \,.\, \rho(x, y) \vee x = y$

- $\mathsf{transitiveClosure}(\rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{Rel}(\mathsf{A})$

  Transitive closure of a relation.

  $\mathsf{transitiveClosure}(\rho : \mathsf{Rel}(\mathsf{A})) : \mathsf{Rel}(\mathsf{A}) :=$

  $\quad \lambda x, y : \mathsf{A}\ \exists \mathit{list} : \mathsf{A}^{+} \,.\, \mathit{list}[0] = x \,\&\, \mathsf{isOrderedBy}(\mathit{list}, \rho) \,\&\, \mathit{list}[\mathit{list}.\mathsf{len} - 1] = y$

- $\mathsf{mapOrder}(\rho : \mathsf{Rel}(\mathsf{A}), f : \mathsf{A} \to \mathsf{B}) : \mathsf{Rel}(\mathsf{B})$

  Maps order ρ by a function f.

  $\mathsf{mapOrder}(\rho : \mathsf{Rel}(\mathsf{A}), f : \mathsf{A} \to \mathsf{B}) : \mathsf{Rel}(\mathsf{B}) :=$

  $\quad \lambda a, b : \mathsf{B}\ \exists x, y : \mathsf{A} \,.\, f(x) = a \,\&\, f(y) = b \,\&\, \rho(x, y)$

  For example, $\mathsf{mapOrder}(<, -)(3, 1)$

- $\mathsf{passOrder}(\rho : \mathsf{Rel}(\mathsf{B}), f : \mathsf{A} \to \mathsf{B}) : \mathsf{Rel}(\mathsf{A})$

  Dual function to mapOrder.

  $\mathsf{passOrder}(\rho : \mathsf{Rel}(\mathsf{B}), f : \mathsf{A} \to \mathsf{B}) : \mathsf{Rel}(\mathsf{A}) := \lambda x, y : \mathsf{A} \,.\, \rho(f(x), f(y))$

- $\mathsf{respects}(\rho : \mathsf{Rel}(\mathsf{A}), \sigma : \mathsf{Rel}(\mathsf{A})) : \mathsf{Bool}$

  $\mathsf{respects}(\rho : \mathsf{Rel}(\mathsf{A}), \sigma : \mathsf{Rel}(\mathsf{A})) : \mathsf{Bool} :=$

  $\quad \forall x, y : \mathsf{A} \,.\, \sigma(x, y) \Rightarrow \rho(x, y)$

D.0.6 Other

- higher-order polymorphic connectives generalizing the usual connectives ($\&, \vee, \neg$) from boolean values to predicates:

  $\& : (\mathsf{A} \to \mathsf{Bool}) \times (\mathsf{A} \to \mathsf{Bool}) \to (\mathsf{A} \to \mathsf{Bool})$

- higher-order polymorphic operators generalizing the usual set operators $(\cup, \cap)$ to functions yielding sets:

  $\cup : (\mathsf{A} \rightarrow \mathsf{Set}(B)) \times (\mathsf{A} \rightarrow \mathsf{Set}(B)) \rightarrow (\mathsf{A} \rightarrow \mathsf{Set}(B))$

BIBLIOGRAPHY

Ackerman, Farrell, and Gert Webelhuth (1998). *A Theory of Predicates.* Stanford: CSLI Publications.

Anderson, Stephen R. (1992). *A-Morphous Morphology.* Cambridge: Cambridge University Press.

— (1993). “Wackernagel’s Revenge: Clitics, Morphology, and the Syntax of Second Position”. In *Language* 69. 68–98.

— (1994). “Parsing Morphology: “Factoring” Words”. In *Language computations : DIMACS Workshop on Human Language, March 20-22, 1992.* Ed. by Eric Sven Ristad. Vol. 17. Dimacs Series in Discrete Mathematics and Theoretical Computer Science. Providence, RI: American Mathematical Society. 167–183.

Avgustinova, Tania (2000). “Gaining the perspective of language-family-oriented grammar design: predicative special clitics in Slavic”. In *Proceedings of GLiP-1, Workshop on Generative Linguistics in Poland, Warszawa, Poland, 13-14 November 1999.* IPI PAN, Institute of Computer Science, Polish Academy of Sciences. 5–14.

Avgustinova, Tania, and Karel Oliva (1995). The Position of Sentential Clitics in the Czech Clause. 68. CLAUS Report. Universität des Saarlandes.

Avgustinova, Tania, Alla. Bémová, Eva Hajičová, Karel Oliva, Jarmila Panevová, Vladimír Petkevič, Petr Sgall, and Hana Skoumalová (1995). Linguistic problems of Czech. Project Peco 2924. Tech. rep. Prague: Charles University.

Bayer, Joseph (1984). “Comp in Bavarian Syntax”. In *The Linguistic Review* 3. 209–274.

Bell, J. L. (1982). “Categories, Toposes and Sets”. In *Synthese* 51.3. 293–335.

Böhmová, Alena, Jan Hajic, Eva Hajičová, and Barbora Hladká (2001). “The Prague Dependency Treebank: Three-Level Annotation Scenario”. In *Treebanks: Building and Using Syntactically Annotated Corpora.* Ed. by Anne Abeillé. Kluwer Academic Publishers.

Bonet, Eulàlia (1991). “Morphology after Syntax: Pronominal Clitics in Romance”. PhD thesis. MIT.

— (1994). “The Person-Case Constraint: A Morphological Approach”. In *MIT Working Papers in Linguistics* 22. 33–52.

Bresnan, Joan (2001). *Lexical Functional Syntax.* Blackwell.

Brun, Dina (2000). “Discourse structure and definiteness in Russian”. In *Proceedings of ConSOLE 8.* SOLE, Leiden.

Brun, Dina (2001). “Information Structure and the Status of NP in Russian”. In *Theoretical Linguistics* 27.2/3. 109–136.

Cardelli, Luca, and Peter Wegner (1985). “On understanding types, data abstraction, and polymorphism”. In *Computing Surveys* 17.4. 471–522. DOI: http://doi.acm.org/10.1145/6041.6042.

Carpenter, Bob (1999). “The Turing Completeness of Multimodal Categorial Grammars”. In *Papers Presented to Johan van Benthem in honor of his 50th Birthday.* Utrecht: European Summer School in Logic, Language, Information (ESSLLI).

Carstairs, Andrew (1981). *Notes on affixes, clitics and paradigms.* Bloomington, IN: Indiana University Linguistics Club.

Chomsky, Noam (1981). *Lectures on Government and Binding.* Mouton de Gruyter.

Chomsky, Noam, and H. Lasnik (1993). “Principles and Parameters Theory”. In *Syntax: An International Handbook of Contemporary Research.* de Gruyter.

Chung, Chan (1998). “Argument Composition and Long-Distance Scrambling in Korean: An Extension of the Complex Predicates Analysis”. In *Complex Predicates in Nonderivational Syntax.* Ed. by Erhard Hinrichs, Andreas Kathol, and Tsuneko Nakazawa. Vol. 30. Syntax and Semantics. San Diego: Academic Press. 159–220.

Church, Alonzo (1940). “A Formulation of the Simple Theory of Types”. In *The Journal of Symbolic Logic* 5.2 (June 1940). 56–68.

Comrie, Bernard (1981). *Language universals and linguistic typology.* Chicago, IL: The University of Chicago Press.

Crole, Roy L. (1993). *Categories for Types.* Cambridge University Press.

Curry, Haskell B. (1961). “Some Logical Aspects of Grammatical Structure”. In *Structure of Language and Its Mathematical Aspects.* Ed. by Roman Jakobson. 56–68.

Curry, Haskell B., and Robert Feys (1958). *Combinatory Logic.* Vol. 1. Second edition, 1968. North Holland.

Daneš, František (1974). “Functional sentence perspective and the organization of the text”. In *Papers on Functional Sentence Perspective.* Ed. by František Daneš. Praha: Academia. 106–128.

Daneš, František, Miroslav Grepl, and Zdeněk Hlavsa (1987). *Mluvnice češtiny 3 – Skladba [Grammar of Czech 3 – Syntax].* Ed. by Jan Petr. Praha: Academia.

Daniels, Michael (2001). “On a Type-Based Analysis of Feature Neutrality and the Coordination of Unlikes”. In *Proceedings of the 8th International Conference on Head-Driven Phrase Structure Grammar, CSLI Online Proceedings.* Stanford: CSLI Publications. 137–147.

de Groote, Philippe (2001). “Towards Abstract Categorial Grammars”. In *Proceedings of the 39th Annual Meeting of Association for Computational Linguistics and 10th Conference of the European Chapter, Toulouse, France.* 148–155.

— (2002). “Tree-Adjoining Grammars as Abstract Categorial Grammars”. In *TAG+6, Proceedings of the sixth International Workshop on Tree Adjoining Grammars and Related Frameworks.* Università di Venezia. 145–150.

De Kuthy, Kordula (2002). *Discontinuous NPs in German — A Case Study of the Interaction of Syntax, Semantics and Pragmatics*. Studies in Constraint-Based Lexicalism. Stanford: CSLI Publications.

De Kuthy, Kordula, and Walt Detmar Meurers (2001). "On Partial Constituent Fronting in German". In *Journal of Comparative Germanic Linguistics* 3.3. 143–205. URL: http://www.ling.osu.edu/~dm/papers/dekuthy-meurers-jcgl01.html.

Donohue, Cathryn, and Ivan A. Sag (1999). "Domains in Warlpiri". In *Sixth International Conference on HPSG–Abstracts. 04–06 August 1999*. Edinburgh. 101–106.

Dotlačil, Jakub (2006). "Why Clitics Cannot Climb out of CP: A Discourse Approach". In *Proceedings from FASL 15, May 12-14, 2006, Toronto (in press)*.

Dowty, David R. (1996). "Toward a minimalist theory of syntactic structure". In *Discontinuous Constituency*. Ed. by Harry Bunt and Arthur van Horck. Berlin, New York: Mouton de Gruyter. 11–62.

Dowty, David, R. Wall, and S. Peters (1981). *Introduction to Montague Semantics*. Dordrecht: Reidel.

Drach, Erich (1937). *Grundgedanken der deutschen Satzlehre*. 4th edition, Darmstadt: Wissenschaftliche Buchgesellschaft, 1963. Frankfurt: Diesterweg.

Erdmann, Oskar (1886). *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung. Erste Abteilung*. Stuttgart , Germany: Verlag der J. G. Cotta'schen Buchhandlung.

Esvan, François (2000). "Česká klitika z hlediska typologického [Czech clitic from a typological view]". In *Čeština univerzália a specifika 2, Sborník konference ve Šlapanicích u Brna 17.-19.11.1999*. Brno. 141–148.

Filip, Hana (1999). *Aspect, Eventuality Types and Nominal Reference*. New York: Garland Publishing, Taylor & Francis Group, Routledge.

Firbas, Jan (1957). "Some thoughts on the function of word order in old English and modern English". In *Sborník prací Filosofické fakulty Brněnské university. A, Řady jazykovědné [Papers from the Faculty of Arts of the Brno University, A, Linguistic series]*. 5. 72–100.

— (1992). *Functional sentence perspective in written and spoken communication*. Cambridge, England: Cambridge University Press.

Forsberg, Markus (2004). "Applications of Functional Programming in Processing Formal and Natural Languages". MA thesis. Chalmers, Götebork University.

Forsberg, Markus, and Aarne. Ranta (2004). "Functional Morphology". In *Proceedings of the Ninth ACM SIGPLAN International Conference of Functional Programming (ICFP'04), September 19-21, 2004, Snowbird, Utah*.

Franks, Steven, and Tracy Holloway King (2000). *A Handbook of Slavic Clitics*. Oxford University Press.

Fried, Mirjam (1994). "Second-position clitics in Czech: Syntactic or phonological?". In *Lingua* 94. 155–175.

Fronek, Josef (1999). *English-Czech/Czech-English Dictionary.* Contains an overview of Czech grammar. Praha: Leda.

Gabelentz, Georg von der (1891). *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* Leipzig, Germany: T.O. Weigel Nachfolger.

Gallin, D. (1975). *Intensional and Higher Order Modal Logic.* Amsterdam: North Holland.

George, Leland, and Jindřich Toman (1976). "Czech Clitics in Universal Grammar". In *Papers from the 12th Regional Meeting of the Chicago Linguistic Society, Chicago.* Chicago Linguistic Society.

Gordon, M. J. C. (1989). "HOL: A Proof Generating System for Higher-Order Logic". In *Current Trends in Hardware Verification and Automated Theorem Proving.* Ed. by G. Birtwistle and P. A. Subrahmanyam. Springer-Verlag. 73–128.

Hajič, Jan, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alla Bémová (1999). A Manual for Analytic Layer Annotation of the Prague Dependency Treebank (English translation). Tech. rep. ÚFAL MFF UK, Prague, Czech Republic.

Hajičová, Eva, and Jarka Vrbová (1982). "The Role Of The Hierarchy Of Activation In The Process Of Natural Language Understanding". In *Proceedings of the Ninth International Conference on Computational Linguistics, Coling 1982.* North-Holland Publishing Company/Academia.

Hajičová, Eva, Petr Kuboň, and Vladislav Kuboň (1990). "Hierarchy of Salience and Discourse Analysis and Production". In *Proceedings from the 13th International Conference on Computational Linguistics (COLING 1990).* Vol. 1.

— (2004). "Issues of Projectivity in the Prague Dependency Treebank". In *The Prague Bulletin of Mathematical Linguistics* 81.

Halliday, M.A.K. (1967). "Notes on transitivity and theme in English.". In *Journal of Linguistics* 3. 199–244.

Halpern, Aaron (1995). *On the Placement and Morphology of Clitics.* Disertations in Linguistics. Stanford, California: CSLI Publications.

— (1996). In *Approaching Second: Second position clitics and related phenomena.* Ed. by Aaron L. Halpern and Arnold M. Zwicky. CSLI Publications.

— (1998). "Clitics". In. Ed. by Andrew Spencer and Arnold M. Zwicky. Oxford, UK, Malden, MA: Blackwell. 101–122.

Hana, Jirka (2004). "Czech clitics in Higher Order Grammar". In *Working Papers in Slavic Studies.* Columbus, Ohio: Department of Slavic and East European Languages and Literatures. URL: http://ling.osu.edu/~hana/bib.html.

Harkins, William E. (1953). *A modern Czech grammar.* New York: King's Crown Press.

Harris, Alice C. (2002). *Endoclitics and The Origin of Udi Morphosyntax.* Oxford University Press.

Havelka, Jiri (2007). "Beyond Projectivity: Multilingual Evaluation of Constraints and Measures on Non-Projective Structures". In. Paper accepted to ACL 2007.

Hays, David G. (1960). Grouping and dependency theories. Tech. rep. RAND Research.

Hays, David G. (1964). “Dependency Theory: A Formalism and Some Observations”. In *Language* 40.4. 511–525.

Henkin, L. (1950). “Completeness in the theory of types”. In *Journal of Symbolic Logic* 15. 81–91.

Herling, Simon Heinrich Adolf (1821). “Über die Topik der deutschen Sprache”. In *Abhandlungen des frankfurtischen Gelehrtenvereines für deutsche Sprache*. 296–362, 394.

Hickey, Jason (2001). “The MetaPRL logical programming environment”. PhD thesis. Cornell University.

Hickey, Jason, et al. (2003). “MetaPRL - A Modular Logical Environment”. In *Proceedings of the 16th International Conference on Theorem Proving in Higher Order Logics (TPHOLs 2003)*. Vol. 2758. Lecture Notes in Computer Science. Springer-Verlag. 287–303.

Hinrichs, Erhard, and Tsuneko Nakazawa (1994). “Linearizing AUXs in German Verbal Complexes”. In *German in Head-Driven Phrase Structure Grammar*. Ed. by John Nerbonne, Klaus Netter, and Carl J. Pollard. CSLI Lecture Notes 46. Stanford University: CSLI Publications. 11–37.

Höhle, Tilman N. (1986). “Der Begriff ‘Mittelfeld’. Anmerkungen über die Theorie der topologischen Felder”. In *Kontroversen alte und neue. Akten des VII. Internationalen Germanistenkongresses Göttingen 1985*. Ed. by A. Schöne. Tuebingen: Niemeyer. 329–340.

— (1999). “An Architecture for Phonology”. In *Slavic in Head-Driven Phrase Structure Grammar*. Ed. by R. D. Borsley and A. Przepiórkowski. Stanford, CA: CSLI Publications. 61–90.

Holan, Tomáš, Vladislav Kuboň, Karel Oliva, and Martin Plátek (1998). “Two useful measures of word order complexity”. In *Proceedings of COLING-ACL ’98 Workshop “Processing of Dependency-Based Grammars”*. Ed. by Sylvain Kahane and Alain Polguère. Montreal: University of Montreal.

— (2000). “On Complexity of Word Order”. In *Traitement automatique des langues* 41.1. 273–300.

Howard, William A. (1980). “The formulas-as-types notion of construction”. In *To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus, and Formalism*. Ed. by J. P. Seldin and J. R. Hindley. Reprint of 1969 article. Academic Press. 479–490.

Hughes, John (1989). “Why Functional Programming Matters”. In *The Computer Journal* 32.2. 98–107.

Humayoun, Muhammad (2006). “Urdu Morphology, Orthograpgy and Lexicon Extraction”. MA thesis. Chalmers University of Technology, Goteborg University. URL: http://www.lama.univ-savoie.fr/~humayoun/UrduMorph/index.html.

Jackendoff, Ray (1972). *Semantic Interpretation in Generative Grammar*. MIT Press.

Janda, Laura A., and Charles E. Townsend (2002). *Czech*. 2002. URL: http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=2.

Jech, Thomas J (2003). *Set theory*. Berlin ; New York: Springer.

Joshi, A. K., L. S. Levy, and M. Takahashi (1975). “Tree adjunct grammars”. In *Journal Computer Systems Science* 10.1.

Kadmon, Nirit (2001). *Formal Pragmatics.* Blackwell Publishers.

Karlík, Petr, Marek Nekula, and Z. Rusínová (1996). *Příruční mluvnice češtiny [Concise Grammar of Czech].* Praha: Nakladatelství Lidové Noviny.

Kathol, Andreas (1995). "Linearization-Based German Syntax". PhD thesis. The Ohio State University.

— (2000*a*). *Linear Syntax.* Oxford University Press.

— (2000*b*). "Syntactic categories and positional shape alternations". In *Journal of Comparative Germanic Linguistics* 3.2. 59–96.

Kathol, Andreas, and Carl Pollard (1995). "Extraposition via complex domain formation". In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics.* Cambridge, Massachusetts: Association for Computational Linguistics. 174–180.

Kepser, Stephan (2004). "On the Complexity of RSRL". In *Proceedings FG-MOL 2001.* Electronic Notes in Theoretical Computer Science 53. Elsevier Science. 43–52.

Keselj, Vlado (2002). "Modular Stochastic HPSGs for Question Answering". CS-2002-28. PhD thesis. Waterloo, Ontario, Canada: University of Waterloo.

Klavans, Judith L. (1982). *Some problems in a theory of clitics.* Bloomington, IN: Indiana University Linguistics Club.

— (1985). "The Independence of Syntax and Phonology In Cliticization". In *Language* 61.1. 95–120.

— (1995). *On Clitics and Cliticization: The Interaction of Morphology, Phonology, and Syntax.* Outstanding Dissertations in Linguistics. Garland Science.

Králíková, Květa (1981). "Reflexívnost sloves z hlediska automatické analýzy češtiny. [Reflexivity of verbs from the point of automatic analysis of Czech]". In *Slovo a Slovesnost* 42.4. 291–298.

Kupść, Anna (2000). "An HPSG Grammar of Polish Clitics". Ph.D. thesis. Warszawa, Poland: Institute of Computer Science, Polish Academy of Sciences, Université Paris 7. URL: http://www.sfs.uni-tuebingen.de/hpsg/archive/bibliography/papers/kupsc-diss.ps.

Lambek, Joachim (1988). "Categorial and categorical grammars". In *Categorial Grammars and Natural Language Structures.* Ed. by R. Oehrle, E. Bach, and D. Wheeler. Dordercht: Reidel. 297–317.

— (1999). "Deductive systems and categories in linguistics". In *Logic, Language, and Reasoning.* Ed. by H. Ohlbach and U. Reyle. Essays in Honor of Dov Gabbay. Dordercht: Kluwer. 279–294.

Lambek, Joachim, and P. J. Scott (1986). *Introduction to Higher Order Categorical Logic.* Cambridge, Great Britain: Cambridge University Press.

Lenertová, Denisa (2001). "On Clitic placement, topicalization, and CP-structure in Czech". In *Current Issues in Formal Slavic Linguistics.* Ed. by Gerhild Zybatow, Uwe Junghanns, Grit Mehlhorn, and Luka Szucsich. Vol. 5. Linguistik International. Frankfurt am Main, Germany: Lang. 294–305.

Levine, Robert, and Thomas E Hukari (2006). *The Unity of Unbounded Dependency Constructions.* CSLI lecture notes. Stanford, CA: Center for the Study of Language, Information. 406.

Mathesius, Vilém (1915). "O passivu v moderní angličtině. [On Passive in Modern English]". In *Sborník filologický* 5. 198–220.

— (1929). "Zur Satzperspektive im modernen Englisch. [On Sentence Perspective in Modern English]". In *Archiv für das Studium der neueren Sprachen und Literaturen* 155. 202–210.

— (1939). "O tak zvanm aktuálním členění větném [On the so called articulation of the sentence]". In *Slovo a Slovesnost* 5. 171–174. Published in English as (Mathesius 1975).

— (1975). "On Information-Bearing Struture of the Sentence". In *Harvard Studies in Syntax and Semantics.* Ed. by S Kuno. Vol. 1. Originally published in Czech as (Mathesius 1939). 467–480.

Mendelson, Elliott (1997). *An Introduction to Mathematical Logic.* 4th ed. London: Chapman & Hall.

Meurers, Walt Detmar (2005). "On the use of electronic corpora for theoretical linguistics. Case studies from the syntax of German". In *Lingua* 115.11. 1619–1639. http://ling.osu.edu/~dm/papers/meurers-03.html.

Meyer, Roland (2005). "VP-Fronting in Czech and Polish – A Case Study in Corpus-Oriented Grammar Research". In *Heterogeneity in Focus: Creating and Using Linguistic Databases.* Ed. by S. Dipper, M. Götze, and M. Stede. Vol. 2. Interdisciplinary Studies on Information Structure (ISIS). 87–115. URL: http://www.sfb632.uni-potsdam.de/publications/isis02_5meyer.pdf.

Milner, Robin, Mads Tofte, Robert Harper, and David MacQueen (1997). *The Definition of Standard ML (Revised).* Cambridge, MA: MIT Press.

Montague, Richard (1970). "English as a formal language". In *Linguaggi nella Società e nella Tecnica.* Ed. by Bruno Visentini et al. Reprinted in (**montague:1974**, pp. 181-221). Milan, Italy: Edizioni di Comunità. 189–224.

— (1973). "The Proper Treatment of Quantification in Ordinary English". In *Approaches to Natural Language.* Ed. by J. Hintikka, J. Moravcsik, and P. Suppes. Reprinted in (**montague:1974**, pp. 247–270). Dordrecht: Reidel.

Moortgat, Michael (1997). "Categorial Type Logics". In *Handbook of Logic and Language.* Ed. by Johan van Benthem and Alice ter Meulen. Elsevier. 93–177.

Morrill, Glyn V. (1994). *Type Logical Grammar: Categorial Logic of Signs.* Dordrecht: Kluwer.

Müller, Stefan (2002). "Multiple Frontings in German". In *Proceedings of Formal Grammar 2002.* Trento. 113–124. URL: http://www.cl.uni-bremen.de/~stefan/Pub/mehr-vf.html.

— (2003). "Mehrfache Vorfeldbesetzung". In *Deutsche Sprache* 31.1. 29–62. URL: http://www.cl.uni-bremen.de/~stefan/Pub/mehr-vf-ds.html.

— (2005). "Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung". In *Linguistische Berichte* 203. 297–330. URL: http://www.cl.uni-bremen.de/~stefan/Pub/mehr-vf-lb.html.

Muskens, Reinhard (2001*a*). “Categorial Grammar and Lexical-Functional Grammar”. In *Proceedings of the LFG01 Conference, University of Hong Kong*. Stanford, CA: CSLI Publications. 259–279.

— (2001*b*). “Lambda Grammars and the Syntax-Semantics Interface”. In *Proceedings of the Thirteenth Amsterdam Colloquium*. Amsterdam. 150–155.

— (2003). “Language, Lambdas, and Logic”. In *Resource Sensitivity in Binding and Anaphora*. Ed. by Geert-Jan Kruijff and Richard Oehrle. Studies in Linguistics and Philosophy. Kluwer. 23–54.

— (2004). “Separating Syntax and Combinatorics in Categorial Grammar”. In *Proceedings of the International Conference on Categorial Grammars (CG2004)*. Montpellier, France.

Naughton, James (2005). *Czech: An Essential Grammar*. Oxon, Great Britain, New York, NY, USA: Routledge. 288.

Nevis, J. A., B. D. Joseph, D. Wanner, and A. M. Zwicky (1994). *Clitics: A Comprehensive Bibliography, 1892-1991*. John Benjamins.

Oliva, Karel (1998). “Just Czech clitic data, or a closer look at the ”Position paper: Clitics in Slavic”. In *Presented at the Comparative Slavic Morphosyntax Workshop, Bloomington, IN, June 1998*.

Panevová, Jarmila (1980). *Formy a funkce ve stavbě ceské věty [Forms and functions in the structure of the Czech sentence]*. Prague, Czech Republic: Academia.

— (1994). “Valency frames and the meaning of the sentence”. In *The Prague School of structural and functional linguistics. A short introduction*. Ed. by P. A. Luelsdorff. Amsterdam – Philadelphia: John Benjamins. 223–243.

— (1999). “Česká reciproční zájmena a slovesná valence [Czech reciprocal pronouns and verbal valency]”. In *Slovo a slovesnost* 60. 269–275.

Penn, Gerald (1999*a*). “A Generalized-Domain-Based Approach to Serbo-Croatian Second-Position Clitic Placement”. In *Constraints and Resources in Natural Language Syntax and Semantics*. Ed. by Gosse Bouma, Erhard Hinrichs, Geert-Jan M. Kruijff, and Richard Oehrle. Stanford: CSLI Publications. 119–136.

— (1999*b*). “An RSRL Formalization of Serbo-Croatian Second Position Clitic Placement”. In *Tübingen Studies in Head-Driven Phrase Structure Grammar*. Ed. by Valia Kordoni. 132. Universitaet Stuttgart, Universitaet Tuebingen. 177–197. URL: http://www.sfs.uni-tuebingen.de/sfb/reports/berichte/132/penn/penn.dvi.ps.

Pesetsky, David M. (1982). “Paths and Categories”. PhD thesis. Cambridge, Mass: MIT.

Peters, P. Stanley, and Robert W. Ritchie (1973). “On the Generative Power of Transformational Grammars”. In *Information Sciences* 6. 49–83.

Petkevič, Vladimír. “Neprojektivní konstrukce [Nonprojective constructions]”. An unpublished list of various Czech sentences involving non-projectivities (discontinuities).

— (1998). “Special Cases of Non-Projective Constructions in the Structure of Czech Sentence.”. In *České přednášky pro XII. mezinárodní sjezd slavistů Krakov 1998 [Czech talks at the 12th International Slavistic Congress in Krakow 1998]*. Slovansk stav. 61–66.

Petr, Jan (1987). *Mluvnice češtiny.* Praha: Academia.

Plátek, Martin, Tomáš Holan, Vladimír Kubon, and Karel Oliva (2001). "Word-order relaxations and restrictions within a dependency grammar". In *Proceedings of the 7th International Workshop on Parsing Technologices (IWPT).* Beijing: Tsinghua University Press. 237–240.

Pollard, Carl J., and Ivan A. Sag (1994). *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press.

Pollard, Carl (2001*a*). "Cleaning the HPSG Garage: Some Problems and some proposals". Summer School on Constraint-based Grammar. Trondheim, Norway. 2001. URL: http://www.ling.hf.ntnu.no/HPSG2001/summer_school/cp_garage.ps.

— (2001*b*). "Higher-order grammar: a categorical foundation for type-logical constraint-based grammar". In *Proceedings of Formal Grammar.* Helsinki.

— (2004*a*). "Higher-Order Categorial Grammar". In *Proceedings of the International Conference on Categorial Grammars (CG2004).* Montpellier, France.

— (2004*b*). "Higher Order Grammar. NASSLLI04 course materials". 2004.

— (2004*c*). "Type-Logical HPSG". In *Proceedings of Formal Grammar 2004 (Nancy).* 107–124.

— (2005). "Hyperintensional semantics in a higher order logic with definable subtypes". In *Proceedings of the Second Workshop on Lambda Calculus, Type Theory, And Natural Language, King's College London.* 32–45.

— (2006). "Higher Order Grammar: A Tutorial". Dec. 2006.

— (to appear). "Hyperintensions". In *Journal of Logic and Computation.*

Pollard, Carl, and Jiri Hana (2003). "Ambiguity, neutrality, and coordination in higher order grammar". In *Proceedings of Formal Grammar.* Vienna. 125–136.

Ranta, Aarne (2004). "Grammatical Framework: A Type-Theoretical Grammar Formalism". In *The Journal of Functional Programming* 14.2. 145–189.

Reape, Mike (1994). "Domain Union and Word Order Variation in German". In *German in Head-Driven Phrase Structure Grammar.* Ed. by John Nerbonne, Klaus Netter, and Carl J. Pollard. CSLI Lecture Notes 46. Stanford University: CSLI Publications. 151–197.

— (1996). "Getting things in order". In *Discontinuous Constituency.* Ed. by Harry Bunt and Arthur van Horck. Natural language processing 6. Berlin, New York: Mouton de Gruyter. 209–253.

Rejzek, Jiří (2001). *Český etymologický slovník [Czech etymological dictionary].* Leda.

Rezac, Milan (2005). "The syntax of clitic climbing in Czech". In *Clitic and Affix Combinations: Theoretical perspectives.* Ed. by Lorie Heggie and Francisco Ordónez. Linguistik Aktuell/Linguistics Today 74. 103140. URL: http://minimalism.linguistics.arizona.edu/AMSA/PDF/AMSA-202-0602.pdf.

Richter, Frank (2000). "A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar". Version of 2004. Phil. Dissertation. Eberhard-Karls-Universität Tübingen.

Rivero, María Luisa (2005). “Topics in Bulgarian morphology and syntax: a minimalist perspective”. In *Lingua* 115.8 (Aug. 2005). 1083–1128. DOI: 10.1016/j.lingua.2004.02.006.

Roberts, Craige (1996). “Information Structure: Towards an integrated formal theory of pragmatics”. In *Papers in Semantics*. Ed. by Jae Hak Yoon and Andreas Kathol. Vol. 49. OSU Working Papers in Linguistics.

— (1998). “Focus, Information Flow, and Universal Grammar”. In *The Limits of Syntax*. Ed. by Peter W. Culicover and Louise McNally. Vol. 29. Syntax and Semantics. San Diego, CA: Academic Press. 109–160.

Rosen, Alexandr (1994). “Grammar Formalisms and the Description of Word Order Variations”. Unpublished manuscript. 1994. URL: http://utkl.ff.cuni.cz/~rosen/public/worep.ps.

— (2001). “A constraint-based approach to dependency syntax applied to some issues of Czech word order”. PhD thesis. Prague: Charles University. URL: http://utkl.ff.cuni.cz/~rosen/public/THESIS/.

Selkirk, Elisabeth O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, Mass: MIT Press.

Sgall, Petr, Ladislav Nebeský, Alla Goralčíková, and Eva Hajičová (1969). *A Functional Approach to Syntax in Generative Description of Language*. New York: American Elsevier Publishing Company.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Prague, Czech Republic/Dordrecht, Netherlands: Academia/Reidel Publishing Company.

Sgall, Petr, O. E. Pfeiffer, W. U. Dressler, and M. Půček (1995). “Experimental Research on Systemic Ordering”. In *Theoretical Linguistics* 21. 197–239.

Short, David (1993). “Czech”. In *The Slavonic Languages*. Ed. by Bernard Comrie and Grevilled G. Corbett. Routledge Language Family Descriptions. Routledge. 455–532.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, J. Pierrehumbert, J. Hirschberg, and P. Price (1992). “TOBI: A Standard Scheme for Labeling Prosody”. In *Proceedings of the International Conference on Spoken Language 92, Banff, Oct 12-16 1992*.

Søgaard, Anders (2007). “Propositional and first order verification of linguistic structures”. In *Proceedings of the 2nd International Workshop on Typed Feature Structure Grammars*. Tartu, Estonia.

Spencer, Andrew (1991). *Morphological Theory: An Introduction to Word Structure in Generative Grammar*. Blackwell Textbooks in Linguistics. Blackwell.

Steedman, Mark (1991). “Structure and Intonation”. In *Language* 67.2. 260–296.

— (2000*a*). “Information Structure and the Syntax-Phonology Interface”. In *Linguistic Inquiry* 31.4. 649–685.

— (2000*b*). *The Syntactic Process*. Cambridge, MA: MIT Press.

Steedman, Mark, and Jason Baldridge (2003). "Combinatory Categorial Grammar". Unpublished tutorial. Draft 5.0, April 19, 2007. 2003.

Stemberger, Joseph Paul (1981). "Morphological Haplology". In *Language* 57.4 (Dec. 1981). 791–817.

Svoboda, Aleš (2000). "Klitika z hlediska funkční větné perspektivy [Clitic in the point of view of Functional Sentence Perspective]". In *Proceedings of the conference at Šlapanice u Brna, 17-19 November 1999*. Brno: Masarykova univerzita. 149–159.

Svoboda, Karel (1969). "Poznámky k problematice doplňku". In *Slovo a slovesnost* 30. 309–320.

Szudzik, Matthew (2005). "Recursive Function". In *MathWorld–A Wolfram Web Resource*. Ed. by Eric W. Weisstein. Version of September 19, 2005. URL: http://mathworld.wolfram.com/RecursiveFunction.html.

Thompson, Simon (1997). "Higher-order + Polymorphic = Reusable". May 1997. URL: http://www.cs.kent.ac.uk/pubs/1997/224.

Thorpe, Alana Irene (1991). "Clitic placement in complex sentences in Czech". PhD thesis. United States – Rhode Island: Brown University.

Toman, Jindřich (1980). "Weak and Strong: Notes on be in Czech". In *Wege zur Universalien Forschung*. Ed. by G. Brettschneider and Lehmann. Tubingen: Gynter Narr Verlag.

— (1986). "Cliticization from NPs in Czech and compareable phonomena in French and Italian.". In *Syntax and Semantics 19: The Syntax of Pronominal Clitics*. Ed. by H. Borer. New York: Academic Press.

— (1996). "A note on clitics and prosody". In *Approaching second. Second position clitics and related phenomena*. Ed. by Aaron L. Halpern and Arnold M. Zwicky. CSLI Lecture Notes 61. Stanford, California: CSLI Publications. 505–510.

— (2000). "Prosodické spekulace o klitikách v nekanonických pozicích [Prosodic speculations about clitics in non-canonical positions]". In *Čeština univerzália a specifika 2*. Brno.

Trávníček, František (1951). *Mluvnice spisovné češtiny [Grammar of Literary Czech]*. Slovanské nakladatelství.

— (1962). *Historická mluvnice česká III – Skladba [Historic Grammar of Czech III – Syntax]*. Praha: SPN.

Uhlířová, Ludmila (1972). "On the Non-Projective Constructions in Czech". In *Prague Studies in Mathematical Linguistics*. 171–181.

— (1987). *Knížka o slovosledu [A book about word-order]*. Academia.

Úličný, Oldřich (1969). "K syntaktické a transformační charakteristice doplňku". In *Slovo a Slovesnost* 30. 11–22.

— (1970). "Ještě k pojetí doplňku". In *Slovo a Slovesnost* 31. 271–278.

Vallduví, Enric (1993). "The Informational Component". PhD.. University of Pennsylvania.

Vallduví, Enric, and Maria Vilkuna (1998). “On rheme and kontrast”. In *The Limits of Syntax*. Ed. by Peter W. Culicover and Louise McNally. Vol. 29. Syntax and Semantics. San Diego, CA: Academic Press. 79–108.

Veselá, Kateřina, Nino Peterek, and Eva Hajičová (2003). “Topic-Focus Articulation in PDT: Prosodic Characteristics of Contrastive Topic”. In *The Prague Bulletin of Mathematical Linguistics* 79-80.

Veselovská, Ludmila (1995). “Phrasal Movement and X-Morphology: Word Order Parallels in Czech and English Nominal and Verbal Projections”. PhD thesis. Olomouc, Czechia: Palacký University.

Večerka, Radoslav (1989). *Altkirchenslavische (altbulgarische) Syntax I. Die lieneare Satzorganisation.* Freiburg i. Br.: U.W. Weiher.

Šmilauer, Vladimír (1947). *Novočeská skladba [Syntax of Contemporary Czech].* 3rd edition in 1969 by SPN: Prague. Praha: Ing. Mikuta.

Štícha, F. (1996). “Křížení vět v češtině”. In *Naše řč* 1. 26–31.

Wackernagel, Jacob (1892). “Über ein Gesetz der indogermanischen Wortstellung”. In *Indogermanische Forschungen* 1. 333–436.

Weil, Henri (1844). *De l’ordre des mots dans les langues anciennes comparees aux langues modernes : question de grammaire general.* All citations based on the English translation (Weil 1887 [1844]). Paris.

— (1887 [1844]). *The order of words in the ancient languages compared with that of the modern languages.* Boston: Ginn & Company.

Yatabe, Shûichi (1996). “Long-distance scrambling via partial compaction”. In *Formal Approaches to Japanese Linguistics 2 (MIT Working Papers in Linguistics 29).* Ed. by Masatoshi Koizumi, Masayuki Oishi, and Uli Sauerland. Cambridge, Massachusetts. 303–317.

Zeman, Daniel (2004). Neprojektivita v Pražském závislostním korpusu (PDT) [Nonprojectivity in Prague Dependency Treebank (PDT)]. TR-2004-22. Tech. rep. Prague: ÚFAL/CKL, Charles University.

Zikánová, Šárka, Miroslav Týnovský, and Jiří Havelka (2007). “Identification of Topic and Focus in Czech: EvaluationofManualParallel Annotations”. In *The Prague Bulletin of Mathematical Linguistics.* in press.

Zlatić, Larisa (1997). “The Structure of the Serbian Noun Phrase”. PhD thesis. University of Texas at Austin. URL: http://www.lztranslation.com/zlatic_publications.html.

— (to appear). “Slavic Noun Phrases are NPs not DPs”. In *Comparative Slavic Morphosyntax.* Ed. by George Fowler. Paper presented at the Workshop on Comparative Slavic Morphosyntax, Bloomington, Indiana, June 6 1998. Slavica Publishers. URL: http://www.lztranslation.com/zlatic_publications.html.

Zwicky, Arnold M. (1977). On Clitics. Tech. rep. Reproduced by the Indiana University Linguistics Club. Ohio State University.

— (1985). “Clitics and Particles”. In *Language* 61.2. 283–305.

Zwicky, Arnold M., and Geoffrey K. Pullum (1983). “Cliticization vs. Inflection: English N’T”. In *Language* 59.3. 502–513.

INDEX OF CITATIONS