Univerzita Karlova v Praze
Filozofická fakulta
Ústav teoretické a komputační lingvistiky

Alexandr Rosen

A constraint-based approach to dependency syntax applied to some issues of Czech word order

Extended abstract of the thesis

filologie – matematická lingvistika

vedoucí práce a předseda oborové rady
– prof. PhDr. Petr Sgall, DrSc.

2001

1 Introduction

The thesis has two main objectives. The first is to integrate two different components into a coherent linguistic framework: *Functional Generative Description* (FGD), a dependency-based linguistic theory rooted in the Prague School tradition, and *Relational Speciate Re-entrant Logic* (RSRL), a formal language suitable for constructing constraint-based grammars. The second aim is to show that this combination may be used to describe surface word order as conditioned by several factors, originating at various levels of the language system.

In more concrete terms, the goal is to provide a declarative, constraint-based account of a number of Czech word order phenomena using FGD as the theoretical foundation. Both surface-level constraints and *deep word order*, a concept reflecting functional sentence perspective, can be expected to receive adequate treatment in such an account.

An adequate description of the interaction between the underlying syntactic structure and its surface realization is an important goal of theoretical linguistic research. As the starting point, the main premises of FGD have been adopted, namely the structuring of language description into levels, the distinction between the system of language and its semantic and pragmatic interpretations, the relevance of communicative dynamism and topic-focus articulation for *linguistic meaning* – a notion corresponding to the level of underlying syntax, where the structure of a sentence is presented in the shape of a dependency tree with annotated content words as nodes (Sgall, Hajičová, and Panevová, 1986).

According to the commonly used formalization of the theory, a language is described in by a grammar generating the underlying structure, and transducing components, including movement rules, providing the interface between formally distinct levels of description.

This stratificational specification with some procedural aspects has its drawbacks: it is biased towards one direction of processing (generation) and does not allow simultaneous access to information at all levels of description or interpretation of partial expressions. It has been replaced for the purpose of this thesis by a declarative formulation, allowing for parallelism in representing and describing sentences, clauses, constituents and words at different language levels. For this purpose it is useful to adopt a formal language assumed in constraint-based theories such as HPSG – see Pollard and Sag (1987) and (1994). This kind of formal language comes with a proper definition of its syntax and semantics, using as its main descriptive device a system of types, ordered within an inheritance hierarchy and supplemented by attribute-value pairs with the possibility of value sharing. A grammar formalized in this way constrains typed feature structures, which serve as models of events or objects in the linguistic reality.

The other challenge comes from the empirical domain: word order in Czech is a topic that still poses unanswered questions. According to Vilém Mathesius (Mathesius, 1975), factors of various kind are responsible for word order not only in Czech, but also in English (and probably in human languages generally). The differences between languages with the so-called free and fixed word order are due to the relative weight of these factors. This view is compatible with a constraint-based formalism, and is very close to the view of FGD, where discourse-related factors interact with surface-level regularities in determining word order and prosody.

Thus, my claim is that by adopting a constraint-based formalism for FGD, word order (and prosodical) phenomena (at least) in Czech can be solved more easily than in the stratificational approach with procedural aspects.

Because the task involves marrying two different linguistic traditions, a substantial amount of introductory reading is provided, with the aim to build a solid basis for the subsequent grammar writing exercise. The latter task consists of laying the necessary compositional foundations and then specifying how deep word order is reflected in the surface string (and the other way round) and how it interacts with surface-level constraints. This task can be solved on a rather abstract level, given the wealth of FGD research results concerning the issues of functional sentence perspective and related topics. It is more difficult to find a firm empirical ground and an adequate formal analysis for surface-level phenomena.

In addition to capturing the relation between surface ordering and the underlying representation, I have attempted to cover to a various degree of detail several other phenomena governed by surface ordering constraints: the regular continuous cases, discontinuous comparison constructions, some cases of long-distance dependencies, split prepositional phrases, and the position of function words. The difficult class of clausal clitics has received a more detailed treatment.

2 Organization

The thesis consists of eight chapters and a four-parts appendix. The first three chapters provide introduction and background: after explaining what kind of problems this thesis attempts to solve and why (Chapter 1), assumptions about the choice of linguistic theory and formal metalanguages are discussed and a comparison with a few previous dependency-based approaches to word order issues is provided (Chapter 2). Next, the theory and formalism underlying the system are presented in detail (Chapter 3).

The following three chapters are the core of the thesis. Chapter 4 presents theoretical and empirical issues which will be addressed in the formal description: relevant features of FGD are specified in detail, and a few notes concerning some linguistic concepts are made: the relation between linguistic meaning and semantics, the difference between governor and head, and possibilities for solving coordination and apposition. The most extensive note deals with word order: starting from the word order principles of Vilém Mathesius, a classification of word order phenomena is developed, which is employed later in the proposed formal account. Next, the key components of the formal description are presented: objects modelling the underlying (tectogrammatical) level and the surface (morphemic) level. Briefly explained is the role and the organization of lexicon (Chapter 5). After that, the stage is ready for the formal definitions. This involves (i) general principles governing the setup of linguistic objects, (ii) principles constraining representations of tectogrammatical and morphemic levels and the relation between them, and, finally, (iii) a number of surface-level constraints specific to Czech (Chapter 6).

In Chapter 7, the framework is used to describe the ordering behaviour of Czech clausal clitics with a number of examples, and in the final chapter the whole enterprise is evaluated and some perspectives in this line of research are sketched.

In the Appendix, the formal descriptions are presented again for reference, together with definitions of all relations used and a list of tectogrammatical functions.

3 Theory and its formalization

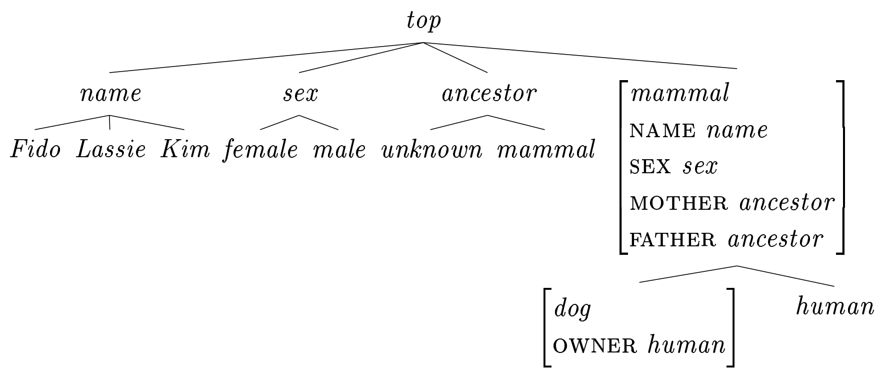
Taking the requirement of descriptive adequacy as the starting point, there are a number of arguments in favour of dependency-based underlying syntax as the level which is (i) sufficiently abstract and thus devoid of surface phenomena such as agreement, or the form and position of function words, and (ii) still firmly belonging to the language system. A representation of this sort is proposed by FGD (Sgall, Hajičová, and Panevová, 1986), (Sgall, 1992). In addition to the elementary relations of dependency between content words in a sentence, for every dependent the representation also specifies an underlying function (comparable to θ-roles). Furthermore, the representation expresses FSP of the sentence: all nodes in the dependency tree are distinguished as being either *contextually bound* or *contextually non-bound*. The nodes are ordered according to the hierarchy of communicative dynamism, where the most dynamic items (usually carrying new information) come last.

The level of underlying syntax is just one of the levels in FGD. The relation between this level and the string of letters or sounds is usually described in a procedural way by means of transducing components and movement rules operating successively from a higher to a lower level of description. This approach does not support easy interpretation of the description during analysis of the surface string and makes descriptive parallelism and partial interpretation difficult. For these reasons, the relation between the level of underlying syntax and the surface string is described in this thesis in a declarative way. Syntactic units are modelled as typed feature structures, defined by a system of constraints – the grammar and the lexicon. In this setup, every syntactic unit is modelled and described as a single object with several dimensions, corresponding to description levels.

The adopted formalism (Richter, 2000) was originally developed for HPSG, but its properties do not contradict any premises of FGD – indeed, RSRL is a powerful tool, which can be used in many contexts outside its initial destination. In fact, it may be accused of being not restrictive enough, as evidenced by its computational properties. One possible answer to this objection is that it is the theory which should restrict possible options, not the formalism. Implementational issues must then be viewed as a separate problem.

An RSRL grammar consists of two parts: *signature* and *theory*. *Signature* defines what kinds of objects are being described as parts of the model. These objects are typed feature structures interpreted as representating sets of linguistic events and their properties in the real world. For simplicity, I will give a non-linguistic example.

(1)



The signature in (1) defines a hierarchy of types in a model of a simple world of dogs and their masters. Some types have attributes whose appropriate values are again types. A type inherits all attributes and appropriate values from its supertype. Feature structures in the model may include only the lowest, maximally specific types with all appropriate attributes present and with values of these attributes set again to maximally specific types.

The feature structure in (2) satisfies the signature by modelling a dog Fido who has the same owner (Kim) as its mother (the boxed number $\boxed{1}$ coindexes identical parts of the structure).

(2) $$\left[\begin{array}{ll} \textit{dog} & \\ \text{NAME} & \textit{Fido} \\ \text{SEX} & \textit{male} \\ \text{MOTHER} & \left[\begin{array}{ll} \textit{dog} & \\ \text{NAME} & \textit{Lassie} \\ \text{SEX} & \textit{male} \\ \text{MOTHER} & \textit{unknown} \\ \text{FATHER} & \textit{unknown} \\ \text{OWNER} & \boxed{1} \end{array}\right] \\ \text{FATHER} & \textit{unknown} \\ \text{OWNER} & \boxed{1}\left[\begin{array}{ll} \textit{human} & \\ \text{NAME} & \textit{Kim} \\ \text{SEX} & \textit{female} \\ \text{MOTHER} & \textit{unknown} \\ \text{FATHER} & \textit{unknown} \end{array}\right] \end{array}\right]$$

(3) $$[\text{MOTHER} \mid \text{SEX } \boxed{1}] \rightarrow \boxed{1}\ \textit{female}$$

(4) $$\left[\begin{array}{ll} \textit{human} & \\ \text{MOTHER} & \boxed{1}\ \textit{mammal} \end{array}\right] \rightarrow \boxed{1}\ \textit{human}$$

(5) $$[\text{OWNER } \boxed{1}] \rightarrow [\text{MOTHER} \mid \text{OWNER } \boxed{1}]$$

The feature structure also satisfies the whole 'grammar': the *theory* is empty, so there are no further constraints. Note that the signature allows some objects which should better be ruled out, such as a male mother (2). This can be remedied by the implication in (3), which is the first statement of the *theory*, one of its *descriptions*. Each description must be satisfied by all objects in the model, so (3) means: for every object with the attribute MOTHER (i.e., a *mammal* type object) and the value of this attribute set to an object with the attribute SEX (again, this can only be a *mammal* type object), the value of SEX must be *female*.
A similar description can be used to exclude a dog as the mother of a human (4). Finally, (5) asserts that each dog owner also owns the dog's mother.

The formal language of RSRL also includes the usual logical connectives, variables, negation, and the possibility to express quantification over components of an object.

4 Factors determining word order

Vilém Mathesius proposed the idea of interacting and mutually competing word order principles. These principles are universal, but have different roles in different languages. For every language, a partial order of these principles can be specified. This order predicts which principles win in case several principles compete for different word orders.

FGD views one of the principles, namely the *functional sentence perspective* (FSP) principle, as primary. A specific FSP can be expressed in various ways: word order, stress pattern, syntactic constructions. FSP of a specific utterance is represented together with other aspects of linguistic meaning at the tectogrammatical level. If grammar consists of (i) a set of well-formedness constraints on possible tectogrammatical representations,

(ii) a set of well-formedness constraints on possible surface strings, and (iii) a set of constraints on correspondences between the two, the latter represents the crucial part which mediates between the underlying representation of FSP and its surface realization, as conditioned by other factors. FSP is manifested in the surface expression wherever possible, i.e., unless defeated by another constraint. Since the formalism allows to use all information in parallel, the various kinds of factors can interact as required. This reasoning results in the following preliminary statements concerning word order:[1]

a. The relative SWO of every two content words corresponds to DWO of the corresponding semantemes, unless any of the cases in the list of Special SWO conditions apply (see below).
b. A function word is ordered adjacently to its host, their order being determined by a syntactic constraint, unless any of the cases in the list of Special SWO conditions apply (see below).
c. For the relative SWO of every two function words in a single ordering domain the list of Special SWO conditions applies.

Special SWO conditions:

a. The word ordered first in SWO is the intonation centre of the utterance and corresponds to focus proper.
b. A syntactic constraint requires otherwise.
c. A stress pattern requires otherwise.
d. A word is ordered first in a SWO domain, the domain is larger than that of its corresponding tectogrammatical subtree, and the word corresponds to topic proper or to contrastive topic.

If two or more conditions compete for different orders, language-specific priorities are applied with the possibility of multiple outcomes. With all of the special SWO conditions, specific constraints on locality must be satisfied.

5 The architecture

The issue of how to describe the relation between surface string and tectogrammatical representation in a compositional way is decided by adopting a flat derivation structure with function words standing as sisters to dependents and the head (except for cases where recursive hosting of function words by other function words is appropriate, as in analytical verbal morphology). An expression corresponding to tectogrammatical node or subtree is represented as a feature structure of type *sign*. Its setup is shown schematically in (6).

[1] SWO: surface word order, DWO: deep word order.

(6)
$$\begin{bmatrix} \textit{sign} & \\ \text{PHONOLOGY} & \textit{list(phonstring)} \\ \text{SYNSEM} & \begin{bmatrix} \textit{synsem} & \\ \text{LOCAL} & \begin{bmatrix} \textit{local} & \\ \text{CATEGORY} & \begin{bmatrix} \textit{category} & \\ \text{HEAD} & \textit{head} \\ \text{VALENCY} & \textit{list(synsem)} \\ \ldots & \end{bmatrix} \\ \text{DEEP} & \begin{bmatrix} \textit{deep} & \\ \text{STATUS} & \textit{status} \\ \text{TREE} & \textit{d-list} \end{bmatrix} \\ \text{CONTENT} & \textit{content} \end{bmatrix} \\ \text{NONLOCAL} & \textit{nonlocal} \end{bmatrix} \\ \text{SURFACE} & \textit{s-list} \end{bmatrix}$$

The structure immediately resembles the type *sign* in HPSG, in order to allow easy adoption of solutions available in that theory and adequate within the context of FGD. However, nothing substantial hinges on this similarity.

The type *deep* represents the level of tectogrammatics and *surface* the level morphemics (string of objects representing morphemes, ordered according to the surface word order). The type *status* has two subtypes: *embedded* and *unembedded*, the latter includes attributes relevant to the utterance as a whole. Additionally, there are parts expressing the syntactic combinatory potential of the expression: *category* and – optionally – its semantic interpretation (*content*).

The type *sign* has also two subtypes: *lexical* and *non-lexical* (similar to the HPSG's sorts *word* and *phrase*). The *non-lexical* type has two additional attributes, which record immediate syntactic components of the expression, mimicking the local derivation tree: a *sign*-valued attribute HEAD-DAUGHTER and a *list(sign)*-valued attribute NONHEAD-DAUGHTERS. The actual string of phonemes (or – for the present purpose – graphemes) of the expression is represented as the value of the attribute PHONOLOGY.

Tectogrammatical representation is formalized as a recursive structure, a list (*d-list*) consisting of a non-list structure (*d-node*) representing the governing node and other lists of the same kind (*d-list*), representing dependent subtrees.

(7) Máňa šla tancovat
Máňa went to dance

The sentence (7) is represented as a tree in Fig. 1 and receives the linearized tectogrammatical representation schematically represented as in (8). The angle brackets enclose a list.

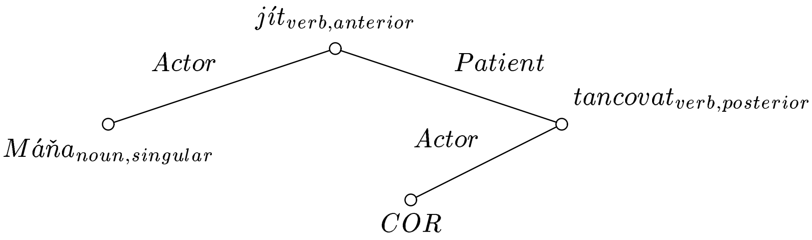


Figure 1: A tectogrammatical tree

(8) $\langle\ \langle[\mathit{Máňa}]\rangle,\ [\mathit{jít}],\ \langle\ \langle[\mathit{COR}]\rangle,\ [\mathit{tancovat}]\ \rangle\ \rangle$

The representation of tectogrammatical level by means of embedded lists relies on the fact that tectogrammatical trees are projective. In (9) the content of the nodes is shown. The nodes still include only the basic information: tectogrammatical function, the binary-valued property of contextual boundness, tectogrammatical word class, lemma and one of the grammatemes appropriate to the word class.

(9)
$$\left\langle \begin{array}{l} \left\langle \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{actor} \\ \text{CB} & \textit{yes} \\ \text{CORE} & \boxed{1} \begin{bmatrix} \textit{d-noun} & \\ \text{LEMMA} & \textit{Máňa} \\ \text{D-NUMBER} & \textit{sg} \end{bmatrix} \end{bmatrix} \right\rangle, \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{fun} \\ \text{CB} & \textit{no} \\ \text{CORE} & \begin{bmatrix} \textit{d-verb} & \\ \text{LEMMA} & \textit{jít} \\ \text{D-TENSE} & \textit{anterior} \end{bmatrix} \end{bmatrix}, \\ \left\langle \left\langle \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{actor} \\ \text{CB} & \textit{yes} \\ \text{CORE} & \boxed{1} \end{bmatrix} \right\rangle, \begin{bmatrix} \textit{d-node} & \\ \text{FUN} & \textit{patient} \\ \text{CB} & \textit{no} \\ \text{CORE} & \begin{bmatrix} \textit{d-verb} & \\ \text{LEMMA} & \textit{tancovat} \\ \text{TENSE} & \textit{posterior} \end{bmatrix} \end{bmatrix} \right\rangle \end{array} \right\rangle$$

Surface string of words is formalized as a simple list (*s-list*). Its setup and the setup of its members is based on *domain lists* of (Penn, 1999a), where embeddable *topological fields* and structure sharing (represented as coindexation) are used to determine position and adjacency of list members (*domain objects*). In (11) as the representation of the surface string of (10), such information is encoded within structures corresponding to the individual words (*s-node*s) as values of the attribute F(IELD) and R(EGION). The attribute I-C stands for *information centre* and represents a very basic account of prosody.

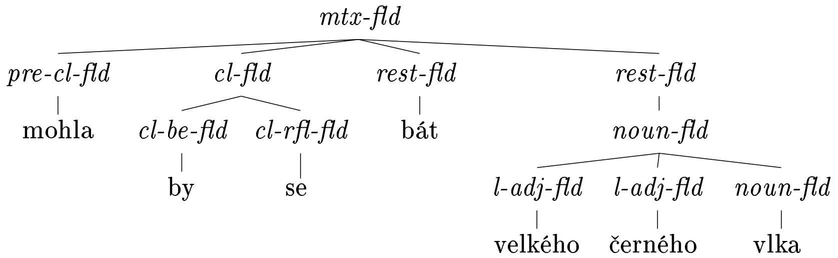
(10) Mohla by se bát velkého černého vlka
could AUX-COND REFL be afraid of big black wolf
'She could be afraid of a big black wolf.'

(11)

$$\left\langle \begin{array}{l}
\left[\begin{array}{l} \textit{s-node} \\ \text{PHON } \langle \textit{mohla} \rangle \\ \text{F } \left[\begin{array}{l} \textit{pre-cl-fld} \\ \text{R } \boxed{1}\ \textit{mtx-fld} \end{array}\right] \\ \text{I-C } \textit{no} \end{array}\right],
\left[\begin{array}{l} \textit{s-node} \\ \text{PHON } \langle \textit{by} \rangle \\ \text{F } \left[\begin{array}{l} \textit{cl-be-fld} \\ \text{R } \boxed{2} \left[\begin{array}{l} \textit{cl-fld} \\ \text{R } \boxed{1} \end{array}\right] \end{array}\right] \\ \text{I-C } \textit{no} \end{array}\right],
\left[\begin{array}{l} \textit{s-node} \\ \text{PHON } \langle \textit{se} \rangle \\ \text{F } \left[\begin{array}{l} \textit{cl-rfl-fld} \\ \text{R } \boxed{2} \end{array}\right] \\ \text{I-C } \textit{no} \end{array}\right],
\left[\begin{array}{l} \textit{s-node} \\ \text{PHON } \langle \textit{bát} \rangle \\ \text{F } \left[\begin{array}{l} \textit{rest-fld} \\ \text{R } \boxed{1} \end{array}\right] \\ \text{I-C } \textit{no} \end{array}\right], \\
\left[\begin{array}{l} \textit{s-node} \\ \text{PHON } \langle \textit{velkého} \rangle \\ \text{F } \left[\begin{array}{l} \textit{l-adj-fld} \\ \text{R } \boxed{3} \end{array}\right] \\ \text{I-C } \textit{yes} \end{array}\right],
\left[\begin{array}{l} \textit{s-node} \\ \text{PHON } \langle \textit{černého} \rangle \\ \text{F } \left[\begin{array}{l} \textit{l-adj-fld} \\ \text{R } \boxed{3} \end{array}\right] \\ \text{I-C } \textit{yes} \end{array}\right],
\left[\begin{array}{l} \textit{s-node} \\ \text{PHON } \langle \textit{vlka} \rangle \\ \text{F } \left[\begin{array}{l} \textit{noun-fld} \\ \text{R } \boxed{3} \left[\begin{array}{l} \textit{noun-fld} \\ \text{R } \left[\begin{array}{l} \textit{rest-fld} \\ \text{R } \boxed{1} \end{array}\right] \end{array}\right] \end{array}\right] \\ \text{I-C } \textit{yes} \end{array}\right]
\end{array} \right\rangle$$

The setup of topological regions in (11) is easier to see in (12).

(12)



The topological fields and regions used in the example are pre-clitic (initial) field, clitic field (here consisting of an auxiliary and a reflexive) and two 'rest fields', one for the verb and the other for the nominal group. The whole sentence is a single field (or region): matrix. An order is defined for the fields relative to a region and some fields must be adjacent (this applies to the nominal group and the clitic fields, the relevant fields specify the region in which they have to be adjacent by coindexing). The flat list structure allows for imposing constraints on the order in a monotonous way.

6 The 'backbone' constraints

After the theoretical and formal foundations have been specified and the necessary parts of signature defined, constraints on the setup of objects in the model can be formulated. There are four 'backbone' constraints which are responsible for the composition of non-lexical signs: Deep List Composition Principle, Surface List Composition Principle, Valency Principle and Phonology Principle. Their flavour can be tasted in the following examples, concerning the Deep List Composition Principle. Its formal expression (14) can be paraphrased as (13).

(13) In every *non-lexical* sign the mother's *d-list* consists of the head daughter's *d-list* into which the non-head daughters' *d-list*s are inserted.

(14)
$$\textit{non-lexical} \rightarrow \left(\begin{array}{l}\left[\begin{array}{l}\textsc{synsem} \mid \textsc{local} \mid \textsc{deep} \mid \textsc{tree}\ \boxed{5} \\ \textsc{head-daughter} \mid \textsc{synsem} \mid \textsc{local} \mid \textsc{deep} \mid \textsc{tree}\ \boxed{1} \\ \textsc{nonhead-daughters}\ \boxed{2}\end{array}\right] \\ \wedge\ \texttt{collect\_dlists}(\boxed{2}, \boxed{3}) \\ \wedge\ \texttt{append}(\boxed{1}, \boxed{3}, \boxed{4}) \\ \wedge\ \texttt{permute}(\boxed{4}, \boxed{5})\end{array}\right)$$

The notion of 'inserting *d-list*s' is expressed by means of three relations. The relation `collect_dlists/2` extracts a *d-list* from every non-head daughter and puts it on the list $\boxed{3}$. This list of *d-list*s is appended with the head daughter's *d-list* ($\boxed{1}$), yielding $\boxed{4}$, formally again a *d-list*. This list is permuted into the mother's *d-list* ($\boxed{5}$) and is subject to all other constraints on *d-list*s. The sentence (15) and its partial representation in (16) shows the effect of this constraint.[2]

(15)

Pepa	dneska	pase	sousedovu	kozu
Pepa-NOM	today	graze-PRES-3RD-SG	neighbour-POSS	goat-ACC

'Today Pepa is grazing the neighbour's goat'

(16)
$$\left[\begin{array}{l}\textit{non-lexical} \\ \textsc{phon}\ \langle \textit{Pepa,dneska,pase,sousedovu,kozu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \langle\ \boxed{2}\langle[\textit{Pepa}]\rangle,\ \boxed{3}\langle[\textit{dnes}]\rangle,\ \boxed{1}[\textit{pást}],\ \boxed{4}\langle\ \boxed{5}[\textit{koza}],\ \boxed{6}\langle[\textit{soused}]\rangle\ \rangle\ \rangle \\ \textsc{hd}\ \left[\begin{array}{l}\textit{lexical} \\ \textsc{phon}\ \langle \textit{pase} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \langle \boxed{1} \rangle\end{array}\right] \\ \textsc{nhd}\ \left\langle \left[\begin{array}{l}\textit{lexical} \\ \textsc{phon}\ \langle \textit{Pepa} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{2}\end{array}\right], \left[\begin{array}{l}\textit{lexical} \\ \textsc{phon}\ \langle \textit{dneska} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{3}\end{array}\right], \left[\begin{array}{l}\textit{non-lexical} \\ \textsc{phon}\ \langle \textit{sousedovu, kozu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{4} \\ \textsc{hd}\ \left[\begin{array}{l}\textit{lexical} \\ \textsc{phon}\ \langle \textit{kozu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \langle \boxed{5} \rangle\end{array}\right] \\ \textsc{nhd}\ \left\langle \left[\begin{array}{l}\textit{lexical} \\ \textsc{phon}\ \langle \textit{sousedovu} \rangle \\ \textsc{ss} \mid \textsc{l} \mid \textsc{d} \mid \textsc{t}\ \boxed{6}\end{array}\right] \right\rangle\end{array}\right] \right\rangle\end{array}\right]$$

[2] A few abbreviations have been used to make the feature structure more compact: (i) SS|L|D|T stands for the path SYNSEM | LOCAL | DEEP | TREE, HD for HEAD-DAUGHTER, and NHD for NONHEAD-DAUGHTERS, (ii) other than the most relevant attributes are suppressed, (iii) phonology substrings are not co-indexed, and (iv) *d-node*s are abbreviated as lemmas.

7 Three kinds of ordering constraints

Three kinds of ordering constraints can be distinguished: those that apply only to one of the levels – tectogrammatical level or surface level – and those that apply on the relation between the two levels. The constraints interact in the same way as the word order factors. If the relative order of any two items is unspecified by surface-level constraints, their order is determined by deep word order, which in turn is based on systemic ordering and the distinction between contextually bound and non-bound items.

There are four constraints on *d-lists*, which make sure that there is at least one non-bound none in the tree, and that in every subtree the governor is correctly positioned, non-bound nodes come last and are ordered according to systemic ordering.

The deep/surface order relation is defined by a single Deep/Surface Order Principle, which is applied to every pair of nodes in a local tectogrammatical tree, provided that the *s-list* position of none of them is determined by surface-level ordering constraints. This information is inferred from their topological field assignment. There are three disjuncts in the consequent of the principle: (i) the relative order of the two nodes is identical on both levels (Mathesius' FSP principle applies), or (ii) if the first node is non-bound and *d-list*-initial, it is assigned the pre-clitic field and occur in the domain of a higher clause (adjacency may be violated), or (iii) if the second node is *d-list*-final, it can receive an appropriate stress and be placed in a non-final *s-list* position (emphasis principle applies).

The other ordering regularities are imposed by surface-level constraints. There are five general and some more construction-specific constraints on *s-lists*. The general *s-list* constraints are modified versions of the non-parochial constraints on domain lists of (Penn, 1999a): the principles of Matrix Compaction, Planarity, Topological Order, Field Existence and Field Uniqueness. The latter three are generalized to allow separate definitions of the setup of regions: the order of fields in a region and how many fields of one type may or must occur in a region. The definition of the top region *matrix-fld* is shown in Table 1.

Region	Field	Order	Occupancy
matrix-fld	*pre-cl-fld*	1	1
	cl-fld	2	$\leq$1
	rest-fld	3	any
	fin-fld	4	$\leq$1

Table 1: Fields within the top region

Construction-specific constraints on *s-lists* almost always refer to the assignments of topological fields, and – in some cases – also to morphosyntactic properties of the signs involved. They treat a few cases of discontinuous realizations of tectogrammatical subtrees: comparison constructions (*a smaller village than Lhota*), elementary cases of long-distance dependencies, split prepositional phrases (*O jakou se jedná soutěž?*, lit. 'About what is being talked competition?'), and – last but not least – the placement of clausal clitics in the post-initial position, including clitic climbing and haplology. Clitics receive closer attention with a number of examples in a separate chapter.

Suggested solutions are all based on the assignment of topological fields to *s-node*s. Admittedly, this may not be the optimal answer to the issues of long-distance dependencies and especially to clitic climbing and haplology, where a treatment based on syntactic structure rather than on *s-list* may be preferable.

8 Conclusion

There is a lot more to do. Many issues which have been opened in the thesis deserve corpus-based research, deeper analysis, and the suggested solutions would definitely profit from proper evaluation, preferably in an implemented system. Still, I believe I can list a few achievements:

a. Arguments have been presented in favour of the combination of FGD as a well-founded linguistic theory, and RSRL as an adequate formal language. The framework was applied to a range of word order phenomena in Czech. An account of the relation between deep and surface order and its interaction with surface-level constraints has been proposed.
b. The word order principles of Vilém Mathesius were shown to be compatible with the constraint-based formalism. The principal role of functional sentence perspective and the hierarchy of communicative dynamism in determining surface order and prosody, as assumed in FGD, has been embodied in constraints interacting with other ordering constraints.
c. The compositional basis of the relation between surface string and tectogrammatical representation was has been described as a flat derivation structure. Tectogrammatical representation was formalized as a recursive list structure, morphemic string as a non-recursive list with adjacency and ordering information encoded within its individual members.
d. The empirical facts and premises of the theory have been formalized in a way which allows for the interaction of factors conditioning surface word order, where deep word order determines surface ordering when it is underspecified by other constraints. This has been achieved by

applying a few general and some construction-specific constraints to topological fields, tectogrammatical representation and – in some cases – also to morphosyntactic properties of the signs involved.

e. Some regularities concerning Czech clausal clitics have been suggested, confirming the interaction of syntactic, discourse and prosodical factors, with syntactic factors playing the main role.

Unfortunately, not all goals have been achieved. Implementation of the description of a fragment of Czech is previewed as the next step, together with the necessary rephrasing of some of the descriptions into a more computationally tractable form, which, howerver, should be both intuitive an theoretically adequate. This is necessary in order to verify the descriptions, but also to assess chances of further development.

Not all investigated phenomena, have received an adequate description. This applies especially to those presented in the chapter on clitics. This may lead to modification of some more or less essential formal or theoretical aspects. Before such a move is made, a more extensive application of the approach to other phenomena is previewed, not only to those related to word order and not only those present in Czech.

References

Mathesius, Vilém. 1975. On Information-Bearing Structure of the Sentence. In Susumu Kuno, editor, *Harvard Studies in Syntax and Semantics*. pages 467–480.

Pollard, Carl J. and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Pollard, Carl J. and Ivan A. Sag. 1987. *Information-Based Syntax and Semantics*. Number 13 in CSLI Lecture Notes. CSLI, Stanford, California, USA.

Sgall, Petr. 1992. Underlying structure of sentences and its relations to semantics. In Tilmann Reuther, editor, *Festschrift für Viktor Jul'evič Rozenberg*. Wien, pages 273–282. Wiener Slawistischer Almanach, Sonderband 33.

Penn, Gerald. 1999a. Linearization and WH-extraction in HPSG: Evidence from Serbo-Croation. pages 149–182.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Reidel and Academia, Dordrecht and Praha. Editor: Jacob Mey.

Richter, Frank. 2000. A mathematical formalism for linguistic theories with an application in head-driven phrase structure grammar. Dissertation, submitted to the Neuphilologische Fakult t of the Universit t T bingen, Version of April 28th, 2000.