Word order factors as constraints on feature structures

Alexandr Rosen

Abstract

According to Vilém Mathesius, word order is conditioned simultaneously by factors belonging to various levels of the system of language. This view is inherent in the theory of *Functional Generative Description* (FGD), where discourse-related factors interact with surface-level regularities in determining word order and prosody. However, the standard way of describing a language within the theory is by means of a grammar generating the underlying syntactic tree and a series of transducers mediating between formally separate levels of description. In the present contribution, based on a dissertation (Rosen, 2001), it is argued that FGD as a linguistic theory can use a formalism which is better suited to capture the simultaneously functioning word order factors: *Relational Speciate Re-Entrant Logic*, one of the formal tools developed for 'constraint-based' grammars. A data structure representing linguistic objects in the spirit of FGD is proposed and some examples are provided of how word order phenomena in Czech can be described as conditioned by several factors from different levels of the language system.

1 Introduction

The linguistic theory of Functional Generative Description (FGD) is centred around the abstract level of *linguistic meaning*. This level provides *tectogrammatical* or *underlying* representation of language expressions in the form of syntactic dependency trees with annotated nodes corresponding to content words. Like other dependency-based frameworks, FGD maintains an inherent distinction between the components of representation and derivation: representation of an expression at an abstract level can be described independently from how the representation is related to the actual string of phonemes/graphemes. As a result, the abstract representation need not include information about the way it is related to the string. This contrasts with phrase structure grammars and theories such as GPSG, LFG or HPSG, where representation and derivation coincide in a single structure.

An adequate description of the derivation component – the interaction between the underlying syntactic structure and its surface realization – is an important goal of theoretical linguistic research. In standard theoretical work on FGD, the derivation component has the shape of a sequence of transducers, see Panevová (1979) or Plátek, Sgall, and Sgall (1984). A language is then described by a grammar generating the underlying structure, and transducing components, including movement rules, providing the interface between formally distinct levels of description. An implementation of Czech generator based on this proposal is described in Panevová (1982) and Borota (1990).

Stratificational frameworks in which levels of descriptions are derived successively by means equivalent to transformations or movement rules have been subjected to criticism for several shortcomings: they are biased towards one direction of processing (generation), do not allow simultaneous access to information at all levels of description, and prevent interpretation of partial expressions (Sag, 1995). That the standard formalism for FGD is not immune to problems of this sort is noticeable in other FGD-inspired projects, aimed at applications involving analysis, generation or grammar checking. The projects tend to employ formalisms without transducers: the machine translation projects APAČ (Kirschner, 1982) and RUSLAN (Oliva, 1989) are based on Q-systems (Colmerauer, 1970), the grammar checker LATESLAV uses RFODG (Kuboň,

Holan, and Plátek, 1997). Thus, it seems that FGD may be formalized in a different way from that originally proposed and that the search for an alternative formalism may be worthwhile.

Let us assume the hypothesis that FGD can indeed be formalized in a different way.[1] This hypothesis does not imply that a formalism actually used is as suitable as any other: there are certainly some preferred theory/formalism combinations, which are determined by their design and their authors' choice. To explore a new combination might be justified if the theory and the formalism have shown their advantage over the competitors, albeit in a different combination.

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 semantico-pragmatic interpretations, the relevance of 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,[2] according to Sgall, Hajičová, and Panevová (1986) and Sgall (1992).

Instead of the standard stratificational framework, the choice has been made to use a declarative formalism, allowing for parallel description of expressions at different language levels, namely *Relational Speciate Re-Entrant Logic* (RSRL) (Richter, 2000), a formal language assumed in constraint-based theories such as HPSG (Pollard and Sag, 1994). The formalism 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 objects modelling events or objects in the linguistic reality.

The 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 the hierachy of communicative dynamism, should receive adequate treatment in such an account.

According to Vilém Mathesius (Mathesius, 1939; 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, it may be expected 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.

2 Theory and its formalization

There are a number of arguments in favour of dependency-based underlying syntax in the spirit of FGD 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) concerned with the system of language, rather than with extra-linguistic inferences. Another good reason is the detailed treatment of *topic-focus articulation* (TFA): all nodes in the underlying dependency tree are distinguished as being *contextually bound* or *contextually non-bound* and they are ordered according to the hierarchy of *communicative dynamism*, where the most dynamic items (usually carrying new information) come last.

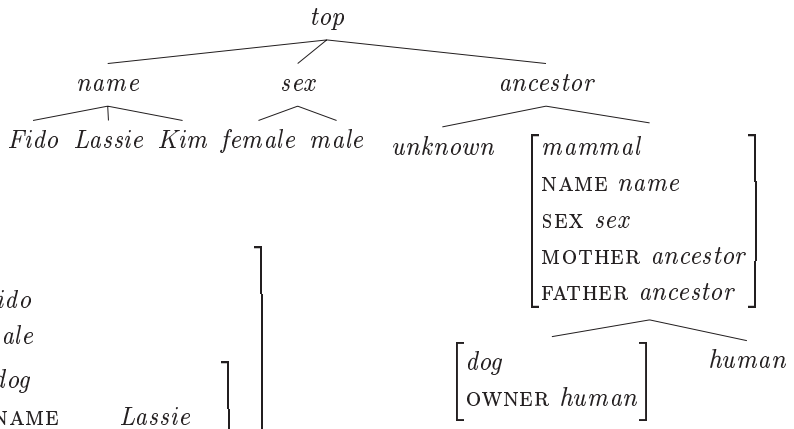
[1] A stronger hypothesis, which will not be commented here, would be that any theory is compatible with any formalism satisfying some minimal criteria.

[2] Coordination and apposition are represented in the third dimension of the dependency tree.

The relation between the underlying level and the string of letters or sounds, in the standard FGD approach described by transducing components and movement rules, receives a fully declarative treatment. Syntactic units are modelled as typed feature structures, defined by a system of constraints – the grammar and the lexicon. 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. An RSRL grammar consists of two parts: *signature* and *theory*. *Signature* defines what kinds of objects are potential parts of the model. These objects are typed feature structures interpreted as representating sets of linguistic events and their properties in the real world. *Theory* puts constraints on potential objects, excluding objects not satisfying the constraints. For simplicity, the use of the formalism is illustrated by a non-linguistic example.

(1)



(2)

$$\begin{bmatrix} \textit{dog} & \\ \text{NAME} & \textit{Fido} \\ \text{SEX} & \textit{male} \\ \text{MOTHER} & \begin{bmatrix} \textit{dog} & \\ \text{NAME} & \textit{Lassie} \\ \text{SEX} & \textit{male} \\ \text{MOTHER} & \textit{unknown} \\ \text{FATHER} & \textit{unknown} \\ \text{OWNER} & \boxed{1} \end{bmatrix} \\ \text{FATHER} & \textit{unknown} \\ \text{OWNER} & \boxed{1}\begin{bmatrix} \textit{human} & \\ \text{NAME} & \textit{Kim} \\ \text{SEX} & \textit{female} \\ \text{MOTHER} & \textit{unknown} \\ \text{FATHER} & \textit{unknown} \end{bmatrix} \end{bmatrix}$$

(3) $\left[\text{MOTHER } \textit{mammal}\right] \rightarrow \left[\text{MOTHER} \mid \text{SEX } \textit{female}\right]$

(4) $\begin{bmatrix} \textit{human} \\ \text{MOTHER } \textit{mammal} \end{bmatrix} \rightarrow \left[\text{MOTHER } \textit{human}\right]$

(5) $\textit{dog} \rightarrow \begin{bmatrix} \text{OWNER } \boxed{1} \\ \text{MOTHER} \mid \text{OWNER } \boxed{1} \end{bmatrix}$

The signature in (1) defines a hierarchy of types in a model of a simple world of dogs and their masters. Types may have attributes whose appropriate values are again types. A type inherits all attributes and appropriate values from its supertype(s). The feature structures in the model are only of 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 1 coindexes identical parts of the structure). 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 specified

again as a *mammal* type object, the value of SEX of the latter object must be *female*. A similar description can be used to exclude a dog as the mother of a human (4). Finally, (5) asserts a less self-evident fact that each dog owner also owns the dog's mother.

The description in (5) includes a variable (1). If a variable is not explicitly bound by a quantifier, then it is bound by implicit existential quantification scoping over the entire formula. Quantification in RSRL is always 'bounded' – restricted to the components of an object being described, rather than quantifying over all entities in the model. RSRL also includes the usual logical connectives, the possibility to define and use relations, and negation (which is interpreted as in classical logic).

In a linguistic application, the typed feature structures would correspond to words and syntagms (phrases) and to the various aspects of their analysis, denoted by the attributes: a structure including its constituent parts, a list of valency requirements, a string of graphemes and its interpretation at various levels of description, values of categories (part of speech, case, gender). The formalism does not restrict the power ("type") of grammar: the surface string may be defined as the concatenation of the terminal yield of the derivation tree, or a more liberal relation may be defined, allowing for solutions to word order phenomena to go beyond context-free grammar, as in *linearization grammars* (Reape, 1994; Kathol, 1995; Penn, 1999).

3 Factors determining word order

Vilém Mathesius proposed the idea of interacting and mutually competing word order principles (Mathesius, 1939; Mathesius, 1975), which are universal, but have different roles in different languages. For every language, a partial order of these principles can be specified, predicting which principles win if several principles compete for different word orders.

FGD views one of the principles, namely the *topic-focus articulation* (TFA) principle, as primary. According to FGD, TFA of a specific utterance includes the distinction between contextually bound and non-bound elements and the hierarchy of communicative dynamism, the *deep word order* (DWO). TFA is represented together with other aspects of linguistic meaning at the tectogrammatical level and is revealed in various surface-level phenomena: *surface word order* (SWO), stress patterns, syntactic constructions.

If a 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 TFA and its surface realization, as conditioned by other factors. TFA 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. Thus, the impact of word order factors can be outlined as follows:[3]

a. For every pair of content words A and B, the relative SWO of A and B corresponds to DWO of the corresponding semantemes, unless any of the cases in the list of Special SWO Conditions apply to A and B (see below).
b. Each function word F is ordered adjacently to its host H, their order being determined by a syntactic constraint, unless any of the cases in the list of Special SWO Conditions apply to F and H (see below).
c. For the relative SWO of each pair of function words F_1 and F_2 in a single ordering domain the Special SWO Conditions apply.

[3]The statements are based on the situation in Czech and probably do not hold across all languages.

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.[4]

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.

4 The architecture

The combinatorial properties of syntactic units are recorded in a flat derivation structure with function words standing as sisters to dependents and their 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).

(6)

sign				
PHONOLOGY	*list(phonstring)*			
SYNSEM	*synsem*			
	LOCAL	*local*		
		CATEGORY	*category*	
			HEAD	*head*
			VALENCY	*list(synsem)*
			...	
		DEEP	*deep*	
			STATUS	*status*
			TREE	*d-list*
		CONTENT	*content*	
	NONLOCAL	*nonlocal*		
SURFACE	*s-list*			

The similarity with the type *sign* in HPSG is intended in order to allow for an easy adoption of solutions to some surface-level phenomena available in that theory and adequate within the context of FGD. However, nothing substantial and relevant for the issues discussed here hinges on this similarity.

The type *deep* represents the level of tectogrammatics and *surface* the level of morphemics (a 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 valency and other surface syntactic properties of the expression: *category* and – optionally – its semantic interpretation (*content*).

[4] This is the case of long-distance dependencies, such as that in '*Him* Mary says Sue told,' where the pronoun *him* is the bearer of contrastive stress and the contrastive topic. Its SWO domain is the whole sentence and its tectogrammatical subtree is the embedded clause.

The type *sign* has two subtypes: *lexical* and *non-lexical*. 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 tree is represented 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. The tectogrammatical tree shown in Fig. 1, representing the sentence (7), is transcribed into a linear notation schematically illustrated in (8). Each pair of angle brackets encloses a list.

(7) Máňa šla tancovat
Máňa went to dance

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

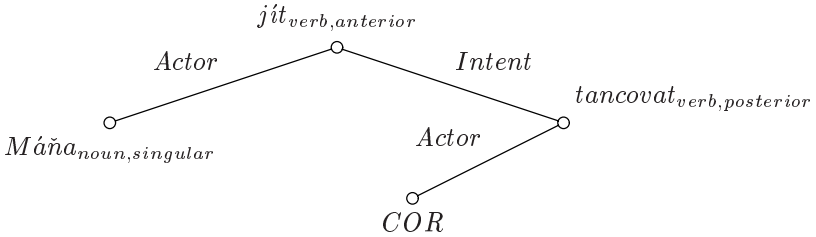


Figure 1: A tectogrammatical tree

The representation of expressions on the tectogrammatical level by means of embedded lists relies on the fact that tectogrammatical trees are projective.[5] 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)

$$\left\langle \left[\begin{array}{lll} \textit{d-node} & & \\ \text{FUN} & \textit{actor} & \\ \text{CB} & \textit{yes} & \\ \text{CORE} & \boxed{1} & \left[\begin{array}{lll} \textit{d-noun} & & \\ \text{LEMMA} & \textit{Máňa} & \\ \text{D-NUMBER} & \textit{sg} & \end{array}\right] \end{array}\right] \right\rangle, \left[\begin{array}{lll} \textit{d-node} & & \\ \text{FUN} & \textit{root} & \\ \text{CB} & \textit{no} & \\ \text{CORE} & \left[\begin{array}{lll} \textit{d-verb} & & \\ \text{LEMMA} & \textit{jít} & \\ \text{D-TENSE} & \textit{anterior} & \end{array}\right] & \end{array}\right],$$

$$\left\langle \left\langle \left\langle \left[\begin{array}{lll} \textit{d-node} & & \\ \text{FUN} & \textit{actor} & \\ \text{CB} & \textit{yes} & \\ \text{CORE} & \boxed{1} & \end{array}\right] \right\rangle, \left[\begin{array}{lll} \textit{d-node} & & \\ \text{FUN} & \textit{intent} & \\ \text{CB} & \textit{no} & \\ \text{CORE} & \left[\begin{array}{lll} \textit{d-verb} & & \\ \text{LEMMA} & \textit{tancovat} & \\ \text{D-TENSE} & \textit{posterior} & \end{array}\right] & \end{array}\right] \right\rangle \right\rangle$$

The representation of an expression on the level of morphemics (string of lexical items) constitutes a simple list (*s-list*). Its setup and the setup of its members is inspired by *domain lists* of (Penn, 1999), where embeddable *topological fields* and structure sharing (represented as coindexation) are used to determine position and adjacency of list members (*domain objects*). The following example illustrates the setup of *s-list*.

[5] Other notations are available if the condition of projectivity is lifted.

(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.'

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). Every *s-node* is appropriate to the binary-valued feature I-C(ENTRE) which stands for *information centre* and represents a very basic account of prosody. For the final items, its value is usually set positive by a constraint on *s-list*s. The value can also be set in the lexicon, usually negative for items that can never bear the intonation centre.

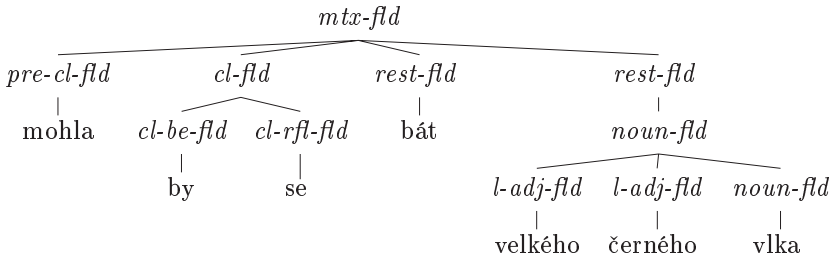
(11)

$$\left\langle \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{mohla} \rangle \\ \text{F} \begin{bmatrix} \textit{pre-cl-fld} \\ \text{R } \boxed{1}\ \textit{mtx-fld} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{by} \rangle \\ \text{F} \begin{bmatrix} \textit{cl-be-fld} \\ \text{R } \boxed{2} \begin{bmatrix} \textit{cl-fld} \\ \text{R } \boxed{1} \end{bmatrix} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{se} \rangle \\ \text{F} \begin{bmatrix} \textit{cl-rfl-fld} \\ \text{R } \boxed{2} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{bát} \rangle \\ \text{F} \begin{bmatrix} \textit{rest-fld} \\ \text{R } \boxed{1} \end{bmatrix} \\ \text{I-C } \textit{no} \end{bmatrix}, \right.$$

$$\left. \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{velkého} \rangle \\ \text{F} \begin{bmatrix} \textit{l-adj-fld} \\ \text{R } \boxed{3} \end{bmatrix} \\ \text{I-C } \textit{yes} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{černého} \rangle \\ \text{F} \begin{bmatrix} \textit{l-adj-fld} \\ \text{R } \boxed{3} \end{bmatrix} \\ \text{I-C } \textit{yes} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHON } \langle \textit{vlka} \rangle \\ \text{F} \begin{bmatrix} \textit{noun-fld} \\ \text{R } \boxed{3} \begin{bmatrix} \textit{noun-fld} \\ \text{R} \begin{bmatrix} \textit{rest-fld} \\ \text{R } \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{I-C } \textit{yes} \end{bmatrix} \right\rangle$$

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

(12)



The specification of the properties of surface order is made possible by each *s-node* being assigned a relative topological field. The field is specified as a part of a region, which in turn can be specified as a field within a higher region. Thus, the position of each field is determined by a path of regions terminating in *matrix-fld*. The flat list structure allows for imposing constraints on the order in a monotonous way.

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. Some fields must be adjacent – this applies to the nominal group and the clitic fields, the relevant fields specify the (continuous) region in which they have to be adjacent by coindexing. In other words, fields are said to 'compact' to a region. In (11) the index $\boxed{1}$ is used to point to the single topmost region of the type *matrix-fld*, to which all the second-level regions/fields (*pre-cl-fld*, *cl-fld* and the two *rest-fld*s) compact. The

two clitics compact to the region *cl-fld*, indexed by 2, and the three components of the nominal group compact to *noun-fld* using the index 3.[6]

A nonlexical sign, representing a string, is composed from other signs, representing substrings which make up the larger sign's string. The setup of all subparts (attribute values) of the nonlexical sign is governed by constraints ('rules') of grammar, making sure that the information in the corresponding subparts of the component signs is combined in a proper way. These 'backbone' constraints determine the elementary properties of signs and their setup: they handle the composition of the value of the attribute PHONOLOGY, the composition of the types *d-list* and *s-list*, and the satisfaction of valency requirements. Other constraints are responsible for more specialized tasks, such as constraining word order.[7] The following example concerns the constraint on deep lists, 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)

$$\textit{non-lexical} \rightarrow \left(\begin{array}{l} \left[\begin{array}{l} \text{SYNSEM} \mid \text{LOCAL} \mid \text{DEEP} \mid \text{TREE } 5 \\ \text{HEAD-DAUGHTER} \mid \text{SYNSEM} \mid \text{LOCAL} \mid \text{DEEP} \mid \text{TREE } 1 \\ \text{NONHEAD-DAUGHTERS } 2 \end{array} \right] \\ \wedge\ \texttt{collect\_dlists}(2, 3) \\ \wedge\ \texttt{append}(1, 3, 4) \\ \wedge\ \texttt{permute}(4, 5) \end{array} \right)$$

The notion of 'inserting *d-list*s' is expressed by means of three relations. The first relation `collect_dlists/2` extracts a *d-list* from every non-head daughter and puts it on the list 3. This list of *d-list*s is appended with the head daughter's *d-list* (1), yielding 4, formally again a *d-list*. This list is permuted into the mother's *d-list* (5) and is subject to all other constraints on *d-list*s. The sentence (15) and its partial representation in Fig. 2 on p. 9 shows the effect of this constraint.

(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'

A few abbreviations have been used to make the picture of the feature structure in Fig. 2 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.

5 Three kinds of ordering constraints

Three kinds of ordering constraints can be distinguished: those that apply to the tectogrammatical level, those that apply to the surface level, and those that apply to the relation between the two. The constraints should interact similarly as word order factors: if the relative order of any two items is unspecified by surface-level constraints, their order is determined by deep word

[6] Each word class has its standard lexically specified field assignment: *noun-fld*, *adj-fld*, *adv-fld*, *prep-fld*. Lexical items usually compact with members of the same syntactic paradigms into regions bearing an identical name.

[7] The proposed organisation of constraints has its theoretical appeal, but has very poor computational properties. For example, in an implemented grammar, word order constraints would have to be integrated with the 'backbone' constraints.

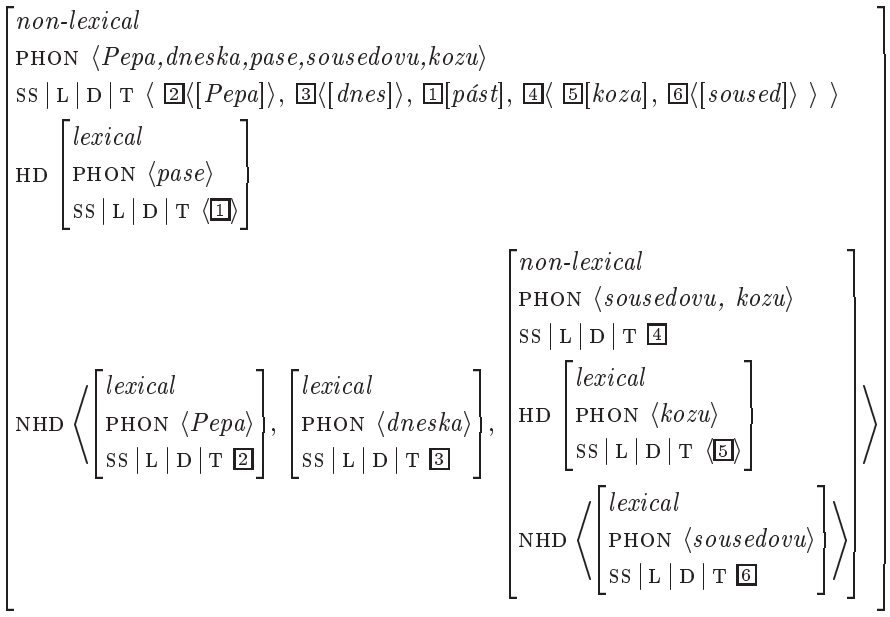


Figure 2: Feature structure representing a sentence

order, which in turn is based on systemic ordering and the distinction between contextually bound and non-bound items.

The deep-level constraints on *d-list*s determine the shape and content of the tectogrammatical tree, including its TFA-related properties. They make sure that there is at least one non-bound node in the whole tree, and they also check that in every subtree the governor is correctly positioned, that non-bound nodes come last and that they are ordered according to systemic ordering.

The deep/surface order relation is defined by a single Deep/Surface Order Principle (DSOP), 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 any of surface-level ordering constraints – this information is inferred from their topological field assignment (see below). There are three disjuncts in the consequent of the principle, corresponding to three options of ordering a pair of nodes A and B, where A precedes B on tectogrammatical level:

Identical Order – the relative order of the two nodes is identical on both levels (tectogrammatical and surface – Mathesius' TFA principle applies),[8] or

Left Dislocation of Topic Proper – if A is contextually-bound and *d-list*-initial, it is assigned the pre-clitic field and occurs in the domain of a higher clause (adjacency may be violated),[9] or

Non-final Placement of Intonation Centre – if B is *d-list*-final, it can receive an appropriate stress and be placed in a non-final *s-list* position (emphasis principle applies).[10]

[8] See above in §3: "For every pair of content words A and B, the relative SWO of A and B corresponds to DWO of the corresponding semantemes, unless any of the cases in the list of Special SWO Conditions apply to A and B (see below)."

[9] Contrastive topic may be treated in this way as well. However, it would need to be explicitly marked as such in its *d-node* and as the recipient of contrastive stress in the corresponding *s-node*.

[10] The rightmost node in the local tectogrammatical tree (can be a dependent or the governor) is naively assumed to be focus proper. The surface counterpart of such a node can either appear in accordance with the Identical Order disjunct in the final position of the corresponding region, or precede other nodes in that region (or even in higher regions, if not compacted to that region), provided that it receives an appropriate stress.

Finally, there are surface-level ordering constraints, which in effect override the DWO/SWO constraints described above. If a pair of nodes is subject to such constraints, none of the SWO/DWO constraints should be applied to the pair. However, given that default constraint application is not possible in RSRL, how can one prevent the application of DWO/SWO constraints to such pairs?

First of all, DWO/SWO constraints are allowed to apply only to some pairs of nodes (namely those which are not subject to surface-level constraints). This is achieved by a condition in the antecedent of DSOP. The appropriate placement of such items is determined by surface constraints using the items' field specifications.

Then comes the issue how to specify the class of items exempt from DWO/SWO constraints. The answer is based on the following hypothesis: DSOP applies whenever surface-level constraints underspecify an item's position in SWO.

For most cases, the notion of surface-level underspecification can be defined as follows: The SWO position of an *s-node* is underspecified by surface-level constraints if its field within a region can be assigned to other *s-node*s within that region, i.e., if in a region the same field can be assigned to multiple *s-node*s.

Rather than enumerating items exempt from DWO/SWO constraints, it seems reasonable to specify items to which the DWO/SWO constraints do apply: they apply only to those items, whose order within a region is not fully determined (is underspecified) by surface-level constraints. Whenever an item's surface position is underspecified by surface-level constraints, the field value for the item relative to a region is one of those values which can be used repeatedly within that region. An example of such a field would be *rest-fld* within *matrix-fld*.

Yet there can be unique positions, fields which can only be assigned to a single item (including a compacted cluster), fillable either by surface-level constraints or by DSOP. Such is the case of *pre-cl-fld*, which can be assigned either to an interrogative expression or to an item according to DSOP.[11] Such fields should also be added to the list of multiply fillable fields, those which allow the application of the consequent of DSOP. Since a field such as *pre-cl-fld* can be assigned only once, an item with the field specified by surface-level constraints will be placed correctly.

Thus, DSOP includes the conditions that both of the two surface nodes must be assigned *rest-fld* or *pre-cl-fld*. On the assumption that these two fields behave in the same way irrespective of their region, the region of these fields is not specified. DSOP is then vacuously satisfied by pairs of nodes whose position is determined by surface-level constraints.

6 A closer look at surface-level constraints

There are five general and some more construction-specific constraints on *s-list*s. The general *s-list* constraints are modified versions of the non-parochial constraints on domain lists of Penn (1999): 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. Definitions of the setup of regions can also be translated into a more readable format. The definition of the top region *matrix-fld* is shown Table 1.

The order of *s-node*s is specified jointly by surface-level constraints and SWO/DWO constraints. Surface-level constraints impose an order on *s-node*s by using two notions: order of fields within a region and continuity of the region. The order of *s-node*s must correspond to the region-specific definition of the order of fields and there may not be any *s-node* in the region whose field

[11] The position of interrogative expressions is not determined by DWO. On the other hand, relative expressions are treated as least dynamic items and are ordered by DSOP.

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

Region	Field	Order	Occupancy
cl-fld	*cl-lis-fld*	1	$\leq$2
	cl-be-fld	2	$\leq$1
	cl-rfl-fld	3$\vee$4	$\leq$1
	cl-ethdat-fld	3$\vee$4	$\leq$1
	cl-freedat-fld	5	any
	cl-dat-fld	6	any
	cl-acc-fld	7	any
	cl-gen-fld	8	any
	cl-ins-fld	9	any
	cl-nom-fld	any	$\leq$1
	cl-uz-fld	any	$\leq$1
	cl-pry-fld	any	$\leq$1
	cl-vsak-fld	any	$\leq$1

Table 2: An overview of topological fields for clitics

specification is not included in the definition of the region. In other words, *s-node*s 'compact' to the region.

In Table 1, the fields in the Fields column compact to the region specified in the leftmost Region column, in the order which is indicated in the Order column. The column Occupancy shows how many times a field can occur within the region. A field may include a number of compacted *s-node*s.

The assignment of a field to *s-node* is conditioned by several factors: lexicon, DWO/SWO constraints, and surface-level constraints. Most surface-level constraints define the order and number of fields within a region and can be informally expressed as tables such as Table 1 and Table 2.

Table 2 defines the setup of *cl-fld*, i.e., the Czech 'second position' clitic cluster.[12] The order of some clitics is fixed. This concerns the clitical conjunction *-li* 'if', forms of the auxiliary *být* 'to be' (*cl-lis-fld* $\prec$ *cl-be-fld*), some unstressed personal pronouns, assigned to fields according to their case (*cl-dat-fld* $\prec$ *cl-acc-fld* $\prec$ *cl-gen-fld* $\prec$ *cl-ins-fld*), and partly to reflexive particles (*cl-rfl-fld*) and 'ethical dative' pronouns (*cl-ethdat-fld*). For other clitics (mostly weak adverbials) placed towards the end of the cluster, the order is less rigid.

[12]For more details, including arguments, examples and a discussion of the phenomena of clitic climbing and haplology, see Rosen (2001), §7.

7 A simple example

In order to show how the application of constraints leads to a fully specified representation of an expression, example analysis of a very simple sentence follows. The sentence in (16) corresponds to an *s-list* consisting of two items:

(16) Děti spí.
children sleep
'The children are sleeping.'

The sign representing the sentence obtains essentially by applying the 'backbone constraints', such as the principle governing valency satisfaction. Another principle makes sure that the sign's *s-list* consists of two *s-node*s.

The field of the *s-node* for *spí* is specified in the lexicon as *pre-cl-fld* or *rest-fld*. The topmost region must be *matrix-fld*, and according to Table 1 the *s-node* for *spí* can become a field within that region. If *spí* were the only *s-node* in the sentence, then its field assignment would necessarily be *pre-cl-fld* – this is the only obligatory field in *matrix-fld*. However, the *s-node* for *spí* is not the only item on the *s-list*, so the choice between *pre-cl-fld* and *rest-fld* cannot be resolved in this way.

The field of the *s-node* for *děti* is lexically specified as *noun-fld*, which can trivially compact to the same regions as the verb: *pre-cl-fld* or *rest-fld*.[13] Since the *matrix-fld* region must be continuous, the region specifications of both *děti* and *spí* point to the same object of type *matrix-fld*, i.e., both *s-node*s compact within the top region.

Now there are two possibilities for the application of Deep/Surface Order Principle: either the Identical Order disjunct applies with *spí* as the bearer of the intonation centre and the focus proper, or the Non-final Intonation Centre Placement disjunct applies with *děti* as carrying the intonation centre and constituting the focus proper. In either of the two cases *pre-cl-fld* must be filled by *děti*: the order of *s-node*s must correspond to the actual order of words in the surface string and must obey the definition of the region *matrix-fld*, more specifically the properties of the field *pre-cl-fld*. Because this field accommodates exactly one occupant, the region in *s-node* for *spí* must be assigned the only remaining option: *rest-fld*.

The *s-list* corresponding to the first possibility with *spí* as the focus proper carrying the intonation centre is shown in (17).[14]

(17)
$$\left\langle \begin{bmatrix} \textit{s-node} \\ \text{PHONOLOGY } \langle \textit{děti} \rangle \\ \text{FIELD} \begin{bmatrix} \textit{noun-fld} \\ \text{R} \begin{bmatrix} \textit{pre-cl-fld} \\ \text{R } \boxed{1}\ \textit{matrix-fld} \end{bmatrix} \end{bmatrix} \\ \text{I-CENTRE } \textit{no} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{PHONOLOGY } \langle \textit{spí} \rangle \\ \text{FIELD} \begin{bmatrix} \textit{rest-fld} \\ \text{R } \boxed{1} \end{bmatrix} \\ \text{I-CENTRE } \textit{yes} \end{bmatrix} \right\rangle$$

A very similar picture can be shown if the verb has more dependents. Each dependent of any word class is eventually assigned two possible regions: *pre-cl-fld* or *rest-fld*. The initial position is licensed by the item being either topic proper or focus proper, in the latter case with an appropriate stress marking.

There is one more field optionally available within *matrix-fld*, which is situated to the right of *rest-fld*: *fin-fld*. This field is used for extraposed and/or phonologically heavy dependents,

[13] Definitions of the regions are omitted for space reasons. If the noun were itself modified, then the field *noun-fld* would compact with all its modifiers into a larger *noun-fld*.

[14] The attribute REGION is abbreviated as R.

such as embedded clauses, and can be multiply filled. Being exempt from Deep/Surface Order Principle, this field is assigned by construction-specific constraints.

8 Discontinuity: a slightly harder problem to solve

The following three points summarize what has been proposed so far to constrain surface word order:

1. Constraints on deep word order, which influence the surface order indirectly through Deep/Surface Order Principle.

2. Deep/Surface Order Principle, which mediates between deep and surface word order. This principle applies only to pairs of content words in a local tectogrammatical tree whose corresponding *s-node*s are assigned *rest-fld* or *pre-cl-fld* within the region corresponding to the tree. The principle relates deep word order with the surface order of compacted items including such *s-node*s, taking into account the intonation centre.

3. The principles of Topological Order, Field Existence and Field Uniqueness apply to *s-list* as a value of SURFACE. They impose an order on fields relative to the lowest region common for all *s-node*s within the *s-list*.

The means enumerated above determine word order in regular cases, where all subtrees are realized continuously. Discontinuously realized subtrees include comparison constructions (*a smaller village than Lhota*), clitic climbing, *wh-* 'movements' and other long-distance dependencies including 'adjunct extraction from NP' as in *Jakou jste mysleli soutěž?*, lit. 'Which did you mean competition?' and the same phenomenon involving 'split-PPs' as in *O jakou se jedná soutěž?*, lit. 'About what is being talked competition?', etc.[15]

Discontinously realized items (*s-node*s) corresponding to a given subtree should not be compacted with other items corresponding to the subtree. Instead, they should be free to compact in a larger region.[16] This concerns items whose (potentially) discontinuous position is obligatory, such as interrogatives. However, a number of constructions have a continuous and a discontinuous variant. In such cases, compaction with other items of the same subtree should be an option rather than necessity, along the alternative to compact in a larger region. Alternatives are appropriate in the two variants of the comparative construction (18a) and (18b), as well as in cases of clitic climbing. In the discontinuous case, the *s-list* looks as in (19).[17]

(18) a. menší vesnice než Lhota
smaller village than Lhota
b. vesnice menší než Lhota
village smaller than Lhota

[15] Construction-specific constraints almost always refer to topological fields to *s-node*s, and sometimes to morphosyntactic properties of the signs involved. Admittedly, this may not be the optimal answer to the issues of long-distance dependencies, especially to clitic climbing and haplology, where a treatment based on syntactic structure rather than *s-list* may be preferable.

[16] This is a solution used in different guises elsewhere, e.g. Kathol (1995), Kupść (2000), Penn (1999).

[17] For definitions of the regions and other details, see Rosen (2001).

(19)
$$\left\langle \begin{bmatrix} \textit{s-node} \\ \text{P}\ \langle \textit{menší} \rangle \\ \text{F}\ \begin{bmatrix} \textit{l-adj-fld} \\ \text{R}\ \boxed{1}\ \textit{noun-fld} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{P}\ \langle \textit{vesnice} \rangle \\ \text{F}\ \begin{bmatrix} \textit{noun-fld} \\ \text{R}\ \boxed{1} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{P}\ \langle \textit{než} \rangle \\ \text{F}\ \begin{bmatrix} \textit{sconj-compar-fld} \\ \text{R}\ \boxed{2}\ \begin{bmatrix} \textit{compar-base-fld} \\ \text{R}\ \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \textit{s-node} \\ \text{P}\ \langle \textit{Lhota} \rangle \\ \text{F}\ \begin{bmatrix} \textit{noun-fld} \\ \text{R}\ \boxed{2} \end{bmatrix} \end{bmatrix} \right\rangle$$

In the following example of the 'split-PP' phenomenon (20) there is a similar choice between two alternatives: a continuous version (20a) and a discontinuous one (20b).

(20) a. O jakou soutěž se jedná?
about what competition REFL is talked about
'What kind of competition is it?'

b. O jakou se jedná soutěž?
about what REFL is talked about competition
'What kind of competition is it?'

The interrogative or relative item can be embedded, as shown in (21a) and (21b). There seems to be a condition that the expression is a dependent of the governing noun in the final position; this condition is not satisfied in (21c).[18] Finally, (21d) shows that the expression need not include an interrogative/relative item, although in such a case the expression must be stressed in order to improve its acceptability.

(21) a. O jak dotovanou soutěž se jedná?
about how financed competition REFL is talked about
'How financed competition is it?'

b. O jak dotovanou se jedná soutěž?
about how financed REFL is talked about competition

c. *O jak se jedná dotovanou soutěž?
about how REFL is talked about financed competition

d. ?O velmi dobře dotovanou se jedná soutěž.
about very well financed REFL is talked about competition
'It is a very well financed competition.'

Additionally, the split PP can be subject to unbounded dependency, as in (22).[19]

(22) O jakou sis myslela, že se jedná soutěž?
about what REFL+AUX-2SG thought-FEM that REFL is talked about competition
'What kind of competition did you think it was?'

Solutions for similar examples from Polish and Serbo-Croatian have been presented by Kupść (2000, §2.4.2) and Penn (1999). In the Polish example of Kupść (2000) (*w dużym mieszka domu*, lit. 'in large she lives house'), the preposition is compacted with the following item (an adjective) only when the whole PP becomes a part of the clause and the noun is free to be ordered independently. Penn (1999) uses a principle applying to signs for NPs and PPs with

[18] The unacceptability of (21c) was observed by Karel Oliva (p.c.).

[19] Sentences where the noun is more deeply embedded are still grammatical, but it is difficult to find some which do not sound awkward.

disjunctive statements, which compact the domain objects of the phrase (i.e., our *s-nodes*) either to *pre-cf* (pre-clitic field) or to *rf* (rest field) in the clause. According to the third option the first *prosodic word*[20] compacts to *pre-cf* and the rest to *post-cf* (post-clitic field) of the next higher region, a matrix clause or an embedded finite clause.

As the present proposal concerning surface order is founded on Penn's approach, I will consider only his solution, which can be adopted with a few modifications. The first point is due to the flat derivation structure. The *s-list* item following the preposition should be available for compaction with the preposition. This condition is satisfied if either no region is defined which consists of a preposition and a nominal group, or if such a region cannot be built, as in our case, where the flat derivation structure for prepositional groups does not allow for compacting a *noun-fld* with a preceding preposition, except when the noun is bare. Such a region can only be formed as an option. On the other hand, because of the flat structure, compaction of a final part of the nominal group does not prevent compaction of the preposition with an initial part into *pre-cl-fld*.

The second point concerns the difference between Czech and Serbo-Croation: in Czech, the equivalent of *post-cf* (post-clitic field) seems to be rather the clause-final field, if any.

The final point concerns the role of prosody. Examples in (21) suggest that in Czech the expression following the preposition is a syntactic rather than a prosodic unit. Thus, if there is any involvement of prosodic factors here at all, it is restricted to the proclitical position of prepositions.[21]

The solution should be in line with the approach pursued so far, which was based solely on defining the setup of regions. Similarly as in the case of discontinuous adjectival group above, it is possible to define two alternative regions: *pre-cl-fld* as the PP-initial region or *pp-fld* as the region compacting preposition with the nominal group. With the initial region compacting into *pre-cl-fld*, the remainder part can be assigned *fin-fld*, which would position the noun at the end of the clause by surface-level rules, or *noun-fld*, which would compact to *rest-fld* and determine its position by Deep/Surface Order Principle. If the noun is the rightmost dependent of the verb, the Identical Order constraint would place its surface counterpart correctly as the last *rest-field*.

If the position of the remainder part of PP is not fixed to *fin-fld*,[22] the constraint on PP compaction can be informally described as follows: If there is a sign headed by *noun* with a non-head daughter whose singleton *s-list* contains an *s-node* corresponding to a preposition, and with another non-head daughter and possibly a rest of the nominal group, then either these two daughters compact to *pre-cl-fld* and the rest of the mother's *s-list* compacts to *noun-fld*, or the second daughter compacts with the rest to *noun-fld*. The *pre-cl-fld* is free to compact within the current finite clause or within a higher clause.[23]

[20] Prosodic word is identified by using a parallel structure for prosodic constituency, with a separate list of items – 'domain objects'. Domain objects can be compacted into prosodic constituents which correspond to prosodic words.

[21] Recall that no separate representation of prosodic structure is assumed here. Admittedly, this makes it difficult to distinguish prosodic and syntactic factors responsible for a specific phenomenon.

[22] This solution could be supported by the marginal acceptability of (i), with the noun being positioned non-finally, which might suggest that a surface-level rule insisting on the final position is not involved here.

(i) ??O jakou se soutěž jedná?
about what REFL involves competition

[23] See again Rosen (2001) for details.

9 Conclusion

The results show that FGD may be combined with RSRL and the framework can be applied to a range of word order phenomena in Czech. Using the framework, a way to define the relation between deep and surface order and its interaction with surface-level constraints has been proposed.

The word order principles of Vilém Mathesius were shown to be compatible with the formalism. The principal role of topic-focus articulation in determining surface order and prosody, as assumed in FGD, has been embodied in constraints interacting with other ordering constraints.

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, using an approach known from linearization grammars with topological fields.

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. This is necessary in order to verify the descriptions, but also to assess chances of further development.

References

Borota, Jan. 1990. *A Procedure of Syntactic Synthesis of Czech*, volume XVIII of *Explizite Beschreibung der Sprache und automatische Textbearbeitung*. Charles University, Faculty of Mathematics and Physics, Prague.

Colmerauer, Alain. 1970. Les Systémes Q ou un Formalisme pour Analyser at Synthétiser des Phrases sur Ordinateur. Internal publication 43, Université de Montréal.

Kathol, Andreas. 1995. *Linearization-based German Syntax*. Dissertation, submitted to the Ohio State University.

Kirschner, Zdeněk. 1982. *A Dependency-Based Analysis of English for the Purpose of Machine Translation*, volume IX of *Explizite Beschreibung der Sprache und automatische Textbearbeitung*. Charles University, Faculty of Mathematics and Physics, Prague.

Kuboň, Vladislav, Tomáš Holan, and Martin Plátek. 1997. A Grammar Checker for Czech. FAL Technical Report TR-1997-02, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.

Kupść, Anna. 2000. *An HPSG Grammar of Polish Clitics*. Dissertation, submitted to the Institute of Computer Science at the Polish Acadamy of Sciences, and to the Université Paris 7.

Mathesius, Vilém. 1939. O tak zvaném aktuálním členění větném. *Slovo a slovesnost*, (5):171–174.

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

Oliva, Karel. 1989. *A Parser for Czech Implemented in Systems Q*, volume XVI of *Explizite Beschreibung der Sprache und automatische Textbearbeitung*. Charles University, Faculty of Mathematics and Physics, Prague.

Panevová, Jarmila. 1979. *Transducing Components of Functional Generative Description 1: From Tectogrammatics to Morphemics*, volume IV of *Explizite Beschreibung der Sprache und automatische Textbearbeitung*. Charles University, Faculty of Mathematics and Physics, Prague.

Panevová, Jarmila. 1982. Random Generation of Czech Sentences. In *Proceedings of COLING 82*, pages 295–300, Praha.

Penn, Gerald. 1999. Linearization and WH-extraction in HPSG: Evidence from Serbo-Croation. In Robert D. Borsley and Adam Przepiórkowski, editors, *Slavic in HPSG*, Studies in constraint-based lexicalism. CSLI Publications, Stanford, pages 149–182.

Plátek, Martin, Jiří Sgall, and Petr Sgall. 1984. A Dependency Base for a Linguistic Description. In Petr Sgall, editor, *Contributions to Functional Syntax, Semantics, and Language Comprehension*. Academia, Praha, pages 63–97.

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

Reape, Mike. 1994. Domain Union and Word Order Variation in German. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in CSLI Lecture Notes. CSLI Publications, Stanford, USA, pages 151–198.

Richter, Frank. 2000. *A Mathematical Formalism for Linguistic Theories with an Application in Head-driven Phrase Structure Grammar*. Dissertation, submitted to the Neuphilologische Fakultät of the Universität Tübingen, Version of April 28th.

Rosen, Alexandr. 2001. *A constraint-based approach to dependency syntax applied to some issues of Czech word order*. Ph.D. thesis, Faculty of Philosophy, Charles University, Prague.

Sag, Ivan A. 1995. Taking Performance Seriously. Unpublished manuscript.

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.

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.