# Condition of Projectivity in the Underlying Dependency Structures

**Kateřina Veselá, Jiří Havelka, Eva Hajičová**
Center for Computational linguistics
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, 11800 Prague 1, Czech Republic
{vesela,havelka,hajicova}@ckl.mff.cuni.cz

## Abstract

The claim made in this paper is that in a formal description of language, it is possible and useful to work with dependency-based underlying representations of sentences (tectogrammatical representations) meeting the condition of projectivity. The reasons for the inclusion of this condition into the definition of the tectogrammatical representations are both formally and empirically sound (Section 1). An analysis of the material offered by the Prague Dependency Treebank with annotations of the underlying syntactic structure of sentences (described in Section 2) has led to an interesting classification of non-projective constructions in Czech (Section 3). It documents that most (types of) constructions that appear to be non-projective in the surface shape of sentences can be described by means of projective trees. The realization of the surface word order (with the use of movement rules) is then relegated to the morphemic level, where the representation of the sentence has the shape of a string rather than a tree.

## 1 Condition of projectivity and tectogrammatical representations of sentences

It may be assumed that in most different languages the prototypical means expressing grammatical values are morphemes (endings, affixes, function words). On the other hand, word order (with intonation) expresses first of all the Topic-Focus Articulation (TFA); this holds also in English, as will be illustrated. The grammatical values themselves are (i) values of morphological categories (number, definiteness, tense, aspect, modality, etc.), and (ii) syntactic relations. The latter may be specified on the basis of valency, as is the case e.g. with FrameNet (Fillmore et al., 2003). As was shown already by Robinson (1970), a valency based approach makes it possible to use dependency syntax (with the relations between a head and its modifiers) as the backbone for a specification of sentence representations. The valency frame of a lexical unit specifies its obligatory and optional dependents, i.e. arguments (such as Actor, Objective, Addressee, Origin, Effect) and adjuncts (Means, Manner, Locative, several directional and temporal ones, and so on). In the descriptive framework of Functional Generative Description (FGD), elaborated in Prague (see Sgall et al. (1986)), the above mentioned set of dependency relations is used to describe the underlying structure, which can serve as an appropriate input to semantic(-pragmatic) interpretation.

The dependency trees, which constitute the prototypical underlying (tectogrammatical) representations of sentences in FGD, meet the condition of projectivity. The condition of projectivity of a rooted tree has been defined in several ways, some of which have been shown by Marcus (1965) to be equivalent, cf. also the concept of adjacency in Hudson (1984).

### 1.1 Formal definition of projectivity as a property of dependency tree structures

We present a definition of projectivity and an algorithm for projectivizing/testing the projectivity of a (sub)tree. (In devising this approach, we were motivated by the practical purposes of the annotation of TFA within the Prague Dependency Treebank, PDT.)

**Definition** A subtree $S$ of a rooted dependency tree $T$ is *projective* iff for all nodes $a$, $b$ and $c$ of the subtree $S$ the condition (P) holds:

$$\begin{aligned}&\left(b \downarrow a \ \&\ b < a \ \&\ c \downarrow\downarrow b \Longrightarrow c < a\right)\\ &\lor \left(b \downarrow a \ \&\ b > a \ \&\ c \downarrow\downarrow b \Longrightarrow c > a\right) . \qquad \text{(P)}\end{aligned}$$

(Here $b \downarrow a$ means that $b$ is immediately dependent on $a$, $c \downarrow\downarrow d$ means that $c$ is subordinated to $d$—the relation of subordination $\downarrow\downarrow$ is the irreflexive transitive closure of the relation of im-

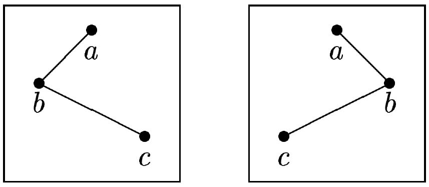

Figure 1: Forbidden configurations

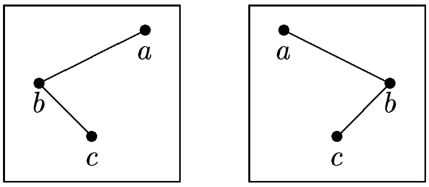

Figure 2: Configurations projectivized

mediate dependency ↓. The symbols <, > denote the relation of linear ordering on the nodes corresponding to the underlying word order.)

A subtree is called *non-projective* iff it does not satisfy condition (P).

To make the notion of projectivity more tangible, in Figure 1 we present the configurations (subtrees of a dependency tree) forbidden by the Definition (lines represent immediate dependency and nodes are ordered from left to right according to the linear ordering on nodes). It is easy to prove that in condition (P) it is enough to work with immediate dependency only, so for a subtree to be projective it suffices to check configurations where three nodes form a chain in the relation of immediate dependency. The edge between the two lower nodes in such a non-projective configuration will be called *non-projective*. For a (sub)tree to be projective, neither of the configurations in Figure 1 may appear in it.

Our definition of projectivity is equivalent to other definitions when applied to the whole dependency tree—then the forbidden configurations cannot appear anywhere in the tree (cf. Sgall et al. (1986), p. 152, and works cited above).

The definition of projectivity presented above lends itself readily to algorithmization. It can be used not only for checking whether a particular subtree is projective, it can also be easily adapted to a procedure for projectivizing the subtree (i.e. transforming the potentially non-projective subtree into a projective one by rearranging its nodes in the linear ordering).

We give a simplified imperative pseudo-code of a recursive version of the algorithm for projectivizing a subtree:

```
procedure Projectivize(node) {
  foreach child in node->children do
    Projectivize(child);
  Rearrange_subtree(node);
}
```

Let us describe the algorithm in more detail: the parameter of the procedure is the root of the subtree we want to projectivize; the procedure first recursively projectivizes the subtrees of nodes immediately depending on the current node (its "children"), and then rearranges the subtree of the current node in such a way that the relative order of the current node and its children remains unaltered, but the whole subtrees are moved right before and after the current node in the linear ordering. In other words, nodes in the subtree to be projectivized are moved as closely to their parent node as possible preserving the relative ordering of all nodes with respect to their parent nodes. (For lack of space we do not give details of data stuctures used for representing rooted dependency trees, but we hope that the exposition is clear enough to be easily understandable.)

Figure 2 shows the result of projectivizing the forbidden configurations from Figure 1.

For checking the projectivity of a subtree using the algorithm, it suffices to projectivize a copy of the subtree and compare it with the original subtree.

The complexity of the algorithm depends on the data representation of rooted dependency trees and the usage of auxiliary data stuctures. If the recursion is transformed to iteration and an auxiliary data structure is used, we can get linear complexity with respect to the number of nodes of the input (sub)tree.

### 1.2 Formal and empirical substantiation of the condition of projectivity

The condition of projectivity is a very strong restriction laid on the tectogrammatical representations, but we believe there are very good reasons to postulate it, both formal and empirical. From the formal side, the more restricted is a formal framework the more interesting it is. In addition, projective rooted trees allow for a straightforward one-to-one linearisation. From the linguistic point of view, such a representation makes it possible to interpret the left-to-right order of nodes of the tree as the basic (underlying) word order and thus to capture the description of the TFA of the sentences at this level. TFA as a semantically relevant opposition can be then defined on the basis of deep word order (or, more precisely, of the opposition of contextual boundness and non-boundness, see Section 2.2 below), and Topic and Focus can be described as continuous parts of the sentence.

### 1.3 Projectivity and deviations from it in theoretical description

In FGD, we proceed from the projective core with tectogrammatical representations (TRs) treated as projective rooted trees and view the deviations from projectivity (as well as many other marked cases and exceptions) as differences between underlying and morphemic structures. Most (types of) non-projective constructions can be described by means of projective trees, leaving the realization of the surface word order to the morphemic level, where the representation of the sentence has the shape of a string rather than a tree (possibilities of a specification of such a transition are illustrated by examples of movement rules in Hajičová and Sgall (2003)). Deviations of all kinds are determined by contextual restrictions (definable by lists, e.g. a list of quasi-modal predicates), by specific indices in node labels (contrast) and by specific behavior of certain items (lists, analogy, additional rules, e.g. those of word-order shifts), see Hajičová et al. (in press) and Section 3 below.

## 2 Annotation of the underlying sentence structure in the Prague Dependency Treebank

The Prague Dependency Treebank (containing 100,000 sentences of Czech text samples; a subcollection of the Czech National Corpus) is manually annotated on three levels: morphological layer (lowest), analytical layer (intermediate)–surface syntactic annotation, and tectogrammatical layer (highest)–underlying structure of the sentence. In this paper we are concerned with the tectogrammatical layer of annotation, which apart from dependency relations captures also the TFA of sentences.

### 2.1 Dependency relations

As already mentioned in Section 1, we work with two sets of dependency relations, namely

1. arguments, i.e. Actor, Objective, Addressee, Origin, Effect, which, in the prototypical case, are obligatory with their heads, and

2. adjuncts (Means, Manner, Locative, several directional and temporal ones, as well as Cause, Condition, Substitution, Amount, Norm, Criterion, etc.), which prototypically are optional, although in certain cases they may be found to be obligatory (e.g. in *to arrive at a place*).

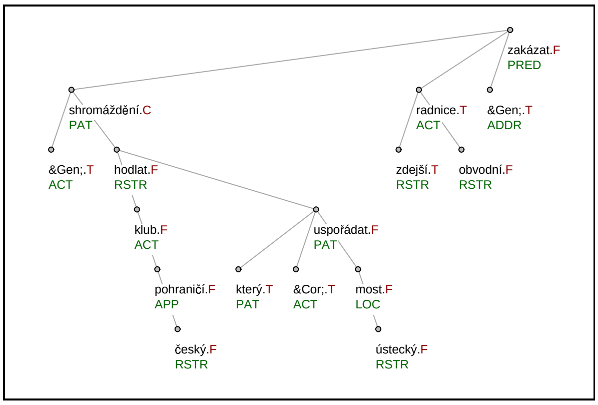

Figure 3: Tectogrammatical tree for the example sentence (1) *Shromáždění, které hodlá uspořádat Klub českého pohraničí na ústeckém mostě, zdejší obvodní radnice zakázala.*

Operational criteria for distinguishing arguments from adjuncts and obligatory complementations from optional ones were discussed by Panevová (1978), taking into account that in certain cases an obligatory dependent may be deletable in the surface form of the sentence (as *to behave somehow*, *to last for how long*).

Let us illustrate some of the mentioned phenomena by the example 1, in which, in Czech, the Actor (identical with the Subject in the active construction) follows the verb, the latter being preceded by the Objective; in the English counterpart, this is rendered by passivization, making the SVO order possible. In the subordinated (relative) clause, the surface (morphemic) word order differs from the underlying order to such a degree that the condition of projectivity appears not to be met by the outer form of the sentence (cf. the tectogrammatical representation in Figure 3; the syntactic labels are transparent, the labels T, C and F stand for contextually bound, contrastively contextually bound, and contextually non-bound, respectively).

(1) *Shromáždění, které hodlá uspořádat*
LIT. assembly$_{\mathrm{Acc}}$, which intends to-organize
*Klub českého pohraničí na ústeckém*
Club of-Czech borderland on Ústí
*mostě, zdejší obvodní radnice zakázala.*
bridge, local district Council prohibited.
TR. The assembly, which the Club of Czech Borderland intends to organize on the Ústí bridge, was prohibited by the local district Council.

### 2.2 Topic-Focus Articulation

In PDT, the sentence structure is specified not only in its "static" layers (as a "predicate-argument" pattern), but also in its information structure (in its TFA). It has been argued for long decades in the context of many linguistic trends (cf. Hajičová et al. (1998)) that sentences exhibiting a different TFA differ systematically in their structures, not just in their contextual combinability. In Czech linguistics, this view was clearly formulated already by Mathesius (1929). Moreover, examples such as (2)(a) and (b) or (3)(a) and (b) document that such sentence pairs in some cases also differ in their semantics, even in their truth conditions.

(2)(a) She was writing her dissertation on the weekends.
(b) On the weekends, she was writing her dissertation.

(3)(a) English is spoken in the Shetlands.
(b) In the Shetlands, English is spoken.

As has been discussed in the context of FGD, the individual nodes of a dependency tree can be characterized as being either contextually bound or non-bound. Among the criteria useful to distinguish between the two members of this opposition, there are different kinds of pitch, as well as, in languages such as Czech, the difference between weak and strong forms of certain pronouns (e.g. *ho* vs. *jeho* 'him'). This view is now being checked on examples from PDT.

The left-to right order of the nodes in the TR, i.e. the scale of communicativre dynamism of Firbas (1992) starts with Topic proper (the least dynamic item) and proceeds to Focus proper (most dynamic). The surface (i.e. morphemic) word orderThis order differs from the underlying one e.g. in cases determined by such shallow rules like "adjective before noun", or in placing the carrier of the (typically falling) sentence stress in a marked (not clause-final) position. Within Focus, the order of the nodes in a TR is fixed, corresponding to the systemic ordering; for a detailed examination of this ordering in German, Czech, and partly also English, see Sgall et al. (1995).

## 3 Deviations from projectivity due to TFA

Non-projective constructions in the surface realization of a sentence can arise under these two conditions: the dependency tree of the sentence contains at least one indirect subordination (i.e. two nodes where one is subordinated but not immediately dependent on the other), and one of the two nodes is moved into a non-projective position (i.e. it brings about a non-projective configuration in the dependency tree).

In Czech, the following types of nodes can appear in an indirectly subordinated position:

1. attributes of participants of the sentence structure, and nodes subordinated to them;
2. complements of infinitives, and nodes subordinated to them;
3. complements of nominal parts of compound predicates, and nodes subordinated to them;
4. complements of predicates of subordinated clauses, and nodes subordinated to them.

Movements of nodes into non-projective positions arise either due to word-order rules of the given language (in our case Czech), or due to TFA. We consider word-order rules as phenomena belonging to the analytical (morphological) layer of the sentence, and therefore we are not concerned with such types of deviations from projectivity. On the other hand, TFA as a semantically relevant feature of the sentence is in our view a component of the underlying sentence structure, and as such a key issue in the study of the conditions for deviations from projectivity in Czech. A description of the types of deviations caused by TFA follows. (We concentrate on declarative sentences, the main reason being that the information structure of questions has not yet been sufficiently elaborated upon.)

### 3.1 Classification

Our classification of the deviations from projectivity is based mainly on the morpho-syntactic features of nodes connected by a non-projective dependency edge.

### 3.1.1 Constructions with attributes

Two types of deviations from projectivity with a nominal node and its attribute connected by a non-projective edge can be distinguished:

**1A** – the attribute is non-projectively moved to the left

(4) *Studené mám pivo nejradši.*
Lit. Cold I-have beer the-most.
Tr. As for beer, I like it best cold.

(5) *O dietě jsem napsal knihu.*
Lit. Aboud diet I-am written a-book.
Tr. As for diet, I have written a book about it.

**1B** – the node governing the attribute is non-projectively moved to the left

(6) *Sportovec je Pavel dobrý.*
Lit. Sportsman is Paul good.
Tr. As for sport, Paul is good at it.

(7) *Těch stromů porazili třicet.*
Lit. Those trees they-felled thirty.
Tr. As for the trees, they felled thirty of them.

### 3.1.2 Constructions with infinitives

There are two types of non-projective edges between an infinitive and its complement.

**2A** – the complement of the infinitive is non-projectively moved to the left

(8) *Karla jsme zamýšleli poslat*
Lit. Charles we-are intended to-send
*do Ameriky.*
to America.
Tr. As for Charles, we intended to send him to America.

**2B** – the infinitive is moved to the left

(9) *Pozvat jsem se rozhodl jen rodinu.*
Lit. To-invite I-am refl. decided only family.
Tr. Speaking of invitation, I have decided to invite only the family members.

### 3.1.3 Compound predicates

Compound predicates are formed by a delexicalized verb and a typically deverbal noun, and are usually synonymous with a single verb. For example *prokázat úctu* 'to show respect' (equivalent to *uctít* 'to honour').

Again, there are two types of non-projective constructions with compound predicates.

**3A** – the valency complement of the nominal part of the compound predicate is non-projectively moved to the left

(10) *K Martinovi cítil úctu.*
Lit. To Martin he-felt respect.
Tr. As for Martin, he felt respect for him.

**3B** – the nominal part of the compound predicate is moved to the left

(11) *Zájem jevil především*
Lit. Interest he-expressed mostly
*o matematiku.*
about mathematics.
Tr. He expressed interest mostly in mathematics.

## 3.2 Factors causing deviations from projectivity

In the above listed types of non-projective constructions it is necessary to establish the conditions for deviations from projectivity and to further specify and describe the above mentioned types. Since issues relevant for the presence of non-projective constructions are general and do not apply to single types of the constructions, we describe them separately and relate them to the individual types of non-projective constructions. If a deeper embedded node is contextually bound, it can either stay in the same position as in the underlying word order, or it can move to the left so as to become a part of the Topic in the surface realization of the sentence.

### 3.2.1 Motivation for non-projective constructions

All movements of nodes considered in our study are movements to the left from a position in the underlying word order. One of the most important factors causing movement of a node to the initial position in the surface word order is the relation of "contrastive contextual boundness". We use the expression "contrastive Topic" for such a node (denoted in the examples by C), which is characterized by several specific features: although it lies in the Topic part of a sentence, it is necessary to use a strong morfological form if the contrastive node is represented by a pronoun (cf. ex. 12) and it can carry the typical rising "contrastive" stress; semantically, it refers to a choice from a set of alternatives and it can be in contrastive relation to some part of the preceding context (cf. ex. 13).

(12) *Jemu.*C *jsem to neřekl (, ale tobě ano).*
Lit. Him I-am it not-said (, but you yes).
Tr. I haven't said it to him (, but I have said it to you).

(13) (*Jirku jsem neviděl, ale*) *Marii.*C
LIT. (George I-am not-seen, but) Mary
*jsem viděl.*
I-am seen.
TR. I have not seen George, but I have seen Mary.

A contrastive node has quite a strong tendency to stand in the initial position in the surface word order, no matter how deep it is embedded in the underlying structure of a sentence. In cases corresponding to types 1A (ex. 4), 1B (ex. 6), 2B and 3B, a non-projective word-order variant is acceptable only if the non-projective left-moved node is contrastively contextually bound. The utterances *Sportovec je Pavel dobrý* and *Pavel je dobrý sportovec* are realizations of two different underlying structures—in the first case the node *sportovec* is contrastively bound and in the second one it is contextually non-bound.

However, in cases corresponding to types 1A (ex. 5), 1B (ex. 7), 2A and 3A, the non-projective left-moved node can be non-contrastively contextually bound. Such nodes skip over specific kinds of constructions which behave (from the TFA point of view) like a single unit of the underlying structure of a sentence. For this very reason these non-projective surface realizations seem to be the non-marked variants (the utterance *Včera jsme se Karla rozhodli poslat do Ameriky* assumes that the node *Karel* is contextually bound, whereas *Včera jsme se rozhodli poslat Karla do Ameriky* assumes *Karel* to be contextually non-bound). The main grammatical factor bringing about non-projective word-order variants is the compound form of the predicate itself, supported by some other grammatical and semantic factors.

### 3.2.2 Specific features causing non-projective constructions

In this subsection, we would like to describe some semantic and grammatical aspects which in our view constitute conditions causing non-projective constructions.

#### Quasi-modal and quasi-phase verbs

A very important feature of compound-verb constructions with a dependent infinitive is the modal or phase aspect of the governing verb. We call these verbs "quasi-modal" and "quasi-phase", because their meaning consists of more semantic features than just the modal or phase one (e.g. verbs *want*, *decide*, *start*, and some others). If a modal or a phase feature is needed to be added to the meaning of a verb, compound-verb constructions with an infinitival (e.g. *he decided to work at st.*) or nominal dependent (e.g. *to improve the relationship with sb.*) are used. Modal and phase semantic features can be both added to the meaning of a verb—this gives rise to complicated constructions, such as *he wanted to start to work at st.*

#### Semantic feature of quantification

The type 1B (ex. 7) differs from other subtypes of 1, because in this case the non-projective left-moved node does not have to be contrastively bound. This seems to be caused by the fact that the governing node (parent of the non-projective left-moved node) contains the semantic feature of quantification. Such nodes are mostly expressed by numerals or adverbial expressions like *much* or *enough.*

#### Valency of nouns

In the case of verbonominal predicates, the left-moved non-projective node is a dependent of the nominal part of the predicate. Most often it is a complement of a deverbal noun (e.g. *zájem o* 'interest in st.', *úcta k* 'respect for sb.'), but there are also nouns requiring such a complement which are not deverbative (e.g. *kniha o* 'book about st.', *příklad na* 'example of st.'). The dislocation to the left need not be motivated by contrastive boundness (e.g. *Před lety jsem o Komenském publikoval článek*—the node *Komenský* is a complement of the noun *článek* and it is non-projectively moved to the left).

#### Grammatical relation of control

Most constructions with infinitives comply with the grammatical rule called "control"—the subject of the action expressed by an infinitive is identical with one of the complements of the main verb (e.g. *Pavel o té věci slíbil pomlčet* 'Paul promised to be silent about the issue'—the subject of *pomlčet* 'be silent' is *Pavel*, because it has to be identical with the actor of the main verb *slíbit* 'promise'). We hope that the presence of the relation of control will help us to define the set of verbs which (as nodes governing infinitives) participate in non-projective constructions, because the modal and phase semantic features are not sufficient to define this set of a verbs. Also in these cases the non-projective left-moved node does not have to be contrastively bound.

## 4 Treatment of non-projective constructions in PDT

### 4.1 Movement of contrastive Topic to the initial position

The facts described in Section 3.2.1 above demonstrate that there are some cases of deviations from projectivity in Czech word order which require a non-projective left-moved node to be contrastive. For such cases (types 1A, 1B, 2B and 3B) it can be therefore supposed that if there is a more deeply embedded contrastively bound node, it generally moves to the initial surface word-order position in the clause. In the tectogrammatical annotation, such constructions are projectivized and we mark the contrastive node with a special value of contrastive contextual boudness C.

### 4.2 Compound predicates and constructions with an infinitive

For constructions of types 2A and 3A it is evident that the compound construction consisting of a verb and an infinitive or a deverbal noun behaves (from the TFA point of view) as a single unit of the underlying sentence structure. It has to be further checked whether the two words form a single node on the tectogrammatical layer or whether their relation has some specific character unlike the other dependency relations. The nominal parts of compound predicates are anotated by a special functor CPHR, which helps us to delimit the set of cases causing non-projective realizations of sentences with verbonominal predicates. As for constructions with infinitives, it is fundamental to determine modal and phase semantic features and the grammatical relation of control causing non-projective constructions.

### 4.3 Other types of non-projective constructions

The annotation of non-projective word-order variants is not defined yet for cases with quantifying expressions in Focus of the sentence (see ex. 7) and for cases with complements of non-deverbative nouns (see ex. 5). In future we envisage to define lists of such cases based on semantic and morphological features, but first it is necessary not only to delimit, but also to explain why non-projective constructions arise in these cases.

## 5 Acknowledgements

The research reported in this paper has been carried out under the project LN00A063 supported by the Czech Ministry of Education.

## References

Charles J. Fillmore, Christopher R. Robinson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

Jan Firbas. 1992. *Fuctional Sentence Perspective in Written and Spoken Communication.* Cambridge–London University, Cambridge/London.

Eva Hajičová and Petr Sgall. 2003. Dependency Syntax in Functional Generative Description. In *Dependenz und Valenz/Dependency and Valency*, volume 1, pages 570–592. Walter de Gruyter, Berlin–New York.

Eva Hajičová, Barbara Partee, and Petr Sgall. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content.* Kluwer Academic Publishers, Amsterdam, Netherlands.

Eva Hajičová, Jiří Havelka, Petr Sgall, Kateřina Veselá, and Daniel Zeman. (in press). Issues of Projectivity in the Prague Dependency Treebank. *Prague Bulletin of Mathematical Linguistics*, 81.

Richard Hudson. 1984. *Word Grammar.* Blackwell, Oxford.

Solomon Marcus. 1965. Sur la notion de projectivite. *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik*, 11:181–192.

Vilém Mathesius. 1929. Zur Satzperspektive im modernen Englisch. *Archiv für das Studium der neueren Sprachen und Literaturen*, 155:202–210.

Jarmila Panevová. 1978. Inner Participants and Free Adverbials. *Prague Studies in Mathematical Linguistics*, 6:227–254.

Jane J. Robinson. 1970. Dependency structures and transformational rules. *Language*, 46(2):259–285.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects.* Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Petr Sgall, Wolfgang Dressler, Oskar Pfeiffer, and Michael Půček. 1995. Experimental research on Systemic Ordering. *Theoretical Linguistics*, 21:197–239.