The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007

EDITORIAL BOARD

Published twice a year by Charles University in Prague

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: pbml@ufal.mff.cuni.cz

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007

CONTENTS

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007

EDITORIAL

As our regular readers have noticed, this issue of the Prague Bulletin of Mathematical Linguistics (PBML No. 87) appears in a new cover and format. After some hesitation, stemming from our conservative beliefs that a change in the outer appearance of a journal should not be made unless undoubtedly needed, we have arrived at the conclusion that our journal has acquired a considerable propagation, circulation and recognition and therefore it deserves a more modern dress.

However, we still preserve the somewhat historical name of the Bulletin: It came into existence in 1964, the same year when the first international meeting on computational linguistics in our country (and, to our knowledge and memory, one of the first in the whole of Europe) called Colloquium on Algebraic Linguistics was organized in Prague by our research group at Charles University. The name "algebraic" was chosen in order to be distinguished from "mathematical" in its interpretation as "quantitative", prevailing then in our country. However, we wanted the Bulletin to cover a broader field, including statistical and computational studies, as well as formal description of language. ('Computational' linguistics would sound a little bit outrageous as well as conceited, with regard to the relative unavailability of computers in our part of the world at that time).

From now on, the size of the Bulletin will be smaller (though not the number of pages), comparable with the usual size of international journals; the scope, however, will be preserved, and, as we all would wish, the quality of the contributions published will keep its raising standards.

To achieve this ambitious goal, we would like to invite all possible contributors to submit their articles from all domains of computational/mathematical linguistics, theoretical as well as applied, be they of a more or less linguistic, computational, computer science, or language technology orientation. All the manuscripts will be reviewed, as is usual nowadays, by two reviewers, and we will keep the rule that one of the reviewers should be a renowned specialist from abroad. For this purpose, the editorial board has been enlarged by prominent scholars of the field coming from different geographical areas as well as domain of interest.

We use the opportunity of being the local organizers of the 40th ACL Annual Conference in June 2007 in Prague to present this newly dressed PBML at this occasion and we would welcome your comments, suggestions, and first of all, your contributions for publication on our address pbml@ufal.mff.cuni.cz.

Eva Hajičová
Editor-in-Chief
hajicova@ufal.mff.cuni.cz

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 7–26

Towards a Formal Model for Functional Generative Description

Analysis by Reduction and Restarting Automata

Markéta Lopatková, Martin Plátek, Petr Sgall

Abstract

Functional Generative Description (FGD) is a dependency based descriptive system, which has been in development since the 1960s, see esp. Sgall et al. (1969). FGD was originally implemented as a generative procedure, but lately we have been interested in a declarative representation. The object of the present paper concerns the foundations of a *reduction system* which is more complex than a reduction system for a (shallow) syntactic analyzer, since it provides not only the possibility of checking the well-formedness of the (surface) analysis of a sentence, but its underlying (tectogrammatical in terms of FGD) representation as well. Such a reduction system makes it possible to define formally the *analysis* as well as the *synthesis* of a sentence.

We propose a new formal frame, namely a *4-level reduction system* for FGD, which is based on the notion of *simple restarting automata*, see Messerschmidt et al. (2006). This new approach mirrors straightforwardly the so-called *(multi-level) analysis by reduction*, an implicit method used for linguistic research – analysis by reduction allows for obtaining (surface and/or deep) (in) dependencies by the reductions of Czech sentences as well as for describing properly the complex word order of a free word order language, see Lopatková, Plátek, and Kuboň (2005).

1. Introduction

Functional Generative Description (FGD) is a dependency based system for Czech, which has been in development since the 1960s, see esp. Sgall et al. (1969); Sgall, Hajičová, and Panevová (1986). FGD may be of some interest for the description of most Slavic languages, since it is adapted to treat a high degree of *free word order*. It not only specifies surface structures of the given sentences, but also translates them into their underlying representations. These representations (called tectogrammatical representations, denoted TRs) are intended as an appropriate input for a procedure of semantico-pragmatic interpretation in the sense of

Please cite this article as: Markéta Lopatková, Martin Plátek, Petr Sgall, Towards a Formal Model for Functional Generative Description: Analysis by Reduction and Restarting Automata. The Prague Bulletin of Mathematical Linguistics No. 87, 2007, 7–26.

intensional semantics, see Hajičová, Partee, and Sgall (1998). Since TRs are, at least in principle, disambiguated, it is possible to understand them as rendering linguistic (literal) meaning (whereas figurative meaning, specification of reference and other aspects belong to individual steps of the interpretation).

FGD has been implemented as a generative procedure by a sequential composition of pushdown automata, see Sgall et al. (1969); Plátek and Sgall (1978). Lately, as documented e.g. in Petkevič (1995), we have been interested in the formalization of FGD designed in a declarative way. In the present paper we want to formulate a formal framework for the procedure of checking the appropriateness and completeness of a description of a language in the context of FGD. The first step in this direction was introduced in Plátek (1982), where the formalization by a sequence of translation schemes is interpreted as an analytical system, and as a generative system as well. Moreover, requirements for a formal system describing a natural language L have been formulated – such a system should capture the following issues:

1. The set of correct sentences of the language L, denoted by LC.
2. The formal language LM representing all possible tectogrammatical representations (TRs) of sentences in L.
3. The relation SH between LC and LM describing the ambiguity and the synonymy of L.
4. The set of the correct structural descriptions SD representing in a structural way all possible TRs of sentences in L as dependency-based structures (*dependency trees*).

We propose here a new formal frame for checking FGD linguistic descriptions, based on *restarting automata*, see e.g. Otto (2006); Messerschmidt et al. (2006). We fully consider the first three requirements, i.e. LC, LM and SH. The fourth one is not formally treated here.

The main contribution of the new approach consists in the fact that it mirrors straightforwardly the so-called *analysis by reduction*. Analysis by reduction allows for obtaining (in)dependencies by the *correct reductions* of Czech sentences as well as for describing properly the complex word-order variants of a language with a high degree of 'free' word order, see Lopatková, Plátek, and Kuboň (2005). During the analysis by reduction, a (disambiguated) input string is processed, i.e. a string of tokens (word forms and punctuation marks) enriched with metalanguage categories from all linguistic layers encoded in the sentence. Analysis by reduction consists of stepwise correct reductions of the sentence; roughly speaking, the input sentence is simplified until the so called *core predicative structure* of the sentence is reached – section 2.1 provides a brief characterization of analysis by reduction.

Example: The example presented in Fig. 1 outlines the form of the input for analysis by reduction used in this paper, demonstrated on the sentence (1):

(1) *Přišel domů pozdě.*
[came home late]
E. He came home late.

There are four (sub)vocabularies Σ_0, Σ_1, Σ_2, Σ_3, each subvocabulary Σ_i represents the

			[on].ACT
Přišel	*m-přijít*.VpYS-	Pred	*t-přijít*.PRED.Frame1.ind-ant
domů	*m-domů*.Db- - -	Adv	*t-domů*.DIR3
pozdě	*m-pozdě*.Dg- - -	Adv	*t-pozdě*.TWHEN
.	..Z: - - -	AuxK	

Figure 1. A sample input structure for analysis by reduction for sentence (1).

corresponding layer of language description in FGD, namely:[1]

- Σ_0 is the set of Czech written *word-forms* and *punctuation marks* (tokens in the sequel), it is the vocabulary for the language LC from the request 1 above;
- Σ_1 represents the *morphemic layer* of FGD, namely morphological lemma and tag for each token;
- Σ_2 describes surface syntactic functions (as e.g. Subject, Object, Predicate);[2]
- Σ_3 is the vocabulary of the *tectogrammatical layer* of FGD describing esp. 'deep' roles, valency frame for frame evoking words, and meaning of morphological categories.

That means that the automaton has an access to all the information encoded in the processed sentence (as well as a human reader/linguist has all the information for his/her analysis).

In Section 2 we address two basic linguistic phenomena, dependency (subsection 2.2) and word order (2.3), and show the process of the analysis by reduction on examples from Czech.

Now, let us briefly describe the type of restarting automaton that we use for modelling analysis by reduction for FGD (see Section 3). A 4-LRL-*automaton* M_{FGD} is a non-deterministic machine with a finite-state control Q, a finite characteristic vocabulary Σ, and a head (window of size 1) that works on a flexible tape. Automaton M_{FGD} performs:

- *move-right* and *move-left steps*, which change the state of M_{FGD} and shift the window one position to the right or to the left, respectively, and
- *delete steps*, which delete the content of the window, thus shortening the tape, change the state, and shift the window to the right neighbor of the symbol deleted.

At the right end of the tape, M_{FGD} either halts and *accepts* the input sentence, or it halts and *rejects*, or it *restarts*, that is, it places its window over the left end of the tape and reenters the initial state. It is required that before the first restart step and also between any two restart steps, M_{FGD} executes at least one delete operation.

The 4-LRL-automata can be also represented by a final set of so called metarules, see Messerschmidt et al. (2006), a declarative way of representation, which seems to be a very promising tool for natural language description.

[1] The first column in the figure contains symbols from a vocabulary Σ_0, the second one contains symbols from a vocabulary Σ_1 and so on, the convention for displaying examples is specified in Section 2.2.

[2] Note that the layer of surface syntax does not correspond to any layer present in the theoretical specification of FGD but rather to the auxiliary 'analytical' layer of the Prague Dependency Treebank, see Hajič (2005), which is technically useful for a maximal articulation of the process of analysis.

In order to model the analysis by reduction for (FGD) the 4-LRL-automaton M_{FGD} works with a complex characteristic vocabulary Σ that is composed from (sub)vocabularies $\Sigma_0, \cdots, \Sigma_3$.

The basic notion related to M_{FGD} is the notion of the language accepted by M_{FGD}, so called *characteristic language* $L_C(M_{FGD})$. In our approach, it is considered as a language that consists of all sentences from the surface language LC over alphabet Σ_0 enriched with metalanguage information from $\Sigma_1, \Sigma_2, \Sigma_3$. The tectogrammatical language LM as well as the relation SH can be extracted from $L_C(M_{FGD})$.

M_{FGD} was introduced with no ambitions to model directly the procedure of the sentence-generating in the human mind or of the procedure of understanding performed in the human mind. On the other hand, it has a straightforward ambition to model the observable behavior of a linguist performing *analysis by reduction* of Czech sentences on the blackboard or on a sheet of paper.

2. Analysis by reduction for FGD

In this section we focus on the analysis by reduction for Functional Generative Description. We address two basic linguistic phenomena, dependency (subsection 2.2) and word order (2.3), and illustrate the process of the analysis by reduction on examples from Czech.

2.1. Analysis by reduction

The analysis by reduction makes it possible to formulate the relationship between dependency and word order, see also Lopatková, Plátek, and Kuboň (2005). This approach is indispensable especially for modelling the syntactic structure of languages with a high degree of 'free' word order, where the dependency (predicate-argument) structure and word order are very loosely related. The restarting automaton M_{FGD} that models analysis by reduction for FGD is specified in detail in the Section 3.

The *analysis by reduction* is based on a stepwise simplification of a sentence – each step of analysis by reduction consists of deleting at least one word of the input sentence, see Lopatková, Plátek, and Kuboň (2005) for more details.[3] The following principles must be satisfied:

- preservation of syntactic correctness of the sentence;
- preservation of the lemmas and sets of morphological categories;
- preservation of the meanings/senses of the words in the sentence (represented e.g. as an entry in a (valency) lexicon);
- preservation of the 'completeness' of the sentence (in this text only valency complementations (i.e. its arguments/inner participants and those of its adjuncts/free modifications that are obligatory) of frame evoking lexical items must be preserved).

The analysis by reduction works on a sentence (string of tokens) enriched with metalanguage categories from all the layers of FGD – in addition to word forms and punctuation marks, it embraces also morphological, surface and tectogrammatical information.

[3] Here we work only with the deleting operation whereas in Lopatková, Plátek, and Kuboň (2005) the rewriting operation also is presupposed.

The input sentence is simplified until the so called *core predicative structure* of the sentence is reached. The core predicative structure consists of:

- the governing verb (predicate) of an independent verbal clause and its valency complementations, or
- the governing noun of an independent nominative clause and its valency complementations, e.g. *Názory čtenářů.* [Readers' opinions.], or
- the governing word of an independent vocative clause, e.g. *Jano!* [Jane!], or
- the governing node of an independent interjectional clause, e.g. *Pozor!* [Attention!].

2.2. Processing dependencies

Czech is a language with a high degree of so-called free word order. Naturally, (surface) sentences with permuted word order are not totally synonymous (as the word order primarily reflects the topic-focus articulation in Czech), but their grammaticality may not be affected and the dependency relations (as binary relations between governing and dependent lexical items) may be preserved regardless of the word order changes. This means that the identification of a governing lexical item and its particular complementations is not based primarily on their position in the sentence but rather on the possible order of their reductions.

There are two ways of processing dependencies during the analysis by reduction.

- Free modifications (i.e. adjuncts) that do not satisfy valency requirements of any lexical item in the sentence are deleted one after another, in an arbitrary order (sentence (2)).
- The so called reduction components (formed by words that must be reduced together to avoid non-grammaticality, i.e. incompleteness of tectogrammatical representation)[4] are processed 'en bloc' depending on their function in the sentence:
  - Either all members of the reduction component are reduced – this step is applied if the 'head' of the reduction component does not fulfill any valency requirements of any lexical item in the sentence (see sentences (3) and (5) below where the whole components represent optional adnominal free modifications).
  - Or (if the 'head' of the reduction component satisfies the valency frame of some lexical item):
    1. the item representing the 'head' is simplified – all the symbols apart from the functor[5] are deleted; the result of such a simplification can be understood as a zero lexical realization of the respective item, see sentence (4); and
    2. the complementation(s) of the 'head' of the reduction component is/are deleted.

Convention: For the sake of clarity we have adopted the following conventions for displaying examples:

[4] Typically, a reduction component is composed of a frame evoking lexical item together with its valency complementations, see Lopatková, Plátek, and Kuboň (2005). Let us stress here that a reduction component may constitute a discontinuous string.

[5] A functor is the label for syntactico-semantic relation holding between the respective item and its governing lexical item.

- Each column contains a symbol from one part of the (partitioned) vocabulary, that means information on one layer of FGD:[6]
  - the first column contains tokens,
  - the second column contains morphological lemmas (m-lemmas) and morphemic values (i.e. morphological categories),
  - the third column contains (surface) syntactic functions,
  - for autosemantic words,[7] the fourth column contains tectogrammatical lemmas (t-lemmas), functors, frame identifiers and other tectogrammatical categories (so called grammatemes).
- Each individual token and its metalanguage categories are located:
  - in one line if its surface word order position agrees with the deep word order (i.e. word order at the tectogrammatical layer), or the token has no 'separate' tectogrammatical representation (i.e. it is not an autosemantic word);
  - in two lines if its surface word order position disagrees with the deep word order:
    1. one line embraces the token, its m-lemma and morphemic values as well as its (surface) syntactic function, and
    2. the other line contains relevant tectogrammatical information (for autosemantic words).
- The top-down ordering of lines reflects the word order on the respective layer.

Such a two-dimensional convention allows for revealing both (i) a representation of a whole sentence on particular layers (individual columns for particular layers), including relevant word order (columns 1, 2, 3 reflects the surface word order whereas column 4 is organized according to deep word order), and (ii) information relevant for individual tokens (rows).

Let us illustrate the processing of dependencies on sentences (2), (3), (4) and (5).

Example:

(2) *Včera přišel domů pozdě.*
[yesterday came home late]
E. Yesterday he came home late.

The analysis by reduction starts with the input structure specified in Fig. 2 (see the convention above; the metalanguage categories are explained e.g. in Hajič, 2005).

It is obvious that an item of TR (an autosemantic word, see for Note 7) can have zero surface lexical realization (e.g. actor, ACT need not be realized, as Czech is a pro-drop language – the corresponding item is restored in the TR; also different kinds of ellipsis are possible). On the other hand, several word forms can constitute a single item of TR (as e.g. a prepositional group in sentence (3)).

Let us point out the difference between the two types of free modifications in the sentence, namely DIR3 (direction 'to_where') and TWHEN (temporal relation 'when'): (i) whereas the

[6] Here the standard annotation used in the Prague Dependency Treebank is used, see Hajič (2005).

[7] Function words have just functors or grammatemes as their tectogrammatical correlates that are assigned to their governing autosemantic words.

Včera	*m-včera*.Dg- - -	Adv	*t-včera*.TWHEN
			[on].ACT
přišel	*m-přijít*.VpYS-	Pred	*t-přijít*.PRED.Frame1.ind-ant
domů	*m-domů*.Db- - -	Adv	*t-domů*.DIR3
pozdě	*m-pozdě*.Dg- - -	Adv	*t-pozdě*.TWHEN
.	..Z: - - -	AuxK	

Figure 2. The input structure for sentence (2).

(2 steps) →

			[on].ACT
přišel	*m-přijít*.VpYS-	Pred	*t-přijít*.PRED.Frame1.ind-ant
domů	*m-domů*.Db- - -	Adv	*t-domů*.DIR3
.	..Z: - - -	AuxK	

Figure 3. The reduced structure – a core predicative structure for sentence (2).

valency complementation of direction DIR3 is considered to be obligatory for the verb *přijít* [to come] (the speaker as well as the listener must know this, see the dialogue test proposed in Panevová, 1974) and thus fills the relevant slot of the valency frame of the verb *přijít* [to come] (here marked by the label Frame1), (ii) the temporal relation TWHEN is an optional free modification (not belonging to the valency frame Frame1).

The first step of analysis by reduction consists in the deletion of one of the optional free modifications *včera* [yesterday] or *pozdě* [late].[8] These free modifications may be reduced in an arbitrary order, they are mutually independent, see Lopatková, Plátek, and Kuboň (2005). These two reduction steps result in the structure in Fig. 3.

Now, the sentence contains only one reduction component constituted by the finite verb and its valency complementations, i.e. its actor (expressed by a zero form of the pronoun) and its obligatory free modification DIR3 'to_where', [on] *přišel domů* [(he) came home]. This is a core predicative structure, thus the reduction ends successfully.[9]

Example: This example shows the reduction of the whole reduction component that consists of a dependent clause.

(3) *Petr včera přišel do školy, kterou loni postavil minulý starosta.*
[Peter yesterday came to school which last_year built previous mayor]
E. Yesterday Peter came to the school which was built last year by the previous mayor.

The input structure looks as in Fig. 4.

[8] More precisely, the tokens as well as all the metalanguage categories relevant for the particular lexical item are reduced, similarly in the sequel.

[9] Here we leave aside the problems of word order – this domain is briefly addressed in the following subsection.

Petr	*m-Petr*.NNMS1	Sb	*t-Petr*.ACT
včera	*m-včera*.Dg- - -	Adv	*t-včera*.TWHEN
přišel	*m-přijít*.VpYS-	Pred	*t-přijít*.PRED.Frame1.ind-ant
do	*m-do*.RR- - 2	AuxP	
školy	*m-škola*.NNFS2	Adv	*t-škola*.DIR3.basic
,	,.Z: - - -	AuxK	
kterou	*m-který*.P4FS4	Obj	*t-který*.PAT
loni	*m-loni*.Db- - -	Adv	*t-loni*.TWHEN
postavil	*m-postavit*.VpYS-	Atr	*t-postavit*.RSTR.Frame2.ind-ant
minulý	*m-minulý*.AAMS1	Atr	
starosta	*m-starosta*.NNMS1	Sb	*t-starosta*.ACT
			t-minulý.RSTR
.	..Z: - - -	AuxK	

Figure 4. The input structure for sentence (3).

(3 steps) →

Petr	*m-Petr*.NNMS1	Sb	*t-Petr*.ACT
přišel	*m-přijít*.VpYS-	Pred	*t-přijít*.PRED.Frame1.ind-ant
do	*m-do*.RR- - 2	AuxP	
školy	*m-škola*.NNFS2	Adv	*t-škola*.DIR3.basic
,	,.Z: - - -	AuxK	
kterou	*m-který*.P4FS4	Obj	*t-který*.PAT
postavil	*m-postavit*.VpYS-	Atr	*t-postavit*.RSTR.Frame2.ind-ant
starosta	*m-starosta*.NNMS1	Sb	*t-starosta*.ACT
.	..Z: - - -	AuxK	

Figure 5. The simplified structure for sentence (3).

In the first three steps, the three optional free modifications *včera, loni* and *minulý* [yesterday, last_year, previous] are deleted in arbitrary order, see Fig. 5.

Next, the whole component *kterou postavil starosta* [which the mayor built] consisting of the verb and its valency complementations is to be processed. As this component represents an optional adnominal free modification RSTR, it can be simply deleted without the loss of completeness.

After this step, only one reduction component *Petr přišel do školy* [Peter came to school] remains, see Fig. 6 which constitute a core predicative structure – the analysis by reduction ends successfully.

Example: Let us show an analysis of a sentence with a valency complementation realized as an infinitive form of the verb.

→

Petr	*m-Petr*.NNMS1	Sb	*t-Petr*.ACT
přišel	*m-přijít*.VpYS-	Pred	*t-přijít*.PRED.Frame1.ind-ant
do	*m-do*.RR- - 2	AuxP	
školy	*m-škola*.NNFS2	Adv	*t-škola*.DIR3.basic
.	..Z: - - -	AuxK	

Figure 6. The core predicative structure for sentence (3).

Petr	*m-Petr*.NNMS1	Sb	*t-Petr*.ACT
pomáhal	*m-pomáhat*.VpYS-	Pred	*t-pomáhat*.PRED.Frame1.ind-ant
Marii	*m-Marie*.NNFS3	Obj	*t-Marie*.ADDR
			[ona].ACT
uklízet	*m-uklízet*.Vf- - -	Adv	*t-uklízet*.PAT.Frame3
zahradu	*m-zahrada*.NNFS4	Obj	*t-zahrada*.PAT
.	..Z: - - -	AuxK	

Figure 7. The input structure for sentence (4).

(4) *Petr pomáhal Marii uklízet zahradu.*
[Peter helped Mary clean garden]
E. Peter helped Mary to clean the garden.

In this sentence there is a valency complementation realized as an infinitive form of the verb *uklízet* [to clean] and its two valency complementations, [ona] [she] (non-expressed) and *zahradu* [garden],[10] see Fig. 7.

In order to obtain the core predicative structure, the following simplification of the reduction component is used: (i) the complementations [ona] [she] and *zahradu* [garden] of the head verb *uklízet* [to clean] are deleted and (ii) the word form *uklízet* [to clean] and all the categories relevant to this word form apart from its functor (here PAT, patient) are deleted – such a simplified item represents a (saturated) lexical item with zero morphemic form (and thus, the valency requirements remain satisfied.

This step results in the core predicative structure in Fig. 8.

Example: The following construction (called genitive of property, see Šmilauer, 1966, p. 175) is another example of reduction component.

[10] We leave aside the relation of control, i.e. a specific type of grammatical coreference between a complementation of a governing node, called controller – here *Marie* as ADDR (addressee) of the verb *pomáhat* [to help] – and (non-expressed) subject of the infinitive verb, called controllee – here *uklízet* [to clean].

→

Petr	*m-Petr*.NNMS1	Sb	*t-Petr*.ACT
pomáhal	*m-pomáhat*.VpYS-	Pred	*t-pomáhat*.PRED.Frame1.ind-ant
Marii	*m-Marie*.NNFS3	Obj	*t-Marie*.ADDR
			[ ].PAT
.	..Z: - - -	AuxK	

Figure 8. The core predicative structure for sentence (4).

			[on].ACT
Uviděl	*uvidět*.VpYS-	Pred	*t-uvidět*.PRED.Frame4.ind-ant
dívku	*m-dívka*.NNFS4	Obj	*t-dívka*.PAT
vysoké	*m-vysoký*.AAFS2	Atr	
postavy	*m-postava*.NNFS2	Atr	*t-postava*.APP
			t-vysoký.RSTR
.	..Z: - - -	AuxK	

Figure 9. The input structure for sentence (5).

(5) *Uviděl dívku vysoké postavy.*
[saw girl (of) tall figure]

E. He saw a girl with a tall figure.

The adnominal attribute (realized (usually) as a noun in genitive case), here *postavy* [figure], obligatorily requires some modification, here *vysoké* [tall], see Fig. 9.

This means that the whole component *vysoké postavy* [(with a) tall figure] must be processed within one cycle. As the head of the component *postavy* [figure] is not required by the valency of the verb, both parts of the reduction component are simply deleted in one cycle. Thus, the core predicative structure is obtained, see Fig. 10.

→

			[on].ACT
Uviděl	*uvidět*.VpYS-	Pred	*t-uvidět*.PRED.Frame4.ind-ant
dívku	*m-dívka*.NNFS4	Obj	*t-dívka*.PAT
.	..Z: - - -	AuxK	

Figure 10. The core predicative structure for sentence (5).

2.3. Word order

A large effort has been devoted to clearing up the role of word order in so called free-word order languages, see e.g. Hajičová, Partee, and Sgall (1998); Holan et al. (2000); Havelka (2005); Hajičová (2006) for some of the most recent contributions for Czech.

Let us recall two basic principles for the tectogrammatical representation of FGD, see esp. Sgall, Hajičová, and Panevová (1986); Hajičová, Partee, and Sgall (1998):

- The word order in TR (deep word order) reflects the topic-focus articulation – it corresponds to the scale of communicative dynamism (thus it may differ from the surface word order).
- The theoretical research assumes the validity of the principle of projectivity for TRs.[11]

These two principles have important consequences for the analysis by reduction that models the transition from surface form of a sentence to its TR – the surface word order must be modified in order to obtain the deep word order (sentence (6)). This holds particulary for sentences with non-projective surface structure (sentence (7)). It implies that the sentence representation must in general reflect two word orders, the surface and the deep one. Let us repeat here the adopted convention of displaying examples, particularly that for word order – whereas columns 1, 2, 3 depict surface word order, column 4, reflecting tectogrammatical representation, reveals the deep word order.

Example: Let us concentrate here on the topic focus articulation, see esp. Hajičová, Partee, and Sgall (1998) and the writings quoted there.

(6) *Černý kocour se napil ze své misky.* (see Mikulová et al., 2006, Section 10.3.1.)
[black tomcat *refl* drunk from its bowl]
E. The black tomcat drank from its bowl.

According to Mikulová et al. (2006), the most general guideline of representing deep word order in TR is the placing of nodes representing contextually bound expressions to the left from their governing node and the placing of nodes representing contextually non-bound expressions to the right from their governing node. The contextual boundness is described in the attribute 'tfa', the values 'c' (contrastive topic), 't' (contextually bound) and 'f' (contextually non-bound) belong to the metalanguage categories in the tectogrammatical representations. The input structure for analysis is in Fig. 11, the last category in the fourth column, divided by '_', reflects tfa.

The actor, ACT *kocour*_t [tomcat] is contextually bound and it appears to the left of its governing verb *napil_se*_f [drank] in the surface; the contextually non-bound DIR1 complementation *misky*_f [bowl] is to the right of its governing verb; and the contextually bound

[11] A great number of definitions of projectivity appears in literature since the 1960s, more or less formal. In Mikulová et al. (2006) the projectivity is defined as follows: 'if two nodes M and N are connected by an edge and M is to the left from N, then all nodes to the right from M and to the left from N are connected with the root via a path that passes through at least one of the nodes M and N. In short: between a mother and its direct daughter there can be only direct or indirect daughters of the mother.'

Černý	*m-černý*.NNMS1	Atr	
kocour	*m-kocour*.NNMS1	Sb	*t-kocour*.ACT_t *t-černý*.RSTR_f [Gen].PAT_t
se	*m-se*.P7-X4	AuxR	
napil	*m-napít*.VpYS-	Pred	*t-napít_se*.PRED.Frame5_f
ze	*m-z*.RV- - 2	AuxP	
své	*m-svůj*.P8FS2	Atr	[PersPron].APP_t
misky	*m-miska*.NNFS2	Adv	*t-miska*.DIR1.basic_f
.	..Z: - - -	AuxK	

Figure 11. The input structure for sentence (6).

*svůj*_t [his] is to the left from its governing word *miska*_f [bowl] as well – the surface word order agrees in these cases with the deep word order.[12]

On the other hand, the modification *černý*_f [black] is contextually non-bound and it stands before its (bound) governing word *kocour*_t [tomcat] – here the surface word order disagrees with the deep word order. This is the reason why the ordering in the last column (with the tectogrammatical representation) does not replicate the ordering of other columns – the contextually bound modification *černý*_f [black] appears at the second position in the TR of the sentence (just behind the governing item *kocour*_t [tomcat]).

Now, the reduction phase can start, i.e. a stepwise simplification of the sentence according to the principles of analysis by reduction, during which the dependencies are treated and the core predicative structure is obtained, as is described in the previous subsection.

Example: Sentence (7) has non-projective surface realization.

(7) *Karla plánujeme poslat na rok do Anglie.*
[Charles plan to_send for year to England]
(see Sgall, Hajičová, and Panevová, 1986, p. 241)
E. Charles we are planning to send for a year to England. ≈ As for Charles, we are planning to send him for a year to England.

The proper noun *Karla*_c [Charles], which is the contrastive topic of the sentence (tfa = 'c'), is moved away from its governing verb *poslat*_f [to send], which causes a non-projectivity in the surface structure. The theoretical assumption of projectivity of TRs requires a different deep order – the corresponding item *t-Charles*.PAT_c in TR is situated just before its governing item *t-poslat*.PRED.Frame1_f [to send]. The analysis by reduction has the input structure given in Fig. 12.

Now, the reduction phase treating the dependencies can start.

[12] We suppose that also restored ellipses (here [Gen].t_PAT, generalized adverbal patient, PAT) are placed in the respective position in the input string.

Karla	*m-Karel*.NNMS4	Obj	
			[my].ACT_t
plánujeme	*m-plánovat*.VB-P-	Pred	*t-plánovat*.PRED.Frame6.ind-sim_f
			t-Karel.PAT_c
			[my].ACT_t
poslat	*m-poslat*.Vf- - -	Obj	*t-poslat*.PAT.Frame7_f
na	*m-na*.RR- - 4	AuxP	
rok	*m-rok*.NNIS4	Adv	*t-rok*.THL_f
do	*m-do*.RR- - 2	AuxP	
Anglie	*m-Anglie*.NNFS2	Adv	*t-Anglie*.DIR3.basic_f
.	..Z: - - -	AuxK	

Figure 12. The input structure for sentence (7).

3. The 4-LRL-automata

In this section, the formal model for analysis by reduction for FGD is proposed. We use here the standard way of presentation from the theory of automata (our remarks should hopefully help readers not quite familiar with that kind of presentation). This section is partitioned into two subsections. The first one introduces sRL-automata – the basic models of restarting automata we will be dealing with. The important notion of metarules is introduced here; they serve for a more transparent, more declarative description of restarting automata.

The second subsection introduces 4-LRL-automata as a special case of sRL-automata. A four-level *analysis by reduction system*, which is an algebraic representation of analysis by reduction, and the formal languages which represent the individual layers of FGD are introduced here, namely the languages of the first and the last level that correspond to the surface language LC and to the tectogrammatical language LM from Section 1. Further, the *characteristic relation* SH(M) is introduced.

Finally, the SH-*synthesis*, which models FGD as a generative device and specifies the generative ability of FGD, and SH-*analysis*, which fulfills the task of syntactico-semantic analysis of FGD, are introduced here step by step.

3.1. The t-sRL-automaton

Here we describe in short the type of restarting automaton we will be dealing with. The subsection is an adapted version of the first part of Messerschmidt et al. (2006). More (formal) details of the development of restarting automata can be found in Otto (2006).

An sRL-*automaton* (*simple* RL-*automaton*) M is (in general) a nondeterministic machine with a finite-state control Q, a finite characteristic vocabulary Σ, and a head with the ability to scan exactly one symbol (word) that works on a flexible tape delimited by the left sentinel ¢ and the right sentinel \$.

Let us proceed a bit more formally. A simple RL-automaton is a tuple $M = (Q, \Sigma, \delta, q_0, ¢, \$)$,

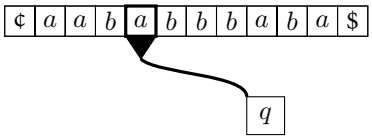

Figure 13. Restarting automaton

where:

- Q is a finite set of states
- Σ is a finite vocabulary (the characteristic vocabulary)
- ¢, \$ are sentinels, $\{¢, \$\}$ do not belong to Σ
- q_0 from Q is the initial state
- δ is the transition relation ≈ a finite set of instructions of the shape : $(q, a) \rightarrow {}_M(p, Op)$, where q, p are states from Q, a is a symbol from Σ, and Op is an operation, where the particular operations correspond to the particular types of steps (move-right, move-left, delete, accept, reject, and restart step).

- move right
- move left
- delete
- restart

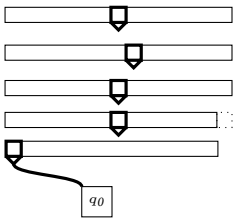

Figure 14. Operations

For an input sentence $w \in \Sigma^*$, the initial tape inscription is ¢w\$. To process this input, M starts in its initial state q_0 with its window over the left end of the tape, scanning the left sentinel ¢.

According to its transition relation, M performs *move-right steps* and *move-left steps*, which change the state of M and shift the window one position to the right or to the left, respectively, and *delete steps*, which delete the content of the window, thus shorten the tape, change the state, and shift the window to the right neighbor of the symbol deleted. Of course, neither the left sentinel ¢ nor the right sentinel \$ may be deleted. At the right end of the tape, M either halts and *accepts*, or it halts and *rejects*, or it *restarts*, that is, it places its window over the left end of the tape and reenters the initial state. It is required that before the first restart step and also between any two restart steps, M executes at least one delete operation.

A *configuration* of M is a string $\alpha q \beta$ where $q \in Q$, and either $\alpha = \lambda$ and $\beta \in \{¢\} \cdot \Sigma^* \cdot \{\$\}$ or

- restarting configuration
- accepting configuration

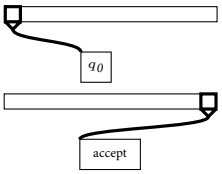

Figure 15. Basic configurations.

$\alpha \in \{¢\} \cdot \Sigma^*$ and $\beta \in \Sigma^* \cdot \{\$\}$; here q represents the current state, $\alpha\beta$ is the current content of the tape, and it is understood that the window contains the first symbol of β. A configuration of the form $q_0¢w\$$ is called a *restarting configuration*.

We observe that each computation of an sRL-automaton M consists of certain phases. Each part of a computation of M from a restarting configuration to the next restarting configuration is called a *cycle*. The part after the last restart operation is called the *tail*. We use the notation $u \vdash_M^c v$ to denote a cycle of M that begins with the restarting configuration $q_0¢u\$$ and ends with the restarting configuration $q_0¢v\$$; the relation $\vdash_M^{c^*}$ is the reflexive and transitive closure of $\vdash_M^c$.

An input $w \in \Sigma^*$ is *accepted* by M, if there is an accepting computation which starts with the (initial) configuration $q_0¢w\$$. By $L_C(M)$ we denote the *characteristic language* consisting of all strings accepted by M; we say that M *recognizes (accepts) the language* $L_C(M)$. By $S_C(M)$ we denote the *simple language* accepted by M, which consists of all strings that M accepts by computations without a restart step. Obviously, $S_C(M)$ is a regular sublanguage of $L_C(M)$. By sRL we denote the class of all sRL-automata.

A t-sRL-automaton ($t \geq 1$) is an sRL-automaton which uses at most t delete operations in a cycle and any string of $S_C(M)$ has no more than t symbols (tokens).

Remark: The t-sRL-automata are two-way automata which allow, in any cycle, to check the whole sentence before reduction (deleting). This reminds us of the behavior of a linguist who can read the whole sentence before choosing the reduction. The automaton should be non-deterministic in general in order to be able to change the order of deleting cycles. That serves for witnessing the independence of some parts of the sentence, see the section about the analysis by reduction. Another message from this section is that there is a t which creates a boundary for the number of deletions in a cycle and for the size of the accepted irreducible strings.

Based on Messerschmidt et al. (2006), we can describe a t-sRL-automaton by *metainstructions* of the form

$$(¢ \cdot E_0, a_1, E_1, a_2, E_2, \ldots, E_{s-1}, a_s, E_s \cdot \$)\ ,\ 1 \leq s \leq t\ ,\ \text{where}$$

- $E_0, E_1, \ldots, E_s$ are regular languages (often represented by regular expressions), called the *regular constraints* of this instruction, and
- $a_1, a_2, \ldots, a_s \in \Sigma$ correspond to letters that are deleted by M during one cycle.

In order to execute this metainstruction, M starts from a configuration $q_0 ¢ w \$$; it will get stuck (and so reject), if w does not admit a factorization of the form $w = v_0 a_1 v_1 a_2 \cdots v_{s-1} a_s v_s$ such that $v_i \in E_i$ for all $i = 0, \ldots, s$. On the other hand, if w admits factorizations of this form, then one of them is chosen nondeterministically, and the restarting configuration $q_0 ¢ w \$$ is transformed into $q_0 ¢ v_0 v_1 \cdots v_{s-1} v_s \$$. To describe also the tails of the accepting computations, we use accepting metainstructions of the form $(¢ \cdot E \cdot \$, \mathsf{Accept})$, where E is a regular language (finite in this case). Moreover, we can require that there is only a single accepting metainstruction for M.

Example: Let $t \geq 1$, and let $L_{R_t} = \{ c_0 w c_1 w c_2 \cdots c_{t-1} w \mid w \in \{a, b\}^* \}$. For this language, a t-sRL-automaton M_t with a vocabulary $\Sigma_t = \{c_0, c_1, \ldots, c_{t-1}\} \cup \Sigma_0$, where $\Sigma_0 = \{a, b\}$, can be obtained through the following sequence of metainstructions:

$$
\begin{array}{ll}
(1) & (¢ c_0, a, {\Sigma_0}^* \cdot c_1, a, {\Sigma_0}^* \cdot c_2, \ldots, {\Sigma_0}^* \cdot c_{t-1}, a, {\Sigma_0}^* \cdot \$), \\
(2) & (¢ c_0, b, {\Sigma_0}^* \cdot c_1, b, {\Sigma_0}^* \cdot c_2, \ldots, {\Sigma_0}^* \cdot c_{t-1}, b, {\Sigma_0}^* \cdot \$), \\
(3) & (¢ c_0 c_1 \cdots c_{t-1} \$, \mathsf{Accept}).
\end{array}
$$

It follows easily that $L(M_t) = L_{R_t}$ holds.

We emphasize the following properties of restarting automata.

Definition: (Error Preserving Property) A t-sRL-automaton M is *error preserving* if $u \notin L_C(M)$ and $u \vdash_M^{c^*} v$ imply that $v \notin L_C(M)$.

The following property plays an important role in our applications of restarting automata.

Definition: (Correctness Preserving Property) A t-sRL-automaton M is *correctness preserving* if $u \in L_C(M)$ and $u \vdash_M^{c^*} v$ imply that $v \in L_C(M)$.

It is rather obvious that each t-sRL-automaton is error preserving, and that all deterministic t-sRL-automata are correctness preserving. On the other hand, one can easily construct examples of nondeterministic t-sRL-automata that are not correctness preserving.

3.2. The 4-LRL-automata and related notions

Let us finally introduce the model of automaton proposed for modelling of analysis by reduction for FGD. A 4-LRL-*automaton* (*4-level* sRL-*automaton*) M_{FGD} is a (correctness preserving) t-sRL-automaton, where its characteristic vocabulary Σ is composed from four subvocabularies $\Sigma_0, \ldots, \Sigma_3$. M_{FGD} deletes at least one symbol from Σ_0 in each cycle.

Remark: The correctness and error preserving properties of M_{FGD} should ensure a good simulation of the linguist performing the analysis by reduction. Similarly as the linguist, the automaton M_{FGD} should not make a mistake during analysis by reduction, otherwise there is something wrong, e.g. the characteristic language is badly proposed. This situation can be improved by adding some new categories (symbols). The correctness preserving property can be automatically tested. This may be useful for checking and improving a language description in

the context of FGD. The request of the deletion of at least one surface wordform in any cycle represents the request of the (generalized) lexicalization of FGD.

Let us inherit the notions $L_C(M_{FGD})$, *characteristic language* of M_{FGD} and $S_C(M_{FGD})$, *simple language* from the previous subsection. All the notions introduced below are derived from these notions.

As the first step, we introduce an *(analysis by) reduction system* involved by M_{FGD}, and by the set of level alphabets $\Sigma_0, \ldots, \Sigma_3$. It is defined as follows:

$$\mathsf{RS}(M_{FGD}) := (\Sigma^*, \vdash^c_{M_{FGD}}, S_C(M_{FGD}), \Sigma_0, \cdots, \Sigma_3).$$

The reduction system (by M_{FGD}) formalizes the notion of the analysis by reduction of FGD in an algebraic, non-procedural way. Observe that for each $w \in \Sigma^*$, we have $w \in L_C(M_{FGD})$ if and only if $w \vdash^{c^*}_{M_{FGD}} v$ holds for some string $v \in S_C(M_{FGD})$.

A *language of level j recognized by* M_{FGD}, where $0 \leq j \leq 3$, is the set of all sentences (strings) that are obtained from $L_C(M_{FGD})$ by removing all symbols which do not belong to Σ_j. We denote it $L_j(M_{FGD})$. Particularly, $L_0(M_{FGD})$ represents the surface language LC defined by M_{FGD}; similarly, $L_3(M_{FGD})$ represents the language of tectogrammatical representations LM defined by M_{FGD} (see Section 1).

Now we can define the *characteristic relation* $\mathsf{SH}(M_{FGD})$ given by M_{FGD}.

$\mathsf{SH}(M_{FGD}) = \{(u, y) \mid$ there is a $w \in L_C(M_{FGD})$ such that $u \in L_0(M_{FGD})$ and u is obtained from w by deleting the symbols not belonging to Σ_0, and $y \in L_3(M_{FGD})$ and y is obtained from w by deleting the symbols not belonging to $\Sigma_3\}$.

Remark: The characteristic relation represents the basic relations in language description, relations of synonymy and ambiguity in language L. In other words, it embraces the translation of the surface language LC into the tectogrammatical language and vice versa. From this notion, the remaining notions, analysis and synthesis, can be derived.

We introduce the SH-*synthesis by M_{FGD} for any y from* LM as a set of pairs (u, y) belonging to $\mathsf{SH}(M_{FGD})$.

$$synthesis\text{-}\mathsf{SH}(M_{FGD}, y) = \{(u, y) | (u, y) \in \mathsf{SH}(M_{FGD})\}$$

The SH-synthesis associates a tectogrammatical representation (i.e. string y from LM) with all its possible surface sentences u belonging to LC. This notion allows for checking the synonymy and its degree provided by M_{FDG}. The linguistic issue is to decrease the degree of the synonymy by M_{FDG} by the gradual refinement of M_{FDG}.

Finally we introduce the dual notion to the SH-synthesis, the SH-*analysis by M_{FGD} of u*:

$$analysis\text{-}\mathsf{SH}(M_{FGD}, u) = \{(u, y) | (u, y) \in \mathsf{SH}(M)_{FGD}\}$$

The SH-analysis returns, to a given surface sentence u, all its possible tectogrammatical representations, i.e. it allows for checking the ambiguity of an individual surface sentence. This notion provides the formal definition for the task of full syntactico-semantic analysis by

[*Včera*]$_1$	[*m-včera*.Dg- - -]$_2$	[Adv]$_3$	[*t-včera*.TWHEN]$_4$
			[[on].ACT]$_5$
[*přišel*]$_6$	[*m-přijít*.VpYS-]$_7$	[Pred]$_8$	[*t-přijít*.PRED.Frame1.ind-ant]$_9$
[*domů*]$_{10}$	[*m-domů*.Db- - -]$_{11}$	[Adv]$_{12}$	[*t-domů*.DIR3]$_{13}$
[*pozdě*]$_{14}$	[*m-pozdě*.Dg- - -]$_{15}$	[Adv]$_{16}$	[*t-pozdě*.TWHEN]$_{17}$
[.]$_{18}$	[..Z: - - -]$_{19}$	[AuxK]$_{20}$	

Figure 16. The input string for sentence (2).

M_{FDG}. The linguistic task is to refine M_{FDG} gradually, especially with respect to the description of ambiguity of the sentence.

Remark: Fig. 16 illustrates the transformation of the input structures used in Section 2 into the input strings for a M_{FDG} automaton. The individual numbered items in square brackets in Fig. 16 represents the individual symbols on the input tape of M_{FDG}. E.g., [*Včera*]$_1$ is the first symbol on the tape (after the left sentinel) belonging to Σ_0; [*m-včera*.Dg- - -]$_2$ is the second symbol (it is from Σ_1) and so on. [AuxK]$_{20}$ is the last symbol (item) on the tape before the right sentinel.

4. Concluding remarks

The paper presents the basic formal notions that allow for formalizing the notion of analysis by reduction for Functional Generative Description, FGD. We have outlined and exemplified the method of analysis by reduction and its application in processing dependencies and word order in a language with a high degree of free word order. Based on this experience, we have introduced the 4-level reduction system for FGD based on the notion of simple restarting automata. This new formal frame allows us to define formally the characteristic relation for FGD, which renders synonymy and ambiguity in the studied language.

Such a formalization makes it possible to propose a software environment for the further development. It provides a possibility to describe exactly the basic phenomena observed during linguistic research. Further, it allows for studying suitable algorithms for tasks in computational linguistics, namely automatic syntactico-semantic analysis and synthesis.

The presented notions are also useful to show exactly the differences and similarities between the methodological basis of our (computational) linguistic school and the methodological bases of other schools. The basic message given here is to show the possibility of generalizing the principle of lexicalization trough the layers in order to obtain a checking procedure for FGD via analysis by reduction.

Acknowledgement

This paper is a result of the project supported by the grants No. 1ET100300517 and MSM0021620838. A preliminary version of this contribution was presented at the conference

Formal Description of Slavic Languages 6.5, December 2006 (to appear in the Proceedings of FDSL 6.5).

Bibliography

Hajič, Jan. 2005. Complex Corpus Annotation: The Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*. Veda Bratislava, Slovakia, pages 54–73.

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

Hajičová, Eva. 2006. K některým otázkám závislostní gramatiky. *Slovo a slovesnost*, (67):3–26.

Havelka, Jiří. 2005. Projectivity in Totally Ordered Rooted Trees: An Alternative Definition of Projectivity and Optimal Algorithms for Detecting Non-Projective Edges and Projectivizing Totally Ordered Rooted Trees. *The Prague Bulletin of Mathematical Linguistics*, (84):13–30.

Holan, Tomáš, Vladislav Kuboň, Karel Oliva, and Martin Plátek. 2000. On Complexity of Word Order. *Les grammaires de dépendance – Traitement automatique des langues*, 41(1):273–300.

Lopatková, Markéta, Martin Plátek, and Vladislav Kuboň. 2005. Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In *Lecture Notes in Artificial Intelligence, Proceedings of Text, Speech and Dialogue*, volume 3658, pages 140–147, Heidelberg. Springer Verlag.

Messerschmidt, Hartmut, František Mráz, Friedrich Otto, and Martin Plátek. 2006. Correctness Preservation and Complexity of Simple RL-Automata. In *Lecture Notes in Computer Science*, volume 4094, pages 162–172, Heidelberg. Springer Verlag.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report 30, ÚFAL MFF UK, Prague. trans. Součková, K., Böhmová, A., Čermáková, K., Havelka, J., Corness, P.

Otto, Friedrich. 2006. Restarting Automata. In Z. Ésik, C. Martin-Vide, and V. Mitrana, editors, *Recent Advances in Formal Languages and Applications, Studies in Computational Intelligence*, volume 25, pages 269–303, Berlin. Springer.

Panevová, Jarmila. 1974. On Verbal Frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, (22):3–40.

Petkevič, Vladimír. 1995. A New Formal Specification of Underlying Structure. *Theoretical Linguistics*, 21(1).

Plátek, Martin. 1982. Composition of Translation with D-trees. In *Proceedings of COLING'82*, pages 313–318.

Plátek, Martin and Petr Sgall. 1978. A Scale of Context-Sensitive Languages: Applications to Natural Language. *Information and Control*, 38(1):1–20.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel Publishing Company, Dordrecht. ed. Mey, J.

Sgall, Petr, Ladislav Nebeský, Alla Goralčíková, and Eva Hajičová. 1969. *A Functional Approach to Syntax in Generative Description of Language*. American Elsevier Publishing Company, Inc., New York.

Šmilauer, Vladimír. 1966. *Novočeská skladba*. SPN, Praha. 2^{nd} edition.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 27–40

On Reciprocity

Jarmila Panevová, Marie Mikulová

1. Introduction

The description of the reciprocity phenomenon is more tricky than it is supposed in grammatical handbooks: It must cover both the issues of lexicon and of syntax and of their interplay as well.

The lexical counterpart of the English expression *each other* is not the central (core) means for denoting reciprocity in some Slavonic languages, esp. in Czech. The troublemaking Czech reflexive *se*/*si* plays a substantial part of responsibility for reciprocal relations. With some lexical items there is no surface expression of reciprocity, as we will demonstrate later.[1]

The distinction between reciprocity as a part of lexical meaning of particular lexical items (see Section 2) and reciprocity as a syntactic relation between some participants of the syntactic construction seems to be universal. However, a lexical item that is characterized by the feature of inherent lexical reciprocity could be used in asymmetric (syntactically non-reciprocal) constructions (see the asymmetry between *John* and *Mary* in (1) and the symmetry of their respective roles in a syntactically reciprocal construction (2)):

1. *Jan se setkal s Marií v divadle.* [John met Mary in the theatre.]
   [John-Nom se-Refl meet-3sg Prep-with Mary-Instr…]
2. *Jan a Marie se setkali v divadle.* [John and Mary met each other in the theatre.]
   [John-Nom and Mary-Nom se-Refl meet-3pl…]

From the other side, many items lacking the lexical feature of reciprocity can be used in syntactically reciprocal constructions (see (3)):

3. *Jan a Marie se fotografují (navzájem).*[2] [John and Mary photograph each other.] [John

[1] See also the comparison of German *einander* and Czech *jeden – druhý* given by Štícha (2003). The former is evaluated by him as more usual and more neutral than the latter.

[2] The other readings of (3) (without an adverb *navzájem*) are left aside here. They are connected with the ambiguity of *se*-constructions in Czech (see e.g. Panevová, 2001).

Please cite this article as: Jarmila Panevová, Marie Mikulová, On Reciprocity. The Prague Bulletin of Mathematical Linguistics No. 87, 2007, 27–40.

and Mary se-Refl photograph-3pl (each other)]

A sample of the classification of Czech verbs as to the lexical feature of reciprocity is given in Section 2. The syntactic reciprocity (using an operation of reciprocalization with its respective consequences for valency) as well as the problems connected with this approach will be described in Section 3. In Section 4 some results of searching for reciprocity phenomena in the electronic corpora of Czech will be presented.

2. Reciprocal and non-reciprocal verbs in Czech

Czech verbs can be classified from this point of view into three classes. In the meaning of the Czech verbs from the classes A and B (see below) the feature of reciprocity is implied, though they can be used in unreciprocal constructions (see (4) and (5) below) as well. The verbs from the classes A and B differ from each other: the reciprocity of verbs from the class A is inherent; if they belong to the class of "reflexiva tantum", they have no unreciprocal counterpart, while the verbs from the class B are "derived" reciprocals, they have unreciprocal counterparts. The verbs from the class C are not lexically reciprocal.

A. **Inherent reciprocal verbs**:
   Reflexive verbs: *hádat se* [to quarrel], *prát se* [to fight], *utkat se* [to clash with], *přít se* [to quarrel], *setkat se* [to meet], *scházet se* [to meet], *loučit se* [to say good-bye], *domlouvat se* [to agree], *podobat se* [to resemble].
   Irreflexive verbs: *zápasit* [to struggle], *soutěžit* [to compete], *diskutovat* [to discuss], *polemizovat* [to argue], *obchodovat* [to trade], *sousedit* [to neighbor], *splývat* [to blend].
B. **Derived reciprocal verbs**:
   líbat se [to kiss], *objímat se* [to embrace], *potkat se* [to meet], *pozdravit se* [to greet], *seznámit se* [to make an acquaintance], *vítat se* [to greet], *navštěvovat se* [to visit], *spojovat se* [to connect], *lišit se* [to differ].
C. **Lexically non-reciprocal verbs**:
   líbat [to kiss], *objímat* [to embrace], *fotografovat* [to photograph], *napodobovat* [to imitate], *popisovat* [to describe], *obviňovat* [to blame], *oceňovat* [to appreciate], *osočovat* [to smear], *pomlouvat* [to gossip], *udávat* [to denunciate], *vidět* [to see].

In the asymmetric (non-reciprocal) usage of the verbs from A and B the implication that at least two participants involved are included in the same action is highly probable, but it is not certain (see (4) and (5), where the lexical reciprocity is canceled):

4. *Starší syn se **rád** hádá s mladším.* [The older son **gladly** argues with the younger one.]
5. *Jan se **chce** s Marií líbat pokaždé, když ji vidí, ale ona se vzpírá.* [John **wants** to kiss Mary every time when he meets her but she refuses.]

Many verbs belong to the class C, consisting of the lexically non-reciprocal items, which could be syntactically reciprocalized (see Section 3). This class is wide and it seems to be open. It should be noticed that many of them have a reflexive derivation belonging to the class B (e.g. *líbat se* [to kiss], *objímat se* [to embrace]). This step, called by us a derivation, is understood by Chrakovskij (1999) as a difference between dynamic (*líbat* [to kiss]) and static verbs (*líbat se* [to kiss each other]).

3. Syntactic reciprocity

We have described the syntactic operation of reciprocalization earlier Panevová (1999); Panevová (in press) in a more detailed way. Here, we only shortly repeat that the reciprocalization is a syntactic operation on the valency frames of verbs (and other lexical items) in which two valency slots are the bearers of the feature allowing their symmetrical usage, as is illustrated by sentence (11). This feature is technically marked in the valency frame by the superscript R (see (6), (7)).

6. *prát se* [to fight] - ACT^R (Nom), PAT^R (s + Instr) (class A)
7. *vzpomínat* [to remember] - ACT^R (Nom), PAT^R (na + Accus) (class B)

Verbs belonging to the classes A, B and C can be used in syntactically reciprocal construction in which one valency slot of the verb is deleted. The deletion is reflected in the syntactic structure either as a plural noun in subject[3] (single elements of the plural noun participate in the same way in the action, see (8)) or as a coordinated construction of subjects (where the elements participating on the action is separated, see (9))[4]. In Chrakovskij (1999), the term "soprjazhenije rolej" (combining of the roles) is used for the noun in plural or for the members of coordination.

8. *Sourozenci se perou.* [The siblings se-Refl fight.]
9. *Jan a Robert se perou.* [John and Bob se-Refl fight.]

Other conditions for the using of this operation are also described elsewhere Panevová (1999): the homogeneity of the combined participants as to their lexical meaning and as to their position in the topic/focus articulation are required (see (10a), and unacceptability of (10b)) as well as the validity of paraphrases (11a) and/or (11b) for (11) are necessary:

10. (a) *Jan se setkal s námitkami.* [John se-Refl met the objections.]
    (b) **Jan a námitky se setkali.* [John and objections met.]
11. *Jan a Marie se líbají.* [John and Mary se-Refl kiss-3pl] ← (11a), (11b)
    (a) *Jan líbá Marii a (zároveň) Marie líbá Jana.* [John kisses Mary and (simultaneously) Mary kisses John.]
    (b) *Jan se líbá s Marií a (zároveň) se Marie líbá s Janem.* [John se-Refl kisses Mary and (simultaneously) se-Refl Mary kisses John.]

We encounter here a theoretical problem: (11) is described as ambiguous because it has two sources (11a) and (11b). If we take into account the other means for expressing reciprocity in Czech (the expression *jeden – druhý* [each – other]), we actually receive two different paraphrases: (12a) and (12b) for (11a) and (11b), respectively:

12. (a) *Jan a Marie líbají jeden druhého.* [John and Mary kiss-3pl each-Nom other-Accus]
    (b) *Jan a Marie se líbají jeden s druhým.* [John and Mary se-Refl kiss-3pl each-Nom s-Prep other-Instr sg]

[3] Examples in which the reciprocalization does not include the subject position are discussed in Panevová (1999).

[4] The collective (uncountable as well as countable) nouns (e.g. *šlechta* [aristocracy], *dělnictvo* [labour], *mužstvo* [team], *rodina* [family], *vláda* [government]) have to be understood as a semantical plural.

In (12a) the lexical element *each – other* is obligatory (its absence would cause ungrammaticality of this construction), while in (12b) it is optional. In (12a), it stands instead of a reflexive pronoun as a true reflexive in Accusative. In (12b), the elements *each - other* are used with the derived reciprocal verb *líbat se* [to kiss se-Refl] as optional and the syntactic construction (with the Objective moved in the coordinated Actor/Subject) is both grammatical and reciprocal.

The conclusion of this observation may be formulated as follows: There are two different lexical items in Czech lexicon: *líbat* [to kiss] and *líbat se* [to kiss se-Refl] belonging to the classes C and B, respectively; in their valency frames their both Actors and Objectives (Patients) bear a superscript R. The use of the superscript gives at least one common ambiguous output and some paraphrases different for (11a) and (11b).

4. Formal expressions of reciprocity in Czech

Analyzing the formal expressions of the syntactic reciprocity in which the first participant (Actor) and some other participant are involved,[5] we have received a scale of means which are partially grammatical, partially lexical, some of them standing on the boundary between lexicon and grammar.

4.1. With the inherent reciprocal verbs (class A) and derived inherent reciprocal verbs (class B) the change of syntactic structure (i.e. a multiplied subject and a missing valency member) is a sufficient marker of reciprocity in principle and no overt expression for it is needed. However, the material from corpora[6] shows, that the situation is more complicated and differs from one verb to another. With some verbs such zero expression is either ambiguous (see (13), (14)), or strange (up to unacceptability), see (15), (16):

13. *Američtí poradci jednají o Ulsteru.* (PDT) [American advisors negotiate Ulster.]
14. *Všechna mužstva bojují o místo, které zajišťuje start v evropských pohárech.* (SYN2005). [All teams fight for the position guaranteed the start in the European cups.]
15. *?Matka a dcera se podobají.* [Mother and daughter resemble each other.]
16. *?Země EU obchodují.* [The countries of EU trade.]

However, many sentences with zero expression of reciprocity with the verbs from A and B classes sound well enough and their reciprocal interpretation is obvious, see e.g. (17), (18):

17. *Pověření poslanci budou o základních principech důchodového pojištění zřejmě ještě dlouho diskutovat a mohou padnout závažná rozhodnutí.* (PDT) [Charged deputies will discuss basic principles of tax insurance and important decisions may be achieved.]
18. *V těchto místech komické i chmurné stránky počítačové historie splývají.* (PDT) [In these places funny and sad points of the computational history blend.]

[5] Other issues, such as multiplied reciprocity with which several pairs of participants enter the syntactically reciprocal relation (such as *Pavel a Jan spolu mluvili o sobě navzájem.* [Paul and John talked with each other about each other.]), the reciprocity between a participant and a free adverbial as well as the reciprocity between noun complementations are studied elsewhere Panevová (1999); Panevová (in press); Mikulová et al. (2005).

[6] We have used the Czech National Corpus (CNC) in its variant SYN2005 (morphologically tagged corpus) and the syntactically annotated corpus the Prague Dependency Treebank (PDT) in its version 2.0.

The interpretation of the empty valency position probably depends on the semantics of the given verb and on the wider context of the sentence. The verbs *soutěžit* [to compete], *soupeřit* [to compete], for instance, presuppose the existence of the other competitor by their lexical meaning. Therefore, such sentences as (19), (20) are undoubtedly reciprocal:

19. *Firmy by měly soutěžit kvalitou poskytovaného servisu nebo cenami.* (SYN2005) [The companies would compete as to the quality of the delivered services or as to the prices.]
20. *…týmy České republiky, Finska, Ruska a Švédska soupeří o neoficiální titul mistra Evropy.* (PDT) […the teams of Czech Republic, Finland, Russia and Sweden compete for unofficial title of European champions.]

On the contrary, verbs such as *bojovat* [to fight], *zápasit* [to struggle] may have beside the interpretation of "a fight of rivals" also an interpretation "to fight to reach something" (see also Lopatková et al., 2006). Therefore, the empty valency position also opens other interpretations than the reciprocal one (see (14) above and (21), (22); the reciprocal interpretation is, of course, excluded in some cases (see (23), (24)):

21. *Tehdy zde bojovali Mohykánovi bratři a příbuzní.* (SYN2005) [At that time Mohykan's brothers and relatives fought there.]
22. *…národy bojují o území a přírodní zdroje, jednotlivci bojují…* (SYN2005) […nations fight for territories and natural sources, individuals fight for…]
23. *Pohled na malá prasátka, jak zápasí, aby se postavila na vlastní nohy.* (SYN2005) [A view on little pigs how they struggle to stand on their own legs.]
24. *Ale Renovi jezdci zápasili o holý život.* (SYN2005) [However, Reno's riders struggled for their poor life.]

A similar behavior as of the verb *bojovat* [to fight] is proper to the verbs with an Addressee (expressed in Czech by the prepositional phrase s with Instrumental) such as *diskutovat* [to discuss], *polemizovat* [to argue], *mluvit* [to talk], *hovořit* [to talk], *souhlasit* [to agree], but also by such inherent reciprocals as e.g. *prát se* [to fight], *loučit se* [to part], see (25), (26):

25. *Můj mladší i starší syn se ve škole rádi perou.* [My younger son and older one love to fight at school.]
26. *Otec a matka se už loučí, odjíždějí na léto na chatu.* [Our father and mother say good bye, they are leaving for the country cottage for the summer.]

4.2. With the verbs analyzed here, certain optional lexical expressions can be used. In Czech the adverbs *spolu* [together], *navzájem/ vzájemně* [each other], the prepositional construction *mezi sebou* [among/between each other] and the expression with a special agreement *jeden – druhý* [each other] belong to these optional means. Due to the grammatical features of the latter item[7], we classify it as an alternative (semi)grammatical means for the reciprocity in Czech. The items enumerated here are interchangeable in majority of contexts, however, sometimes some of them sound peculiarly; see (27), (28), (29):

[7] In examples (12a) and (12b) above, the mixed character of the agreement of this complex item is illustrated: Its first part *jeden* agrees with the nominative of subject, its latter part *druhý* is required by the missing participant as to its case.

27. *Jak jsem později zjistil, soutěžili mezi sebou, kdo přijde s lepším příběhem.* (SYN2005) [As I have recognized later, they competed with each other, who would bring the better story.]
    (a) *…soutěžili spolu…*
    (b) *…soutěžili vzájemně…*
    (c) *…soutěžili navzájem…*
    (d) *…soutěžili jeden s druhým…*
28. *Po následující dva roky Mirea a Vlad spolu bojovali o valašský trůn.* (SYN2005) [During the next two years, Mirea and Vlad were fighting over the Moravian throne.]
    (a) *…mezi sebou bojovali …*
    (b) *…vzájemně bojovali …*
    (c) *…navzájem bojovali …*
    (d) *…jeden s druhým bojovali …*
29. *Účastnící kongresu se navzájem rozloučili a odjeli do svých domovů.* [The participants of the congress said their farewells to each other and left for their homes.]
    (a) *…se spolu rozloučili…*
    (b) *…se vzájemně rozloučili…*
    (c) *…se mezi sebou rozloučili…*
    (d) *…se jeden s druhým rozloučili…*

4.3. We have analyzed some samples of the occurrences of the verbs studied in this paper in CNC. The frequency of the selected verbs in the corpus SYN2005 is indicated in Table 1. Since these figures reflect all senses and all forms of selected verbs, they are of no great interest; they have partially influenced our selection of the samples studied in detail. Illustrative results of these studies will be described in Sections 4.3.1 - 4.3.6. We have excluded from the detailed analysis the verbs with many senses, such as *mluvit* [to talk] having 10 senses according to Lopatková et al. (2006) and the verbs with low frequency, such as *polemizovat* [to argue].

Table 1. Number of the occurrences of the selected verbs in the corpus SYN2005

mluvit [to talk]	46 213
souhlasit [to agree]	12 040
bojovat [to fight]	8 889
diskutovat [to discuss]	3 074
splývat [to blend]	1 153
soutěžit [to compete]	1 138
zápasit [to wrestle/struggle]	1 136
soupeřit [to compete]	654
polemizovat [to argue]	323

4.3.1. Among three senses of the verb *souhlasit* [to agree] (see Lopatková et al., 2006) we are interested in the sense 1 "somebody agrees with somebody". In the sample of the first 120 occurrences from SYN2005, there are only 2 occurrences of the sense 1 with a possible reciprocity reading, see (30):

30. *Nevěděli jsme, jak ho budeme chytat, ale všichni jsme souhlasili.* (SYN2005) [We did not know, how to catch him, however, we all agreed.]

4.3.2. In the sample of 400 occurrences of the verb *bojovat* [to fight], only 51 examples allow a reciprocal reading, in 3 among them, the adverb *spolu* [together] is present, in 4 the expression *mezi sebou* [among/between each other] is used. In 20 sentences, the general Actor appears and the interpretation "everybody involved fights with everybody"[8] is very probable (see (31)). In the rest of examples, reciprocity is highly probable too, see (32):

31. *Bojovalo se současně ve třech světadílech.* (SYN2005) [It was fought on three continents simultaneously.]
32. *Když se naskytla práce pro jednoho, bojovali o ni všichni nezaměstnaní.* (SYN2005) [When a job for one person appeared, all unemployed fought over it.]

4.3.3. The verb *zápasit* [to struggle/wrestle] in one of its senses, which are interesting from the point of view studied here, is close to the verb *bojovat* [to fight]. Among 150 occurrences of this verb in SYN2005, 50 examples are clearly reciprocal, in 18 sentences, the adverb *spolu* [together] is present (see (33)), in 2 of them, the expression *mezi sebou* [among/between each other] is used (see (34)):

33. *Rvali jsme se a zápasili spolu za měsíčního svitu.* (SYN2005) [We fought and struggle together in the moonlight.]
34. *Je říje, jeleni mezi sebou zápasí.* (SYN2005) [It is rutting season, the stags struggle with each other.]

4.3.4. Among 150 occurrences of the verb *diskutovat* [to discuss] from the SYN2005, the reciprocal relation between Actor(s) and Addressee(s) is present in 99 sentences; however, in 54 of them, it is the case of their generalization (see (35)); the lexical means are present rarely: 2x *spolu* [together] (see (36)) and 1x *mezi sebou* [among/between each other]. However, in some occurrences, esp. from scientific texts this verb looses its meaning "to have a discussion with an opponent" and it has the meaning of simple presentation (see (37)):

35. *Diskutovalo se stále o stejných problémech.* (SYN2005) [The same problems were discussed all the time.]
36. *Diskutovali spolu o schopnostech…* (SYN2005) [They discussed together the abilities of…]
37. *Některé normativní důsledky budeme diskutovat v jedné z následujících kapitol.* (SYN2005) [We shall discuss some normative consequences in one of the following chapters.]

[8] The considerations about the features of general actor, allowing its reciprocal usage, are included in Panevová (2006).

4.3.5. In the sample of 150 occurrences of the verb *soutěžit* [to compete], there are 43 examples enforcing the syntactic reciprocity (the rest of them display the asymmetrical usage, see (38)), among these 43 in 6 sentences the expression *mezi sebou* [among/between each other] is present, in 5 *spolu* [together] and in 1 *navzájem* [each other] occurs, see (39):

38. *To je vyšetřovací zařízení, s nímž irácká tajná služba mohla soutěžit leda ve snu.* (SYN2005) [This is the investigative equipment with which Iranian secret services could compete let above the dream.]
39. *Samci navzájem soutěží o místo na společenském žebříčku.* (SYN2005) [The males compete with each other to reach for a top social position.]

4.3.6. The verb *soupeřit* [to compete] differs from *soutěžit* [to compete] by a stylistic feature, the former is bookish, while the latter is neutral. Among 150 occurrences, 54 sentences display syntactic reciprocity, in 7 of them, the adverb *spolu* [together] is present, in 7, the expression *mezi sebou* [among/between each other] is included, see (40):

40. *Proto tyto ženy soupeřily mezi sebou v umění zalíbit se mužům.* (SYN2005) [Therefore these women competed with each other in their skills to be loved by men.]

4.3.7. We wanted to demonstrate by the illustrative material, described in Sections 4.3.1 to 4.3.6, that the power of combined the lexical and syntactical reciprocity is so strong that the speakers rarely feel the necessity to use an explicit (optional) lexical means for the reciprocity.

4.4. The expressing of the reciprocity with the verbs from the open class C, where some of their participants fulfil the conditions for reciprocalization, is a bit more complicated. The means of expression depend on the original morphemic form of a participant required by the valency frame that is moved to the subject position.

4.4.1. If the participant (Patiens or Addressee) expressed by the accusative is involved in the reciprocity relation, there are two possibilities for the syntactic reciprocalization:

(i) (True) reflexive pronoun *se* is used. The examination of the corpus material did not fully prove that in favor of avoiding ambiguity, the lexical means (*navzájem/spolu/mezi sebou*) are used regularly at least with verbs having a counterpart in a derived reciprocal (B type). Sentences (41), (42) are ambiguous as to the source of the reciprocity in accordance with our assumptions from Section 3; however, their reciprocal meaning is obvious:

41. *Ti dva se tam líbali.* [The couple kissed each other there.]
42. *Seděli vedle sebe, objímali se kolem ramen.* (SYN2005) [They were sitting next to each other and embraced each other around the shoulders.]

For the verbs without a derived reciprocal counterpart, the reciprocal meaning is transparent only in presence of a lexical means for reciprocity (see (43)) while in (44) the reciprocity is not granted:

43. *Dokonce se vzájemně fotografujeme.* (SYN2005) [Eventually, we photograph each other.]
44. *…vedou tudy koleje. Protože vlak nejede, fotografujeme se alespoň u nich.* (SYN2005)

[…there are rails here. Since the train is not coming, we at least photograph ourselves/each other by them.]

Within the sample of 250 occurrences of the verb *fotografovat* [to photograph] from SYN2005, a lexical means for reciprocity was used only twice (see (43) above).

Analyzing all occurrences of the verb *okukovat* [to take a look at] (163 in the SYN2005), we have found only one example of reciprocity, see (45):

45. *Kočky vyčkávají, navzájem se okukují.* (SYN2005) [The cats are waiting, they are taking a look at each other.]

(ii) The other expression of reciprocity is manifested by *jeden – druhý [each other]*, see (46), (47):[9]

46. *Ve škole napodobují jeden druhého.* (SYN2005) [They imitate each other at school.]
47. *…pokoušejí se obelstít jeden druhého.* (SYN2005) […they try to trick/outwit each other.]

4.4.2. The reflexive verbs (so-called "reflexiva tantum") with a participant (Patiens) expressed by genitive or dative, such as *vyhýbat se* [to avoid], *dotknout se* [to touch], *všímat si* [to notice], *zamlouvat se* [to like], *líbit se* [to like], use obligatorily the expression *jeden - druhý* [each other], see (48), (49), (50); its stylistically less natural alternative is also acceptable (see e.g. (48a)):

48. *Sousedé se léta vyhýbali jeden druhému.* [The neighbors avoided each other for whole years.]
    (a) *Sousedé se léta sobě navzájem vyhýbali.*
    (b) *? Sousedé se vyhýbali.* [The neighbors avoided.]
49. *Přistupují tiše a radostně k sobě, aniž by se dotkli jeden druhého.* (SYN2005) [They are approaching silently and happily without touching each other.]
    (a) *?…aniž by se dotkli.* […without touching themselves/each other/something.]
50. *Jan a Marie se líbí jeden druhému.* [John and Mary like each other.]
    (a) **Jan a Marie se líbí.* [*John and Mary like.]

4.4.3. The verbs with an Addressee expressed by dative, such as *blahopřát* [to congratulate], *pomáhat* [to help], *naslouchat* [to listen] have again two alternatives for expressing the syntactic reciprocity:[10]

(i) Dative form of reflexive pronoun *si* (see (51), (52)), optionally combined with one of the expressions *navzájem/vzájemně/spolu/mezi sebou*:

51. *Potvrzují, že obě ženy se navštěvovaly a blahopřály si k narozeninám.* (SYN2005) [They

[9] The expression *sebe/sobě navzájem* [Refl-long form each other] seems to be an alternative for *jeden – druhý* [each other]. They are interchangeable in all of the 53 occurrences from SYN2005. However, this expression often sounds unnaturally: Sentence *(a) Je podivuhodné, jak se mladí chlapci sobě navzájem podobají* (SYN2005) is stylistically worse than *(a') Je podivuhodné, jak se mladí chlapci jeden druhému podobají.* [It is surprising, how the young boys resemble each other.].

[10] However, the issue of *si*-derived reciprocals as an analogy to the class B remains still as an open question. It is necessary to explain why e.g. *tykat (si)* [to be on the first name terms], *vykat (si)* [to be on formal terms] need not any expression more and (a) is undoubtedly reciprocal: *(a) Profesoři a studenti si zpravidla vykají.*

confirm that the two women were visiting and congratulating each other on their birthdays.]

52. *Společně neseme následky krutého dětství a pomáháme si.* (SYN2005) [We bear together the consequences of cruel childhood and we help each other.]

(ii) The expression *jeden - druhý* [each other] in an appropriate form,[11] see (53):

53. *Naslouchali jeden druhému a zapomněli za těchto okolností na čas a prostor.* (SYN2005) [They listened to each other and they forgot the time and the space under those conditions.]

4.4.4. Verbs with the participant expressed by a prepositional case (such as *dívat se na + Accus* [to look at], *narazit na + Accus* [to bump], *křičet na + Accus* [to cry/shout], *volat na + Accus* [to shout], *ptát se na + Accus* [to ask], *předstírat před + Instr* [to pretend], *stydět se před + Instr* [to be ashamed], *smýšlet o + Loc* [to think about], *vědět o + Loc* [to know about]) have again two alternatives, analogically to Section 4.4.3:

(i) reflexive pronoun *se* in an appropriate prepositional case, optionally accompanied by the expressions *navzájem/vzájemně/spolu/mezi sebou*, see (54), (55), (56). Here we encounter the problem considered in Section 4.3 again: Though with these verbs the valency position moved into the subject is filled by the prepositional case of the pronoun *se* and it is not empty as in Section 4.3, the famous ambiguity of the reflexive *se* sometimes suggests the other than reciprocal interpretation. While in (54) a non-reciprocal interpretation would be ridiculous, ex. (55), (56) could be understood also as true-reflexives. The problem is connected with a boundary between a group and sentence coordination. The insertion of the adverb *navzájem* [each other] removes this ambiguity, see (56a):

54. *Jan a Marie na sebe narazili v kuřárně.* [John and Mary bumped at each other in the smoking room.]
55. *Hoši a dívky se před sebou stydí.* [Boys and girls are ashamed in front of each other/themselves.]
56. *Profesor A a profesor B o sobě vědí, že jsou fyzici.* [Professor A and professor B know about each other/themselves that they are physicians.]
    (a) *Profesor A a profesor B o sobě navzájem vědí, že jsou fyzici.* [Professor A and professor B know about each other that they are physicians.]

(ii) Alternatively, the expression *jeden - druhý* [each other] can be used, see (57), (58), (59):

57. *Podívali se jeden na druhého, pokrčili rameny…* (SYN2005) [They looked at each other, shrugged their shoulders…]
58. *Službu chápali oba stejně a jeden to o druhém věděli.* (SYN2005) [They both interpreted the service in the same way and they knew it about each other.]

[11] We have mentioned the peculiarity of the agreement of the parts of this expression in Note 7. There is one more peculiarity: the rule of the gender prominence (see e.g. Havránek and Jedlička, 1960) is kept here: *(a) Jan a Marie/Marie a Jan blahopřejí jeden druhému* [John and Mary/Mary and John congratulate each-Nom sg masc other-Dat sg masc], whereas *(b) Marie a Eva blahopřejí jedna druhé* [Mary and Eva congratulate each-Nom sg fem other- Dat sg fem].

59. *Ti dva mladí pitomci podezírali jeden druhého.* (SYN2005) [These two foolish guys suspected each other.]

4.5. There is one more means that could be taken into consideration as a possible expression of syntactic reciprocity; though it is possible only with some verbs, it crosses the boundary between A, B from one side and C from the other side. Dimitriadis and Milćev (2006) speak about "discontinuous reciprocals" with similar Serbian constructions (however, this term, according to our opinion, does not fit) and they point out the closeness of these constructions to the accompaniment modification. This type of construction is connected with another syntactic problem, namely the use of the *with*-constructions as an alternative expression for the coordination of sentence members. The other participant of reciprocity in these constructions is not coordinated, but it is expressed by the form typical for accompaniment (or subordinated coordination) *s* + *instrumental* [with + instrumental case], although the plural form of the predicate indicates a kind of mutuality (reciprocity). This type occurs in the corpus SYN2005 very rarely: We have found it 1x with *dotýkat se* [to touch], see (60), 4x with *navštěvovat se* [to visit each other], see (61), 4x with *objímat se* [to embrace], see (62), though there are also examples only suspected to be reciprocal (see (63)). With (64), both interpretations are acceptable because in the context there is no indication how many participants in the subject of the dependant clause are involved: it is not clear if it is only the speaker of the main clause together with *Skřivan* (then we have to do with the reciprocal reading); if somebody else is involved in the subject, we face the non-reciprocal (asymmetric) reading.

60. *Když se Stalin s Trumanem takřka dotýkali špičkami nosu, vecpal se mezi ně britský premiér.* (SYN2005) [When Stalin and Truman were nearly touching by the tips of their noses, the British prime minister squeezed between them.]
61. *S Honzou jsme se navštěvovali, jak jen to bylo možné.* (SYN2005) [lit. With Johnie we have visited each other whenever it was possible.]
62. *Objímali se s dívkou kolem pasu, (kdykoli s ní šel do parku).* (SYN2005) [lit. They embraced each other with a girl, whenever he went with her to the park]
63. *Pes vyskakoval na oba chlapce, kteří se objímali s Annou.* (SYN2005) [A dog sprung on the both boys, who embraced Anna.]
64. *Vzpomínám často, jak jsme se loučili se Skřivanem.* (SYN2005) [I often remember, how we said good bye to Skřivan.]

According to our opinion, this construction is possible with some verbs from the class C as well, e.g. *podezírat/podezřívat* [to suspect], *ujišťovat* [to assure], though we have not found any example of that type in the corpus SYN2005. However, the introspective examples (65) and (66) seem to be fully acceptable:

65. *Bratr se sestrou se podezírají, kdo z nich dopil láhev whisky.* [lit. Brother with his sister suspect each other who of them finished the bottle of whisky.]
66. *Otec se s matkou ujišťují, že se mají pořád rádi.* [lit. Father with mother assure each other that they still love each other.]

5. Conclusion

We think that the topic of Czech reciprocals has not yet been exhausted. We have proposed several issues open for further studies, e. g. the distribution of the optional lexical means, their position in word order, behavior of *si*-reflexives etc.

Recalling our ontological considerations on vagueness in syntactic reciprocal relations (see Panevová, in press, Section 4, as well as Chrakovskij, 1999[12]), our insight into the corpus material confirms for the whole domain of reciprocity that there are many vague and ambiguous constructions, interpretation of which strongly depends on inferencese provided by the speech participants with the knowledge of the broader context or situation. Our hypothesis that the use of the optional lexical means in a syntactically reciprocal construction could be redundant for the verbs from the classes A and B, while it is required (or at least preferred) for the verbs from the class C, was not fully confirmed by the corpora material. Therefore, we let speak several figures exploited from the SYN2005: In Tables 2 and 3 the figures in the column I indicate the number of the occurrences of the lemma having *se/si* on the left or on the right (not more than by 4 positions). The occurrence where syntactic reciprocity was applied is shown in the column II; the column III indicates how many occurrences from II are combined with the lexical item for reciprocity (including *jeden – druhý* [each other]).

Table 2. Selected verbs from the classes A, B

	I	**II**	**III**
dotýkat se	280[13]	68	19
objímat se	317	282	8
loučit se	207	113	2
navštěvovat se	167	87	23

Acknowledgement

The research was supported by the Ministry of Education, Youth and Sport (within the project MSM0021620838 and LC 536) and by Grant Agency of the Czech Republic (within the project 405/06/0589).

Bibliography

Chrakovskij, V. S. 1999. Diateza i referentnost. In: Teorija jazykoznanija. Rusistika. Arabistika. Sankt Peterburg, 67-101.

[12] He speaks here Chrakovskij (1999) about holistic/general meaning (celostnoje znachenije), where the particular role of the participants involved is not indicated.

[13] As to this verb, we have provided a selection of 280 occurrences from the whole number of occurrences exceeding 1 500 occurrences.

Table 3. Selected verbs from the class C

	I	**II**	**III**
obdivovat	172	4	4
hodnotit	141	4	4
popisovat	135	1	1
obviňovat	122	67	49
přesvědčovat	101	10	8
chválit	92	7	6
ujišťovat	84	19	12
urážet	76	14	6
oceňovat	67	1	1
udávat	51	5	5
pomlouvat	50	19	12
podezírat/podezřívat	46	5	4
odměňovat	28	0	0
obdarovat	15	13	11
osočovat	9	6	5
blahopřát	9	4	0

Dimitriadis, A. and T. Milćev. 2006. Symmetric and non-symmetric reciprocals in Serbo-Croatian. Handout for FASL 6.5. Nova Gorica.

Havránek, B. and A. Jedlička. 1960. Česká mluvnice [Grammar of Czech]. SPN, Praha.

Lopatková et al., 2006. Valency Lexicon of Czech Verbs: VALEX 2.0. Technical Report TR-2006-34. Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty UK, Praha.

Mikulová et al., 2005. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Technical Report 2006-30. Ústav formální a aplikované lingvistiky Matematicko-fyzikální fakulty UK, Praha.

Panevová, J. 1999. Česká reciproční zájmena a slovesná valence [Czech reciprocal pronouns and valency of verbs]. Slovo a slovesnost, 60, 269-275.

Panevová, J. 2001. Problémy reflexívního zájmena v češtině [Issues of reflexive pronoun in Czech]. In: Sborník přednášek z 44. běhu Letní školy slovanských studií. Univerzita Karlova, Filozofická fakulta, Praha, 81-88.

Panevová, J. 2006. Dvě poznámky k tzv. vágnosti [Two remarks on so-called vagueness]. In: Od fonemu do tekstu. Prace dedykowane Prof. Romanovi. Laskowskiemu (eds. I. Bobrowski, K. Kowalik).Instytut Języka Polskiego PAN, Wyd. LEXIS, Kraków, 301-304.

Panevová, J. in press. Znovu o reciprocitě [Reciprocity revisited]. Slovo a slovesnost, 68.

Štícha, F. 2003. Česko-německá srovnávací gramatika [Czech-German comparative grammar]. ARGO, Praha.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 41–60

Valency Information in VALLEX 2.0

Logical Structure of the Lexicon

Zdeněk Žabokrtský, Markéta Lopatková

Abstract

The Valency Lexicon of Czech Verbs (VALLEX 2.0) is a collection of linguistically annotated data and documentation. It provides information on the valency structure of verbs in their particular meanings / senses, possible morphological forms of their complementations and additional syntactic information, accompanied with glosses and examples. The primary goal of the following text is to briefly describe the content of VALLEX 2.0 data from a structural point of view.

1. Introduction

The Valency Lexicon of Czech Verbs, Version 2.0 (VALLEX 2.0) is a collection of linguistically annotated data and documentation, resulting from an attempt at a formal description of the valency frames of Czech verbs. VALLEX has been developed at the Institute of Formal and Applied Linguistics (ÚFAL) at Faculty of Mathematics and Physics, Charles University in Prague. VALLEX 2.0 is a successor of VALLEX 1.0, see Lopatková et al. (2003), extended in both theoretical and quantitative aspects.

1.1. Basic characteristics

VALLEX 2.0 provides information on the valency structure of verbs in their particular meanings / senses, possible morphological forms of their complementations and additional syntactic information, accompanied with glosses and examples. All lexeme entries in VALLEX are created manually; manual annotation with accent on consistency is highly time consuming and limits the speed of quantitative growth, but allows for reaching the desired quality.

VALLEX is closely related to the Prague Dependency Treebank (PDT) project, see e.g. Hajič (2005). The Functional Generative Description (FGD), being developed by Petr Sgall and his

Please cite this article as: Zdeněk Žabokrtský, Markéta Lopatková, Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. The Prague Bulletin of Mathematical Linguistics No. 87, 2007, 41–60.

collaborators since the 1960s, see esp. Sgall, Hajičová, and Panevová (1986); Hajičová, Partee, and Sgall (1998); Panevová (1974); Panevová (1994), is used as the background theory both in PDT and in VALLEX. In PDT, FGD is being verified by a complex annotation of large amounts of textual data, whereas in VALLEX it is used only for the description of the valency frames of selected verbs.

In VALLEX 2.0, there are roughly 2,730 lexeme entries containing together around 6,460 lexical units ('senses'). It is important to mention that VALLEX 2.0 – according to FGD and unlike traditional dictionaries – treats a pair of perfective and imperfective aspectual counterparts as a single lexeme. Therefore, if perfective and imperfective verbs are counted separately, the size of VALLEX 2.0 virtually grows to 4,250 entries (still without counting iteratives).

The verbs contained in VALLEX 2.0 were selected as follows: (1) We gradually processed around 2500 most frequent Czech verbs, according to the number of their occurrences in a part of the Czech National Corpus.[1] (2) Simultaneously, we added their perfective or imperfective aspectual counterparts (if they were not already present in the list of the most frequent verbs), and occasionally also iterative counterparts.

VALLEX 2.0 is issued in an electronic form available at http://ufal.mff.cuni.cz/vallex/2.0. From the very beginning, it has been designed with emphasis on both human and machine readability. Therefore, both linguists and developers of applications within the Natural Language Processing domain can use and critically evaluate its content (of course, any feedback from them will be a valuable source of information for us, as well as a great motivation for further work). In order to satisfy different needs of these different potential users, VALLEX 2.0 contains the data in the following three formats:

- **Browsable version.** The HTML version of the data allows for an easy and fast navigation through the lexicon. Lexemes and lexical units are organized in several ways, following various criteria. The screenshot of a particular lexeme is given in Figure 2 in Appendix.
- **Printable version.** The graphical layout of the lexeme entries in the printed version of the lexicon is illustrated in Figure 3 in Appendix.
- **XML version.** Programmers can run sophisticated queries (e.g. based on the XPATH query language) on this machine-tractable data, or use it in their applications.

1.2. Structure of the article

The primary goal of the following text is to briefly describe the content of VALLEX 2.0 data from a structural point of view. Linguistic issues requiring an extensive explanation or discussion are mostly left apart. However, more detailed description (and also additional relevant references) can be found in Žabokrtský (2005). Some theoretical issues concerning valency are summarized in Lopatková (2003).

The description of the VALLEX 2.0 structure is slightly simplified here, in order to correspond straightforwardly to the visual form of the lexicon and to be sufficient for its full understanding. It neglects certain features present in the underlying XML version of the lexicon,

[1] http://ucnk.ff.cuni.cz

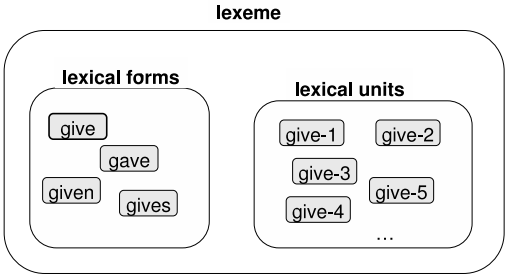

Figure 1. Illustration of the notions of lexeme, lexical form, and lexical unit.

from which both the printed and html version have been generated. Again, the details about the (slightly richer) XML structure are available in Žabokrtský (2005).

This paper is organized as follows: In Section 2, we introduce lexemes, abstract entities that constitute the lexicon on the topmost level. Section 3 deals with lexical forms and lemmas – reflexive verb forms, aspectual counterparts, lemma variants, and homographs are mentioned there. Lexical forms are associated with lexical units briefly described in Section 4. The core valency information is encoded in the valency frame; in Section 5, functors (labels for 'deep roles'), possible morphemic realizations and obligatoriness of particular valency complementations are discussed. Section 6 refers to optional attributes of a lexical unit, namely control, reflexivity, reciprocity, semantic class, and flag for idiomatic usage.

As for terminology, the terms used here either belong to the broadly accepted linguistic terminology, or come from FGD (which we have used as the background theory), or are defined somewhere else in this text.

2. Lexemes

On the highest level, VALLEX 2.0 is composed of **lexemes**. Lexeme is understood as a two-fold abstract entity: it associates a set of possible **lexical forms** (by which the presence of the lexeme is manifested in an utterance, Section 3) with a set of **lexical units** (complexes of syntactic and semantic features, LUs for short, Section 4). In simpler words, lexical forms can be viewed as the conjugated forms of a given verbal lexeme, whereas each LU corresponds roughly to the lexeme used in a specific sense and with specific syntactic combinatorial potential. This view is illustrated in Figure 1.

3. Lexical forms and lemmas

It is usual in dictionaries that the set of all possible lexical forms of a given lexeme is represented only by the infinitive form called lemma.

Lemma in VALLEX 2.0 should be considered as a complex structure:

- it always contains the 'base' infinitive form;
- it is always labeled in superscript with its morphological aspect (Section 3.2);
- it may contain also reflexive particle (e.g. *bát se* – to fear, see Section 3.1);
- it may be also labeled with a Roman number in subscript if it is necessary to distinguish it from its homograph (e.g. *nakupovat*$_I$ – to buy vs. *nakupovat*$_{II}$ – to heap, see Section 3.4).

In VALLEX 2.0, there are typically two or more lemmas listed at the beginning of the lexeme entry. It follows the FGD principle of treating aspectual counterparts (perfective and imperfective verbs expressing the same lexical meaning, Section 3.2) as manifestations of the same lexeme. Another reason for more lemmas being present in the same lexeme might be the existence of orthographic variants (Section 3.3).

3.1. Reflexive lemmas

In VALLEX 2.0, two types of reflexive constructions are distinguished:

- Reflexive lexemes – both 'reflexiva tantum' (e.g. *bát se* – to fear, *smát se* – to laugh) and derived reflexives (e.g. *odpovídat se* – to account, *šířit se* – to spread, *vrátit se* – to return) are represented as separate lexemes, and the reflexive particles *se* or *si* are considered as parts of their lemmas.
- Reflexive usage of irreflexive lexemes – if the reflexive particles/pronouns *se* or *si* have specific syntactic function(s), reflexive forms of particular verbs are treated within irreflexive lexemes and their possible functions are specified (see Sections 6.2 and 6.3) – *se* or *si* can be a part of the reflexive passive form (e.g. in *pátrá se po zloději* – a thief is being looked for); it can be a complementation fulfilling some valency slot of the governing verb (e.g. *mýt se* – to wash oneself, where *se* is PAT (Patient) coreferential with ACT (actor)), or it can mark reciprocity (e.g. *kopat se* in *kopou se vzájemně do nohou* – they kick each other's legs).

3.2. Aspectual counterparts

Imperfective and perfective verb forms are distinguished in Czech (as well as a specific subclasses of iterative verbs and so called biaspectual verbs); this characteristic is called aspect.

In VALLEX 2.0, the value of aspect is attached to each lemma as a superscript label:

- *impf* for imperfective;
- *pf* for perfective;
- *iter* for iterative verbs;
- *biasp* for biaspectual verbs.

There are three ways how aspectual counterparts (verbs with the same or very similar lexical meaning differing in aspect) are formed in Czech (sorted according to productivity):

- *affixation*: an imperfective verb is derived from the perfective one, e.g. by infix *-ova-*, *vypsat* → *vypisovat* – to excerpt, to write off;

- *prefixation*: a perfective verb is derived from the imperfective one by adding a prefix, e.g. *psát* → *napsat* – to write;
- suppletive (phonemically unrelated) couples: *vzít* / *brát* – to take.

Aspectual counterparts of the first and third type constitute a single lexeme in VALLEX 2.0, as e.g. in the case of *nasedat*^{impf}, *nasednout*^{pf}, *nasedávat*^{iter} – to get on.

As already mentioned, a LU typically shares all its lemmas with the other LUs in the lexeme in which it is embedded. However, there are exceptions: the aspectual counterpart(s) need not be the same for all LUs of the particular lexeme. For example, *odpovědět*^{pf} is a counterpart of *odpovídat*^{impf} in the sense 'to answer', but not in the sense 'to correspond'. In such cases, the set of applicable lemmas is specified directly for the LU (and overrides the set of lemmas specified for the whole lexeme).

There might be more than one lemma with the same aspect in a lexeme (without being lemma variants, see Section 3.3). Then the aspect flags are distinguished by Arabic numbers, as e.g. in the lexeme *osušovat*^{impf1}, *osoušet*^{impf2}, *osušit*^{pf} – to dry up, to wipe, or *odřezávat*^{impf}, *odříznout*^{pf1}, *odřezat*^{pf2} – to cut off (unique aspect flags are necessary because they serve also for co-indexing the lemmas with example sentences illustrating the usage of the lexeme).

Some verbs (e.g. *informovat* – to inform, *charakterizovat* – to characterize) can be used in different contexts either as imperfective or as perfective. They are called biaspectual verbs.

Within imperfective verbs, there is a subclass of iterative verbs (iter.). Czech iterative verbs are derived more or less in a regular way by affixes such as *-va-* or *-íva-*, and express extended and repetitive actions (e.g. *číst* – to read → *čítávat*, *chodit* – to walk → *chodívat*). In VALLEX 2.0, iterative verbs containing double affix *-va-* (e.g. *chodívávat*) are completely disregarded, whereas the remaining iterative verbs occur as headword lemmas of the relevant lexeme.

3.3. Lemma variants

Lemma variants (many of which are just spelling variants, i.e. orthographic variants) are groups of two or more lemmas that are interchangeable in any context without any change of the meaning (e.g. *dovědět se/dozvědět se* – to learn). Usually, the only difference is just a small alternation in the morphological stem, which might be accompanied by a subtle stylistic shift (e.g. *myslet/myslit* – to think, the latter one being bookish). Moreover, although the infinitive forms of the variants differ in spelling, some of their conjugated forms might be identical (*mysli* (imper.sg.) both for *myslet* and *myslit*).

There are rare exceptions when only one of the variants can be used, e.g. *plavat* and *plovat* – to swim, are usually considered to be variants, see, e.g. SSJČ (1964), although, in some contexts, only *plavat*, in the sense 'to flounder', can be used (*plavat při zkoušce*, **plovat při zkoušce*). The applicable lemmas must be then listed for the specific LU as in any other cases when a LU imposes a further limitation on the set of lexical forms.

3.4. Homographs

Homographs are lemmas 'accidentally' identical in the spelling but considerably different in their meaning (there is no obvious semantic relation between them). They also might differ as to their etymology (e.g. *nakupovat*$_I$ – to buy vs. *nakupovat*$_{II}$ – to heap), aspect (Section 3.2) (e.g. *stačit*$_I$ pf. – to be enough vs. *stačit*$_{II}$ impf. – to catch up with), or conjugated forms (*žilo* (past.sg.fem) for *žít*$_I$ – to live vs. *žalo* (past.sg.fem) *žít*$_{II}$ – to mow. In VALLEX 2.0, such lemmas are distinguished by Roman numbering in the subscript. These numbers should be understood as inseparable parts of VALLEX 2.0 lemmas.

4. Lexical units

Each lexeme is formed by a set of lexical units that are assigned to respective lexical forms (represented by their lemmas). Following Cruse (1986), we understand lexical units (LUs) as "form-meaning complexes with (relatively) stable and discrete semantic properties". Roughly speaking, LU can be understood as 'a given word in the given sense'. In the Czech tradition, this concept of LU corresponds to Filipec's 'monosemic lexeme', see Filipec and Čermák (1985).

Within each lexeme in VALLEX 2.0, LUs are numbered by Arabic numbers. In the printed and html versions of the lexicon, the LU entry starts with its number.

The ordering of lexical units is not completely random, but it is not perfectly systematic either. So far, it is based only on the following weak intuition: the primary and/or the most frequent meanings should go first, whereas rare and/or idiomatic meanings should go last. (We do not guarantee that the ordering of LUs in VALLEX 2.0 exactly matches their frequency in the contemporary language.)

By default, a LU 'inherits' all lemmas specified for the given lexeme in which it is embedded. However, it might happen that for a given LU not all the forms specified for the whole lexeme are applicable. In such cases, the list of applicable lemmas is specified for the given LU separately.

Available information about each LU entry in VALLEX 2.0 is captured by obligatory and optional attributes. The former ones have to be filled with every LU. The latter ones might be empty, either because they are not applicable (e.g. no control can be applicable for verbs without infinitive complementations), or because the annotation was not finished yet (e.g. attribute class, Section 6.4).

Obligatory LU attributes:

- valency frame (Section 5);
- gloss – verb or paraphrase roughly synonymous with the given sense/meaning; this attribute is not supposed to serve as a source of synonyms or even of genuine lexicographic definition – it should be used just as a clue for fast orientation within the word entry!
- example – sentence(s) or sentence fragment(s) containing the given verb used with the given valency frame.

Optional LU attributes:

- flag for idiom (Section 6.5);
- information on control (Section 6.1);

- possible type(s) of reflexive constructions (Section 6.2);
- possible type(s) of reciprocal constructions (Section 6.3);
- affiliation to a syntactico-semantic class (Section 6.4).

In the printable version (see Figure 3 in Appendix), the gloss is located in parentheses at the beginning of every LU entry, and then the valency frame is printed. Example sentence follows the diamond sign, and the optional attributes (if any) are given after the cross sign. If more lemmas are relevant for the given lexeme (as it is often the case because of aspectual pairs), it might be necessary to give more values also in the attribute (especially in the example attribute). The correspondence between the respective values and the relevant lemmas is captured by superscript labels ^{pf}, ^{impf}, ^{pf1} etc.

5. Valency frames

The core valency information is encoded in the **valency frame**. Within the FGD framework, valency frames (in a narrow sense) consist only of inner participants (both obligatory and optional) and obligatory free modifications, see Panevová (1974); Panevová (1994). In VALLEX 2.0, valency frames are enriched with quasi-valency complementations. Moreover, a few non-obligatory free modifications occur in valency frames too, since they are typically related to some verbs (or even to whole classes of them) and not to others.[2]

In VALLEX 2.0, a valency frame is modeled as a sequence of frame slots. Each frame slot corresponds to one (either required or specifically permitted) complementation of the given verb.

Note on terminology: in this text, the term 'complementation' (dependent item) is used in its broad sense, not related to the traditional argument/adjunct (complement/modifier) dichotomy.

The following attributes are assigned to each slot:

- functor (Section 5.1);
- list of possible morphemic forms (realizations) (Section 5.2);
- type of complementation (Section 5.3).

Some slots tend to occur systematically together. In order to capture this type of regularity, we have introduced the mechanism of slot expansion, Section 5.4 (full valency frame is obtained after performing these expansions).

5.1. Functors

In VALLEX 2.0, functors (labels for 'deep roles'; similar to theta-roles) are used for expressing types of relations between verbs and their complementations. According to FGD, functors are divided into inner participants (actants) and free modifications (this division roughly corresponds to the argument/adjunct dichotomy), see Panevová (1974); Panevová (1994). In

[2] The other free modifications can occur with the given verb too, but they are not contained in the valency frame as their presence in a sentence is not understood as syntactically conditioned in FGD.

VALLEX 2.0, we also distinguish an additional group of quasi-valency complementations, see esp. Lopatková and Panevová (2005).

Functors that occur in VALLEX 2.0 are listed in the following tables (for Czech sample sentences see Lopatková et al., 2002, page 43):

Inner participants:

- ACT (actor): *Peter read a letter.*
- ADDR (addressee): *Peter gave Mary a book.*
- PAT (patient): *I saw him.*
- EFF (effect): *We made her the secretary.*
- ORIG (origin): *She made a cake from apples.*

Quasi-valency complementations:

- DIFF (difference): *The value of shares has risen by 100%.*
- OBST(obstacle): *The boy stumbled over a stump.*
- INTT (intent): *He came there to look for Jane.*

Free modifications:

- ACMP (accompaniment): *Mother came with her children.*
- AIM (aim): *John came to a bakery for a piece of bread.*
- BEN (benefactive): *She made this for her children.*
- CAUS (cause): *She did so since they wanted it.*
- COMPL (complement): *They painted the wall blue.*
- CRIT (criterion): *Peter has to do it exactly according to directions.*
- DIR1 (direction-from): *He went from the forest to the village.*
- DIR2 (direction-through): *He went through the forest to the village.*
- DIR3 (direction-to): *He went from the forest to the village.*
- DPHR (dependent part of a phraseme): *Peter talked horse again.*
- EXT (extent): *The temperatures reached an all time high.*
- HER (heritage): *He named the new villa after his wife.*
- LOC (locative): *He was born in Italy.*
- MANN (manner): *They did it quickly.*
- MEANS (means): *He wrote it by hand.*
- RCMP (recompense): *She bought a new shirt for 25 $.*
- REG (regard): *With regard to George she asked his teacher for advice.*
- SUBS (substitution): *He went to the theater instead of his ill sister.*
- TFHL (temporal-for-how-long): *They interrupted their studies for a year.*
- TFRWH (temporal-from-when): *His bad reminiscences came from this period.*

- THL (temporal-how-long): *We were there for three weeks.*
- TOWH (temporal-to when): *He put it over to next Tuesday.*
- TSIN (temporal-since-when): *I have not heard about him since that time.*
- TTIL (temporal-till-when): *It will last till 5 o'clock.*
- TWHEN (temporal-when): *He will come tomorrow.*

Note 1: Besides the functors listed in the tables above, also value DIR occurs in the VALLEX 2.0 data. It is used only as a special symbol for the slot expansion (Section 5.4).

Note 2: The set of functors as introduced in FGD and used in the Prague Dependency Treebank is richer than that shown above, see Mikulová et al. (2006). We do not use its full (current) set in VALLEX 2.0 due to several reasons. Some functors do not occur with verbs at all (e.g. MAT – material, partitive, as *sklenice piva.MAT* – glass of beer), some other functors can occur there but represent other than dependency relations (e.g. coordination, *Jim nebo.CONJ Jack* – Jim or Jack). And still others can occur with verbs as well but their behavior is absolutely independent of the head verb; thus they have nothing to do with valency frames (e.g. ATT – attitude, *udělal to dobrovolně.ATT* – he did it willingly).

5.2. Morphemic forms

In a sentence, each frame slot can be expressed by a limited set of morphemic means which we call forms. In VALLEX 2.0, the set of possible forms (supposing active verb form) is defined either explicitly, or implicitly.

In the first case (explicitly declared forms), the forms are enumerated in a list attached to the given slot (in the case of arguments and quasi-valency complementations, no other forms can be used; in the case of free modifiers, the possible forms are not necessarily limited to those given in the list).

In the second case (implicitly declared forms), no such list is specified because the set of possible forms is implied by the functor of the respective slot (in other words, all forms possibly expressing the given functor may appear).

5.2.1. Explicitly declared forms

The list of forms attached to a frame slot may contain values of the following types:

- **Pure (prepositionless) case.** There are seven morphological cases in Czech. In the VALLEX 2.0 notation, we use numbering traditional in the Czech linguistics: 1 – nominative, 2 – genitive, 3 – dative, 4 – accusative, 5 – vocative, 6 – locative, and 7 – instrumental.
- **Prepositional case.** Lemma of the preposition (i.e. preposition without vocalization) and the number of the required morphological case are specified (e.g. *z+2*, *na+4*, *o+6*, …). The prepositions occurring in VALLEX 2.0 are the following: *bez, do, jako*[3], *k, kolem,*

[3]Word *jako* is traditionally considered as a conjunction, but it is included in this list as it requires a particular morphological case in some valency frames

mezi, místo, na, nad, o, od, po, pod, podle, pro, proti, před, přes, při, s, u, v, z, za.

- **Infinitive construction.** The abbreviation 'inf' stands for infinitive verbal complementation; 'inf' can appear together with a conjunction (e.g. *než*+inf), but it happens very rarely in Czech.
- **Subordinated clauses.** Subordinated content clauses introduced by subordinating conjunctions are represented by the conjunction lemmas; the following values occur in VALLEX 2.0: *aby, ať, až, jak, zda,*[4] *že*.
  Subordinated content clauses not introduced by a conjunction (e.g. those having the form of an indirect speech with an interrogative pronoun or pronominal adverb) are represented by the abbreviation 'cont'.
- **Construction with adjectives.** Abbreviation 'adj-digit' stands for an adjective complementation in the given case, e.g. adj-1 (e.g. *cítím se slabý* – I feel weak).
- **Constructions with *být*.** Infinitive of verb *být* (to be) may combine with some of the types above, e.g. *být*+adj-1 (e.g. *zdá se to být dostatečné* – it seems to be sufficient).
- **Part of phraseme.** If the set of the possible lexical values of the given complementation is very small (often one-element), we list these values directly (e.g. *napospas* for the phraseme *ponechat napospas* – to expose).

5.2.2. Implicitly declared forms

If no forms are listed explicitly for a frame slot, then the list of possible forms implicitly results from the functor of the slot according to the following (yet incomplete) lists:

- ACMP: bez+2, s+7, společně s+7, spolu s+7, v čele s+7, v souvislosti s+7, ve spojení s+7, včetně+2, …;
- AIM: aby, ať, do+2, k+3, na+4, o+4, pro+4, pro případ+2, proti+3, v zájmu+2, za+4, za+7, že, …;
- BEN: 3, na+4, na účet+2, na úkor+2, na vrub+2, pro+4, proti+3, v+4, ve prospěch+2, v rozporu, s+7, v zájmu+2 …;
- CAUS: 7, aby, adverb, díky+3, jelikož, ježto, k+7, kvůli+3, na+4, na+6, na základě+2, nad+7, následkem+2, od+2, pod+7, pod náporem+2, pod tíhou+2, pod váhou+2, poněvadž, pro+4, proto, protože, v+6, v důsledku+2, v souvislosti s+7, vinou+2, vlivem+2, vzhledem k+3, z+2, z důvodu+2, za+4, za+7, zásluhou+2, že, …;
- CRIT: 7, 2, dle+2, podle+2, na+6, na základě+2, po vzoru+2, přiměřeně+3, v+6, v duchu+2, v rozporu s+7, v souladu s+7, v souhlase s+7, v závislosti na+6, ve shodě s+7, ve smyslu+2, ve světle+2, z titulu+2, …;
- DIR1: adverb, od+2, s+2, z+2, ze strany+2, zpod+2, zpoza+2, zpřed+2, …;
- DIR2: 7, adverb, kolem+2, cestou+2, mezi+7, napříč+7, po+6, podél+2, přes+4, skrz+4, v+6, …

[4] Form *zda* is in fact an abbreviation for the couple of conjunctions *zda* and *jestli*.

- DIR3: 7, adverb, do+2, do čela+2, k+3, kolem+2, mezi+4, mimo+4, na+4, na+6, nad+4, naproti+3, okolo+2, po+4, po+6, pod+4, proti+3, před+4, přes+4, směrem do+2, směrem k+3, směrem na+4, v+4, vedle+2, za+4, za+7, …;
- EXT: adverb, 2, 4, 7, do+2, kolem+2, k+3, na+4, na+6, nad+4, okolo+2, po+6, pod+7, přes+4, v+4, z+2, za+4, …;
- LOC: adverb, blízko+2, blízko+3, daleko+2, do+2, kolem+2, mezi+7, mimo+4, na+4, na+6, na úroveň+2, nad+7, naproti+3, nedaleko+2, okolo+2, po+6, po bok+2, poblíž+2, pod+7, podél+2, proti+3, před+7, přes+4, při+6, stranou+2, u+2, uprostřed+2, uvnitř+2, v+6, v čele+2, v oblasti+2, v rámci+2, v řadě+2, vedle+2, za+4, za+7, …;
- MANN: 7, adverb, do+2, formou+2, na+4, na+6, nad+4, o+4, po+6, pod+7, proti+3, před+7, při+6, přes+4, s+7, v+4, v+6, v podobě+2, ve formě+2, vedle+2, z+2, za+4, za+7, jak, že …;
- MEANS: adverb, 7, cestou+2, díky+3, do+2, na+4, na+6, o+6, po+6, pod+7, pomocí+2, prostřednictvím+2, přes+4, s+7, s pomocí+2, v+6, z+2, za+4, skrz+2, za pomoci+2, že, …;
- REG: adverb, 7, bez ohledu na+4, bez zřetele k+3, k+3, kolem+2, na+4, na+6, na téma+2, nad+7, nezávisle na+6, o+6, ohledně+2, po+6, pro+4, před+7, při+6, s+7, se zřetelem k+3, se zřetelem na+4, s ohledem na+4, u+2, v+6, v otázce+2, v případě+2, v rámci+2, v souvislosti s+7, ve věci+2, ve vztahu k+3, vůči+3, vzhledem k+3, z+2, z hlediska+2, za+4, …;
- SUBS: jménem+2, namísto+2, místo+2, výměnou za+4, za+4, …;
- TFHL: adverb, do+2, na+4, po+2, pro+4, …;
- TFRWH: z+2, od+2, …;
- THL: adverb, 2, 4, 7, až, dokud, do+2, na+4, po+4, po dobu+2, přes+4, v+2, za+4, …;
- TOWH: adverb, do+2, k+3, na+4, pro+4, …;
- TSIN: adverb, od+2, počínaje+7, z+2, …;
- TTILL: adverb, do+2, dokud, k+3, než, po+4, …;
- TWHEN: 2, 4, 7, adverb, až, do+2, jakmile, k+3, když, kolem+2, koncem+2, mezi+7, na+4, na+6, na závěr+2, než, o+6, okolo+2, po+6, počátkem+2, postupem+2, poté co, před+7, předtím než, při+6, s+7, u příležitosti+2, v+4, v+6, v době+2, v období+2, v průběhu+2, v závěru, z+2, za+2, za+4, začátkem, …;

5.3. Types of complementations

Within the FGD framework, valency frames (in a narrow sense) consist only of inner participants (both obligatory[5] and optional) and obligatory free modifications; the dialogue test was introduced by Panevová (1974) as a criterion for obligatoriness, see also Sgall, Hajičová,

[5] It should be emphasized that in this context the term obligatoriness is related to the presence of the given complementation in the deep (tectogrammatical) structure, and not to its (surface) deletability in a sentence (moreover, the

and Panevová (1986). In VALLEX 2.0, valency frames are enriched with quasi-valency complementations. Moreover, a few non-obligatory free modifications occur in valency frames too, since they are typically related to some verbs (or even to whole classes of them) and not to others.

The attribute 'type' is attached to each frame slot and can have one of the following values: 'obl' or 'opt' for inner participants and quasi-valency complementations, and 'obl' or 'typ' for free modifications. In the printed version, optional complementations are marked with '?', whereas typical complementations are marked with '¿'.

5.4. Slot expansion

Some slots tend to occur systematically together. For instance, verbs of motion can be often modified with direction-to and/or direction-through and/or direction-from modifier. We decided to capture this type of regularity by introducing the abbreviation flag for a slot. If this flag is set (in the VALLEX 2.0 notation it is marked with an upward arrow), the full valency frame is obtained after slot expansion.

If one of the frame slots is marked with the upward arrow (in the XML data, attribute 'abbrev' is set to 1), then the full valency frame will be obtained after substituting this slot with a sequence of slots as follows:

- $\uparrow \mathrm{DIR}^{typ} \rightarrow \mathrm{DIR1}^{typ}\ \mathrm{DIR2}^{typ}\ \mathrm{DIR3}^{typ}$
- $\uparrow \mathrm{DIR1}^{obl} \rightarrow \mathrm{DIR1}^{obl}\ \mathrm{DIR2}^{typ}\ \mathrm{DIR3}^{typ}$
- $\uparrow \mathrm{DIR2}^{obl} \rightarrow \mathrm{DIR1}^{typ}\ \mathrm{DIR2}^{obl}\ \mathrm{DIR3}^{typ}$
- $\uparrow \mathrm{DIR3}^{obl} \rightarrow \mathrm{DIR1}^{typ}\ \mathrm{DIR2}^{typ}\ \mathrm{DIR3}^{obl}$
- $\uparrow \mathrm{THL}^{typ} \rightarrow \mathrm{TSIN}^{typ}\ \mathrm{THL}^{typ}\ \mathrm{TTIL}^{typ}$

6. Optional LU attributes

6.1. Control

The term 'control' relates in this context to a certain type of predicates (verbs of control) and two coreferential expressions, a 'controller' and a 'controllee', see also Panevová (1996). In VALLEX 2.0, control is captured in the data only in the situation in which a verb has an infinitive modifier (regardless of its functor). Then the controllee is an element that would be a 'subject' of the infinitive (which is structurally excluded on the surface), and controller is the co-indexed expression. In VALLEX 2.0, the type of control is stored in the frame attribute 'control' as follows:

- if there is a coreferential relation between the (unexpressed) subject ('controllee') of the infinitive verb and one of the frame slots of the head verb, then the attribute is filled with the functor of this slot ('controller');

relation between deep obligatoriness and surface deletability is not at all straightforward in Czech).

- otherwise (i.e., if there is no such coreference), value 'ex' is used.

Examples:

- *pokusit se* – to try, e.g. *Jiří se pokusí přijít* – Jiří will try to come, control: ACT;
- *slyšet* – to hear, e.g. *děti slyší někoho přicházet* – children hear somebody coming, control: PAT;
- *jít*, in the sense *jde to udělat* – it is possible to do it, control: ex.

6.2. Reflexivity

The optional attribute reflexivity (abbreviation 'rfl') indicates possible syntactic functions of the reflexive particles/pronouns *se* or *si*.

The reflexive particles/pronouns *se* or *si* are used in Czech as formal means expressing the following syntactic constructions:

- derived diatheses: the particle *se* is a part of the reflexive passive verb form:
  - for transitive verbs (e.g *plány se připravují* – plans are prepared); marked with the label 'pass';
  - for intransitive verbs (e.g. *pátrá se po zloději* – a thief is being looked for; *v neděli se chodí do kostela* – on Sundays one visits the church); marked with the label 'pass0'.
- grammatical coreference: the pronouns *se* or *si* stands for an inner participant that is coreferential with Actor (e.g. *mýt se* – to wash oneself, coreference between ACT and PAT (in Accusative); *podřídit si zaměstnance* – to bring under the employees, coreference between ACT and ADDR in dative); marked with the labels 'cor3' (in the case of *si*) or 'cor4' (in the case of *se*).

Note that the attribute reflexivity does not cover reflexive verb forms where reflexive particles *se* or *si* are parts of the infinitive forms, i.e. reflexiva tantum (e.g. *bát se* – to fear, *smát se* – to laugh) as well as derived reflexive (e.g. *odpovídat se* – to account, *šířit se* – to spread, *vrátit se* – to return) (as already discussed in Section 3.1), nor the reciprocal function of *se* or *si* pronouns (see the following Section).

6.3. Reciprocity

Reciprocity is understood as a possibility of (two or more) valency complementations to be in relations with each other that may be viewed symmetrically (and their roles are interchangeable).

In Czech, if Actor and some other complementation are reciprocal, then the reflexive verb form is used and these two complementations are expressed either as a coordinated nominal group (as in *Petr a Marie se hádali* – Peter and Mary argued (with one another)), or as a plural noun (*přátelé se navštěvují* – friends visit each other), possibly with additional adverbs *spolu*, *navzájem*, ….

If Actor is not affected, the reciprocity may follow from the plural form or coordination (with no other formal sign), as in *seznámil je* – he introduced them (to each other).

The possibility of reciprocal usage is indicated in the attribute reciprocity ('rcp' for short), the value of which is a pair (or triple) of functors involved, e.g. ACT-ADDR for *hádat se* – to argue, *neustále se spolu hádali* – they argued with each other all the time; or ACT-ADDR-PAT for *mluvit* – to talk, *mluví spolu o sobě* – they talked with each other about themselves.

In the case of derived reflexive lexemes of inherently reciprocal verbs (with the obligatory complementation in the form s+7), both LUs for irreflexive and reflexive lexemes are assigned attribute 'rcp'.

Examples:

- ACT-PAT for *navštěvovat, navštívit* (*impf navštěvovali se vzájemně*, *pf navštívit se navzájem* – they visited each other);
- ACT-PAT for *navštěvovat se* (*navštěvovali se pravidelně celá léta* – they visited each other for all the years).

6.4. Semantic class

A significant part of lexical units (2,903 LUs out of 6,460, i.e. 45% of all LUs) is assigned with semantic classes. These classes were built strictly in a 'bottom-up' way, by grouping LUs with similar syntactic property and with respect to their semantics. The following 22 semantic classes were established:

- appoint verb (23 LUs), e.g. *nominovat* – to nominate, *určovat, určit* – to assign (as in *určila ho za svého zástupce* – she assigned him as her assistant), *ustanovovat, ustanovit* – to appoint, …;
- cause motion (43 LUs), e.g. *hýbat, hnout, hýbnout* – to move (as in *hnul pravou rukou* – he moved his right hand), *mávat, mávnout* – to wave, *vrhat* - to throw, …;
- combining (96 LUs), e.g. *míchat* – to mix, *přidat, přidávat* – to add, *spojit, spojovat* – to join/to combine, …;
- communication (364 LUs), e.g. *číst* – to read, *hovořit* – to talk, *nařizovat, nařídit* – to command, *pochybovat* – to hesitate/to question, …;
- contact (115 LUs), e.g. *dotýkat se, dotknout se* – to contact, *narážet, narazit* – to hit (against sth), *tisknout* – to press, …;
- emission (22 LUs), e.g. *pouštět, pustit* – to run (as in *tričko pustilo barvu* – the shirt lost color), *vysílat, vyslat* – to radiate/to emit, …;
- exchange (177 LUs), e.g. *dávat, dát* – to give, *dostávat, dostat* – to get, *platit* – to pay, *pronajímat, pronajmout* – to let, …;
- expansion (19 LUs), e.g. *pronikat, proniknout* – to spread, *šířit* – to diffuse/to disseminate, …;
- extent (20 LUs), e.g. *činit* – to amount, *dosahovat, dosáhnout* – to reach, *vycházet, vyjít* – to cost/to come to (as in *boty vyjdou na tisíc korun* – shoes come to one thousand crowns), …;
- change (318 LUs), e.g. *budovat* – to build, *klesat, klesnout* – to fall (as in *teplota klesla*

pod bod mrazu – the temperature fell below freezing point), *proměňovat, proměnit* – to change, *růst* – to grow, *vytvářet, vytvořit* – to create, …;

- intervention (10 LUs), e.g. *zasahovat* – to meddle, *mluvit* – to speak/to interfere (as in *do toho nemůžu mluvit* – I have no voice in this), …;
- location (399 LUs), e.g. *doplňovat, doplnit* – to add, *nacházet, najít* – to find, *shromažďovat* – to gather, …;
- mental action (304 LUs), e.g. *cítit se* – to feel (as in *cítit se dobře* – to feel fine), *jásat* – to exult, *mrzet* – to be sorry, …;
- modal verb (15 LUs), e.g. *dovést* – to be able, *chtít* – to want, …;
- motion (309 LUs), e.g. *běžet* – to run, *dorážet, dorazit* – to arrive, *hýbat se* – to move (as in *Nehýbej se!* – Don't move!), …;
- perception (104 LUs), e.g. *hledět* – to look, *pamatovat* – remember, *všímat se, všimnout si* – to notice, …;
- phase of action (80 LUs), e.g. *končit* – to end (as in *zde les končí* – here the forest ends), *vrcholit* – to culminate, *vznikat, vzniknout* – to arise, …;
- phase verb (76 LUs), e.g. *iniciovat* – to initiate, *končit* – to end (as in *končit školu* – to finish the school), *najet* – to cover (as in *najeli aspoň 500 mil* – they covered at least 500 miles), …;
- providing (51 LUs), e.g. *naplňovat, naplnit* – to fill/to replentish, *oloupávat, olupovat, oloupnout, oloupat* – to peel (as in *oloupat ovoce* – to peel fruit), *vybavovat, vybavit* – to equip, …;
- psych verb (83 LUs), e.g. *klamat* – to deceive, *těšit* – to pleasure, …;
- social interaction (86 LUs), e.g. *potkávat se, potkat se* – to meet (as in *potkává se s přáteli v baru* – he used to meet his friends in bar), *spojovat se, spojit se* – to interconnect/to get in touch (as in *spojím se s ním co nejdříve* – I will get in touch with him as soon as possible), *souhlasit* – to agree, …;
- transport (189 LUs), e.g. *donášet, donést* – to bring/to carry, *přemisťovat/přemísťovat, přemístit* – to move, *shrnovat, shrnout* – to heap, ….

We admit that this classification is tentative and should be understood merely as an intuitive gathering of frames, rather than a properly defined ontology. The motivation for introducing such semantic classification in VALLEX 2.0 was the fact that it simplifies systematic checking of consistency and allows for making more general observations about the data.

6.5. Idioms

When building VALLEX, we have focused mainly on primary or usual meanings of verbs. We also noted many LUs corresponding to peripheral usages of verbs. However, their coverage in VALLEX might not be complete. We call such LUs idiomatic and mark them with the label 'idiom'. An idiomatic frame is tentatively characterized either by a substantial shift in

meaning (with respect to the primary sense), or by a small and strictly limited set of possible lexical values in one of its complementations, or by occurrence of another type of irregularity or anomaly.

7. Final remarks

The preparation of the presented version of VALLEX has taken more than five years. The primordial aim of this work was to create a publicly available high-quality NLP-oriented lexical resource focused on valency properties of Czech verbs. We believe that this goal has been achieved: VALLEX 2.0 is a formally structured large-coverage lexicon available in both human readable and machine tractable form. We also hope that our attempt at accumulating dispersed linguistic knowledge relevant for valency, as well as the stress laid on the consistency of the description of regular properties of lexical units, have contributed to the user value of the lexicon. On the other hand, the data-oriented approach to valency inquiries shows that there are still open theoretical questions requiring further linguistic research.

Acknowledgement

When creating VALLEX 2.0, we have used the following Czech dictionaries (some of them via the dictionary browser DEBDict[6]):

- BRIEF, Pala and Ševeček (1997);
- Slovník spisovné češtiny, SSČ (2003);
- Slovník spisovného jazyka českého, SSJČ (1964);
- Slovesa pro praxi, Svozilová, Prouzová, and Jirsová (1997);
- Slovník slovesných, substantivních a adjektivních vazeb a spojení, Svozilová, Prouzová, and Jirsová (2005).

Many thanks for an extensive linguistic and also technical advice go to our colleagues from ÚFAL, especially to Professor Jarmila Panevová.

VALLEX 2.0 has been carried out under the project of the Ministry of Education, Youth and Sports of the Czech Republic No. MSM0021620838 (Objects of Research). It was partially supported also by the project of the Ministry of Education, Youth and Sports of the Czech Republic No. LC 536 and by the projects of the Academy of Science of the Czech Republic, Information Society No. 1ET100300517 and No. 1ET101120503.

Bibliography

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

Filipec, Josef and František Čermák. 1985. *Česká lexikologie*. Academia, Prague.

[6] http://nlp.fi.muni.cz/projekty/deb2/debdict/index.html

Hajič, Jan. 2005. Complex Corpus Annotation: The Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*. Veda Bratislava, pages 54–73.

Hajičová, Eva, Barbara H. Partee, and Petr Sgall. 1998. *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*, volume 71 of *Studies in Linguistics and Philosophy*. Kluwer, Dordrecht.

Lopatková, Markéta. 2003. Valency in the Prague Dependency Treebank: Building the Valency Lexicon. *The Prague Bulletin of Mathematical Linguistics*, (79–80):37–60.

Lopatková, Markéta and Jarmila Panevová. 2005. Recent Developments in the Theory of Valency in the Light of the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*. Veda Bratislava, pages 83–92.

Lopatková, Markéta, Zdeněk Žabokrtský, Karolína Skwarska, and Václava Benešová. 2002. Tektogramaticky anotovaný valenční slovník českých sloves. Technical Report TR-2002-15, ÚFAL/CKL MFF UK, Prague.

Lopatková, Markéta, Zdeněk Žabokrtský, Karolína Skwarska, and Václava Benešová. 2003. VALLEX 1.0 Valency Lexicon of Czech Verbs. Technical Report TR-2003-18, ÚFAL/CKL MFF UK, Prague.

Mikulová, Marie, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report TR-2006-30, ÚFAL MFF UK, Prague.

Pala, Karel and Pavel Ševeček. 1997. Valence českých sloves. In *Sborník prací FFBU*, pages 41–54, Brno.

Panevová, Jarmila. 1974. On Verbal Frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, (22):3–40.

Panevová, Jarmila. 1994. Valency Frames and the Meaning of the Sentence. In Philip A. Luelsdorff, editor, *The Prague School of Structural and Functional Linguistics*. John Benjamins Publishing Company, pages 223–243.

Panevová, Jarmila. 1996. More Remarks on Control. *Prague Linguistic Circle Papers, John Benjamins*, 2:101–120.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

SSJČ. 1964. *Slovník spisovného jazyka českého*. Academia, Prague.

SSČ. 2003. *Slovník spisovné češtiny pro školu a veřejnost*. Academia, Prague. (3^{rd} edition).

Svozilová, Naďa, Hana Prouzová, and Anna Jirsová. 1997. *Slovesa pro praxi*. Academia, Prague.

Svozilová, Naďa, Hana Prouzová, and Anna Jirsová. 2005. *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Academia, Prague.

Žabokrtský, Zdeněk. 2005. *Valency Lexicon of Czech Verbs*. Ph.D. thesis, Charles University, Prague.

Appendix

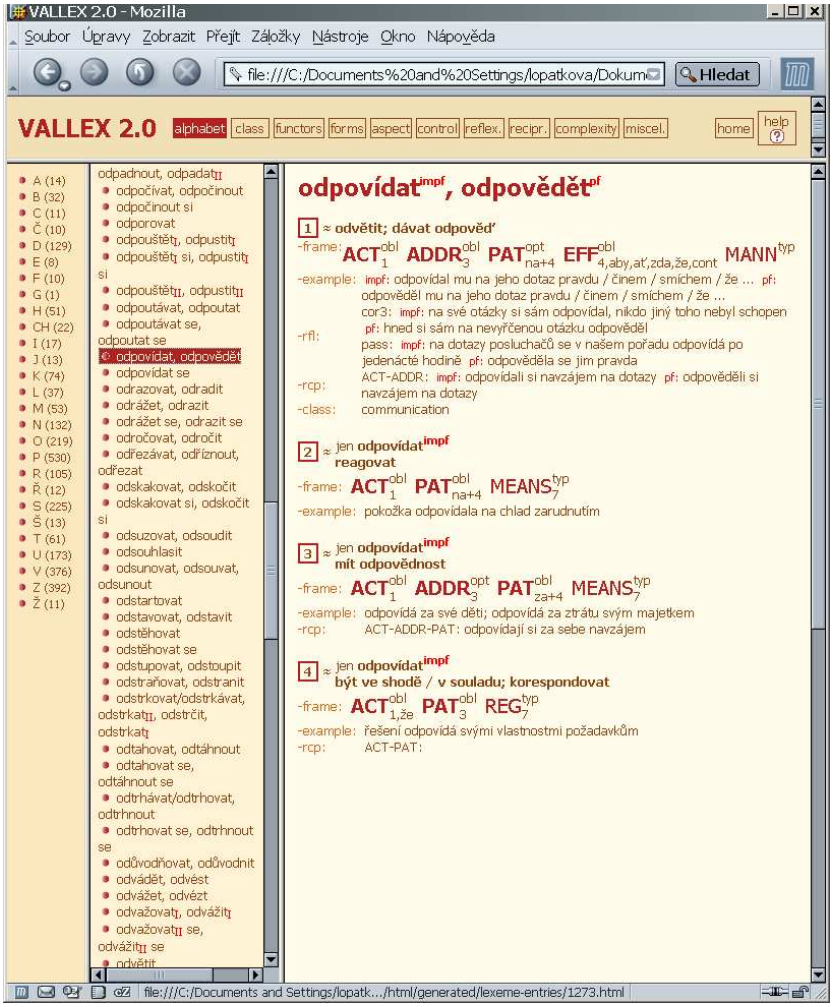

Figure 2. The screenshot of the lexeme odpovídat^impf, odpovědět^pf – to answer/to react/to be responsible/to correspond.

*impf*1 *koncert odpadá; impf2 každou chvíli jím odpadává hodina češtiny; pf*1 *koncert odpadl*

4 jen odpadat *pf*2 *II* (*padáním se oddělit*) ACT(1) ◇*nekvalitní nátěr brzo odpadá*

odpírat / odepírat *impf*, **odepřít** *pf* v (*impf neposkytovat; nečinit něco žádaného; odmítat; odříkat; pf neposkytnout; neučinit něco žádaného; odmítnout; odříci*) ACT (1) ADDR(3) PAT(4|inf|aby) ◇*impf odpíral pomoc / poslušnost; odpíral vojenskou službu; pf odepřel jim pomoc / / poslušnost* ✠ control: ACT, ADDR; rfl: cor3, pass; rcp: ACT-ADDR; class: exchange

odpočinout si *pf* v (*oddechnout si; oddat se klidu*) ACT(1) ?PAT(oł+2) ◇*věděla, že si musí od dětí odpočinout; před cestou si trochu odpočineme* ✠ rcp: ACT-PAT

odpočívat *impf*, **odpočinout** *pf* v

1 jen odpočívat *impf* (*oddychovat si; oddávat se klidu; ležet; spočívat*) ACT(1) ¿LOC ◇*po náročném dni musel chvíli odpočívat; nechte těsto odpočívat; ruce odpočívaly na klíně* ✠ rfl: pass0

2 (*impf poskytovat odpočinek; pf poskytnout odpočinek; pohovět*) ACT(1) PAT(3) ◇*impf odpočívat očím / tělu; pf odpočinout očím / tělu* ✠ rfl: pass0

odporovat *impf* v (*tvrdit opak, protiřečit*) ACT(1) PAT (3) ¿REG(v+6) ◇*pořád tatínkovi odporoval; to tvrzení odporuje zdravému rozumu* ✠ rfl: cor3, pass0; rcp: ACT-PAT

odpouštět *I impf*, **odpustit** *I pf* v

1 (*impf přestávat se zlobit; pf přestat zazlívat*) ACT(1) ADDR(3) PAT(4|že|cont) ◇*impf maminka synovi všechno odpouštěla; pf odpustil mu, že nepřijel* ✠ rfl: cor3, pass; rcp: ACT-ADDR; class: mental action

2 (*impf nevymáhat; pf slevit; prominout*) ACT(1) ADDR (3) PAT(4) ◇*impf odpouštět svým dlužníkům pohledávky; pf odpustil mu dluh* ✠ rfl: pass; rcp: ACT-ADDR; class: exchange

odpouštět *II impf*, **odpustit** *II pf* v (*impf vypouštět část; pf vypustit část*) ACT(1) PAT(4) DIR1 ◇*impf odpouštěli pomalu vodu z rybníku; pf odpustili třetinu vody z nádrže* ✠ rfl: pass; class: location

odpouštět si *I impf*, **odpustit si** *I pf* v idiom (*impf opomíjet; pf opomenout*) ACT (1) PAT (4|aby) ◇*impf neodpouští si poznámky na své blízké; pf nemohl si odpustit tuto poznámku*

odpoutávat *impf*, **odpoutat** *pf* v

1 (*impf odvazovat; uvolňovat; pf odvázat; uvolnit*) ACT(1) PAT(4) DIR1 ◇*impf odpoutávat balon; odpoutávat koně od žlabu; pf odpoutat balon; odpoutat psa ze řetězu* ✠ rfl: cor4, pass; rcp: ACT-PAT

2 (*impf přerušovat spojení/závislost; uvolňovat; pf přerušit spojení/závislost; uvolnit*) ACT(1) PAT(4) ORIG(oł+2) ◇*impf odpoutávat pozornost od vnějšího světa (SSJČ); její přítomnost stále odpoutává jeho pozornost; pf nemohl od ní oči odpoutat* ✠ rfl: pass; rcp: ACT-ORIG

odpoutávat se *impf*, **odpoutat se** *pf* v (*impf uvolňovat se; pf uvolnit se*) ACT(1) PAT(oł+2) ◇*impf odpoutával se od existenčních starostí; pf odpoutal se od svých radostí*

odpovídat *impf*, **odpovědět** *pf* v

1 (*odvětit; dávat odpověď*) ACT(1) ADDR(3) ?PAT(na+4) EFF(4|aby|ať|zda|že|cont) ¿MANN ◇*impf odpovídal mu na jeho dotaz pravdu / činem / smíchem / že ...; pf odpověděl mu na jeho dotaz pravdu / činem / smíchem / že ...* ✠ rfl: cor3, pass; rcp: ACT-ADDR; class: communication

2 jen odpovídat *impf* (*reagovat*) ACT(1) PAT(na+4) ¿MEANS (7) ◇*pokožka odpovídala na chlad zarudnutím*

3 jen odpovídat *impf* (*mít odpovědnost*) ACT(1) ?ADDR (3) PAT(za+4) ¿MEANS(7) ◇*odpovídá za své děti; odpovídá za ztrátu svým majetkem* ✠ rcp: ACT-ADDR-PAT

4 jen odpovídat *impf* (*být ve shodě / v souladu; korespondovat*) ACT(1|že) PAT(3) ¿REG(7) ◇*řešení odpovídá svými vlastnostmi požadavkům* ✠ rcp: ACT-PAT

odpovídat se *impf* v (*být zodpovědný*) ACT(1) ADDR(3) PAT (z+2) ◇*odpovídá se ze ztrát*

odrazovat *impf*, **odradit** *pf* v (*impf odvracet od úmyslu / / zájmu; pf odvrátit od úmyslu / zájmu; odstrašit*) ACT (1|inf|že|cont) ADDR(4) ?PAT(oł+2) ¿MEANS(7) ◇*impf otec ho odrazoval od cesty do Jižní Ameriky; odrazovalo ho, že je celý projekt finančně neprůhledný; pf neúspěch mě nikdy neodradil; odradilo ho stále hrát jen druhé housle* ✠ control: PAT; rfl: pass; rcp: ACT-ADDR; class: psych verb

odrážet *impf*, **odrazit** *pf* v

1 (*impf uvádět do pohybu (jiným směrem); pf uvést do pohybu (jiným směrem)*) ACT(1) PAT(4) ¿MEANS(7) ◇*impf odrážet rukama míče; pf odrazit rukama míč* ✠ rfl: pass; class: contact

2 (*impf potlačovat; zastavovat; pf potlačit; zastavit*) ACT (1) PAT(4) ◇*impf armáda zdatně odrážela útoky; pf armáda odrazila útok* ✠ rfl: pass

3 (*impf odpoutávat se od břehu; pf odpoutat se od břehu*) ACT (1) ↑DIR1 ◇*impf loď odrážela od břehu; pf loď odrazila od břehu* ✠ rfl: pass0; class: motion

4 jen odrážet *impf* (*zobrazovat odrazem*) ACT(1) PAT(4) ◇*mýtus odráží víru*

5 jen odrážet *impf* (*způsobovat optický odraz*) ACT(1) PAT (4) ◇*bílá barva odrážží slunce*

6 idiom jen odrazit *pf* (*nechat zteplat*) ACT(1) PAT(4) ◇ *nechat odrazit mléko* ✠ rfl: pass

odrážet se *impf*, **odrazit se** *pf* v

1 (*impf nárazem dostávat jiný směr; mít odraz; pf nárazem dostat jiný směr; mít odraz*) ACT(1) ↑DIR1 ◇*impf míče se odrážely od stěny; odrážet se od hladiny; pf míč se odrazil od země; odrazit se od hladiny* ✠ class: motion

2 (*impf projevovat se; mít důsledky; pf projevit se; mít důsledky*) ACT(1) PAT(na+6|v+6) ◇*impf zvýšení životní úrovně se odráží i v lepší vybavenosti domácností; pf zvýšení životní úrovně se odrazilo i v lepší vybavenosti domácností*

3 (*mít odraz*) ACT(1) LOC ◇*impf paprsky se odrážely ve vodě; pf obraz se odrazil na hladině* ✠ class: location

odročovat *impf*, **odročit** *pf* v (*impf odkládat; stanovovat na pozdější dobu; pf odložit; stanovit na pozdější dobu*) ACT(1) PAT(4) ¿TFRWH(z+2) ¿TOWH(na+4) ?DIFF(o+4) ◇ *impf nemůžeme stále do nekonečna odročovat jednání; pf soudce odročil jednání o měsíc* ✠ rfl: pass

odřezávat *impf*, **odříznout** *pf*1, **odřezat** *pf*2 v

Figure 3. Sample page from the printed version of the lexicon.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 61–70

Identification of Topic and Focus in Czech

Evaluation of Manual Parallel Annotations

Šárka Zikánová, Miroslav Týnovský, Jiří Havelka

Abstract

This paper presents results of a control annotation of the Topic-Focus Articulation of Czech sentences based on the notion of "aboutness". This is one of the steps testing the hypothesis about the relation between contextual boundness and "aboutness". We suppose that the bipartition of the sentence into its Topic and Focus ("aboutness") can be automatically derived from the values of contextual boundness assigned to each node of the dependency tree representing the underlying structure of the sentence. For the testing of this hypothesis, control manual parallel annotations have been carried out. The principles of the controll annotations are described and preliminary results are reported on.

1. Introduction

The topic-focus articulation of a sentence into its Topic and Focus can be looked upon from two points of view: as derived from a primary notion of contextual boundness or in terms of *"aboutness"* (Focus is "about" Topic, i.e. F(T); with a primary reading of a negative sentence non-F(T); see Mathesius, 1947, p. 235; Sgall, Hajičová, and Panevová, 1986; Firbas, 1992). According to *contextual boundness,* elements in sentences are classified as contextually bound (CB; with a subtype of contrastive contextually bound elements, CCB) or contextually non-bound (CN; Sgall, Hajičová and, Buráňová, 1980; Sgall, Hajičová, and Panevová, 1986), cf. the following example:

(1) *(Maruška se obrátila na lesní víly.) Víly*$_{CB}$ *ji*$_{CB}$ *vyslechly* $_{CN}$.
[lit.: Mary turned to forest fairies.] The_fairies$_{CB}$ received$_{CN}$ her$_{CB}$.

The bipartition of the sentence into Topic and Focus is then derived from the CB/CCB/NB features; for our example, the Topic is *Víly*$_{CB}$ *ji*$_{CB}$ [(The) fairies$_{CB}$ her$_{CB}$] and the Focus is *vyslechly*$_{CN}$ [received$_{CN}$].

The criterion of the "aboutness" divides a sentence into two parts: Topic (T, a part expressing what the sentence is about) and Focus (F, a part giving information about the Topic). Thus, if (2)

Please cite this article as: Šárka Zikánová, Miroslav Týnovský, Jiří Havelka, Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. The Prague Bulletin of Mathematical Linguistics No. 87, 2007, 61–70.

is used in the context of the question "Where are the children? / What are the children doing?", the following assignment of Topic and Focus would hold: the Topic of the sentence is *Děti* [The children] and the Focus is *běhají po ulici* [are running in the street].

(2) *Děti* $_{T}$ *běhají* $_{F}$ *po ulici* $_{F}$.
[lit.: The_children $_{T}$ are running $_{F}$ in the_street $_{F}$.]

It should be noted (cf. Sgall, Hajičová and, Panevová, 1986), that although in principle the CB items belong to the Topic of the sentence and the NB items to the Focus, this is not so when deeply embedded sentence elements are taken into account. See the element *vašeho* [your] in (3) which belongs to the Focus, though it is contextually bound. (The context of the sentence can be "What did you do yesterday?".)

(3) *Včera*$_{CB,\,T}$ *jsem potkal*$_{CN,\,F}$ ***vašeho***$_{\mathbf{CB,\,F}}$ *kolegu*$_{CN,\,F}$.
[lit.: Yesterday$_{CB,\,T}$ I_met$_{CN,\,F}$ **your**$_{\mathbf{CB,\,F}}$ colleague$_{CN,\,F}$.]

2. The framework of the project: from contextual boundness to aboutness

In our project the relations between contextual boundness and aboutness are investigated. According to the underlying hypothesis, the values of aboutness (Topic and Focus) can be derived from the values of contextual boundness Sgall, Hajičová and Panevová (1986). We test this hypothesis on the material from the Prague Dependency Treebank 2.0 (PDT), where approximately 50,000 Czech sentences have been annotated on three levels, one of them being the underlying syntactic level (tectogrammatics). On that level, sentences are represented in a form of dependency trees, in which the nodes represent autosemantic elements of the sentence and the edges represent the types of relations between the governing and the dependent nodes. Every node has been assigned (in addition to other relevant values) one of the values of contextual boundness.

Our study proceeds in the following three steps:

- the formulation of an algoritm transforming the values of contextual boundness into the values of aboutness; implementation of the algoritm on the data from the PDT (Sgall and, Hajičová, 2005; Hajičová, Havelka and, Veselá, 2005);
- manual parallel annotation of the control data according to the aboutness relation (i.e. directly assigning the bipartition of Topic and Focus);
- comparison of the values achieved in the manual annotation with the automatically assigned T-F bipartition and evaluation of the results.

In the present paper, we are concerned with the second point of the overall programme of the project – we describe and evaluate the results of the manual parallel annotations which will later serve as referential data for the evaluation of the automatic recognition of Topic and Focus.

3. The linguistic material

For the control annotation, the texts from the PDT have been used, so that we get data comparable with the results of the automatic procedure. The texts in the PDT come from Czech

newspapers from the beginning of the 1990's; they extend from short remarks to longer essays. The annotators worked with whole texts, since for the correct analysis of the topic-focus articulation, it is necessary to respect the context.

In total, almost 11,000 sentences have been analysed (cf. Tables 1–2). All the annotations have been done in parallel, in order to take into account possible disagreement of the annotators in their interpretation of the topic-focus articulation as well as to be aware of errors by individual annotators. The main part of sentences (almost 10,000 sentences) has been analysed in three parallel annotations; a smaller sample of almost 900 sentences has been annotated in six parallel versions.

4. The method of the annotation

In order not to "spoil" the control data with the hypothesis to be verified, we worked with ten annotators who were not familiar with the previous annotation of the topic-focus articulation in the PDT. If we wanted to get the picture of the common perception of the topic-focus articulation by native speakers, we could not influence annotators with too strict instructions which could be contradictory to their natural intuition. Therefore basic principles of the annotation have been outlined only; later some problematic parts have been discussed in detail (e.g. the analysis of questions, sentences consisting just of one word or sentences with direct speech; cf. Zikánová, 2006).

The annotators worked with a linear (surface shape) form of the texts. There were four possible values, which could be ascribed to individual words in texts:

T	part of the Topic	(what the sentence is about)
F	part of the Focus	(new information about the Topic)
B	Boundary	(a marker of the boundary dividing two structures in which Topics and Focuses should be identified separately, e.g. a conjuction or a punctuation mark within a sentence)
N	Not clear	(problematic words where the annotator is not sure)

The elementary instructions for the annotation included the following points:

- Analyse the structure of the main clause only. Dependent clauses are to be treated as integral elements of the main clause. (Therefore the borderline between the Topic and Focus should not be marked within a dependent clause.)
- Describe the appurtenance of every single word or unit in the main clause to Topic or Focus. It is possible that there is more than one border between these two parts of the sentence, both of these parts can be interrupted with other elements.
- In coordinated clauses, analyse the structure of each main clause separately. (In complex sentences with subordinated clauses, analyse the main clause only.)
- Describe the nominal group as an integrated element (with a preposition, pronoun, adjective or another noun, as the case may be).
- It is possible to assign all the elements of a sentence as belonging to the Focus. (It is not necessary that the sentence contains Topic.)

Generally, when choosing which elements in the linear surface shape of a sentence might have been in the analysis omitted, we have been guided by the principles of the automatic annotation of underlying dependency trees in the PDT as we want to compare these sets of data. The following example presents the way of analyzing sentences in control annotations:

(4) (In the previous context, poor conditions in different world trading zones have been mentioned in general.)
V Indonésii je minimální denní mzda jeden a půl dolaru a někdy za to musí dělníci pracovat 10–12 hod.
[lit.: In Indonesia is minimal daily pay one and half dolar and sometimes for it have workers to_work 10–12 hours.]
[In Indonesia, the minimal daily pay is one and half dolar and sometimes workers have to work for 10–12 hours for it.]

1. There are two coordinated clauses in the sentence; they are to be analysed separately, the conjuction *a* is assigned the value B (Boundary).

2. When setting Topic and Focus, we formulate first a question about the presupposed Topic of the sentence. As for the first clause, we can ask the following questions: *What can we say about Indonesia? What can we say about the minimal daily pay in Indonesia? What can we say about one and half dolar?*

3. Then the analysed sentence is tested as an answer to the formulated question:

(4a) *What can we say about Indonesia? – V Indonésii je minimální denní mzda jeden a půl dolaru.*
In Indonesia, the minimal daily pay is one and half dolar.

(4b) *What can we say about the minimal daily pay in Indonesia? – V Indonésii je minimální denní mzda jeden a půl dolaru.*

(4c) *What can we say about one and half dolar? – V Indonésii je minimální denní mzda jeden a půl dolaru.*

If the answer naturally matches with the question (with respect to the previous context), then the elements repeated from the question are assigned the value T (Topic) and the elements of the part that is the proper answer are assigned the value F (Focus). If the answer does not correspond to the question, the choice of the presupposed Topic is not correct.

In our case, questions (a) and (b) can be answered with the analysed sentence, whereas the question (c) does not correspond to it. Thus, the Topic-Focus values will be assigned in the following way:

(4a') *V Indonésii* $_T$ *je* $_F$ *minimální denní mzda* $_F$ *jeden a půl dolaru* $_F$.

(4b') *V Indonésii* $_T$ *je* $_F$ *minimální denní mzda* $_T$ *jeden a půl dolaru* $_F$.

Since there is a (restricted) variability in matching questions, a certain variability in answers and analyses is admissible, too (4a'–b').

The second clause of the compound sentence is analysed according to the same instructions. The appropriate question to which this clause can be an answer is *What can we say about this daily pay?*

(4d) *Někdy za to musí dělníci pracovat 10–12 hod.*
[lit.: Sometimes workers have to_work for 10–12 hours for it.]

Někdy $_F$ *za to* $_T$ *musí* $_F$ *dělníci* $_F$ *pracovat* $_F$ *10–12 hod.* $_F$.

5. Results and discussion

When evaluating the parallel annotations, we have restricted our attention to certain types of the phenomena observed. With the following elements the assigned value of aboutness has not been taken into account:

- all the words of subordinated clauses except for the verb governing the subordinated (dependent) clause,
- all auxiliary words, which have no corresponding node on the tectogrammatical level of the PDT (functional words such as verbal morphemes, prepositions),
- punctuation marks.

Examining the results, we work with the T/F values of aboutness which have been described above in Sect. 4 and with an additional value "U" – "unannotated" for very sporadic occurrences of words overlooked by mistake by the annotators.

Tables 1 and 2 show the level of agreement among three and six parallel annotations, respectively.

Table 1. Agreement among three parallel annotations

	Occurrence	**Percentage**
Number of sentences	9,825	100.00
Agreement in the annotation of whole sentence	3,553	36.16
Number of words	79,419	100.00
Agreement in the annotation of individual words	60,137	75.72

Table 2. Agreement among six parallel annotations

	Occurrence	**Percentage**
Number of sentences	879	100.00
Agreement in the annotation of whole sentence	232	26.39
Number of words	6,232	100.00
Agreement in the annotation of individual words	4,212	67.59

In Tables 3 and 4, the level of agreement in annotation of individual words is presented in a more detailed way.

Explanations: T and F in the three-letter and six-letter labels refer to the assignment of a word to Topic, or to Focus, respectively, so that e.g. TTT means that a word was considered to be a part of Topic with all the three annotators, or TTT TFF means that a word was considered

to be a part of Topic by four annotators and as belonging to the Focus by two annotators.

Table 3. Types of annotations of individual words in three parallel analyses

	Occurrence	Percentage
FFF	46,099	58.05
TTT	14,036	17.67
TFF	10,575	13.32
TTF	8,287	10.43
TFN	139	0.18
FFN	139	0.18
TTN	67	0.08
Others	77	0.10
Total	**79,419**	**100.00**

Table 4. Types of annotations of individual words in six parallel analyses

	Occurrence	Percentage
FFF FFF	3,332	53.47
TTT TTT	880	14.12
TFF FFF	635	10.19
TTT TTF	367	5.89
TTT FFF	335	5.38
TTF FFF	332	5.33
TTT TFF	288	4.62
FFF FFN	23	0.37
Others	40	0.64
Total	**6,232**	**100.00**

The results of the three parallel annotations in Table 3 as well as of the six parallel annotations in Table 4 indicate that the highest percentage of agreement has been achieved with words belonging to the Focus. The agreement as for the appurtenance of a word to the Topic is not that frequent, nevertheless it is in both annotations the second most common case of agreement. In both annotations, the first two positions in the Tables are occupied by cases of absolute agreement; altogether there are 75.72 % of the absolute agreement in the annotation of individual words at three parallel annotations (cf. Table 1) and 67.59 % of the absolute agreement at six parallel annotations (cf. Table 2). In Table 3, which presents the results of three parallel annotations, the disagreement of annotators is almost the same if the assignment is T or F (lines 3 and 4); with six parallel annotations there is an apparent preference of the annotators to assign F (line 3) rather than T (line 4); actually, this is in accordance with our comments above on lines 1 and 2.

It is interesting to notice that the annotators did not acknowledge much doubt in the assignment of values, although the instructions they received allowed to do so and the reading of some words is not unambiguous: they get much more often in an open disagreement with each other than using the value N (not clear).

The following examples present some results of the parallel annotations. In sentence (5), all the three annotators fully agree in their analysis:

(5) (There is no previous context, the text starts with this sentence.)
Jihlavská radnice *hodlá rázně řešit problém neplatičů nájemného.*
[lit.: **Jihlavian town_council** wants peremptorily to_solve problem of_ bad_payers of_ hire_costs.]

[The town council of Jihlava is about to solve their problem with bad payers of hire costs peremptorily.]

(The parts of Topic are marked with bold characters, the other parts belong to Focus.)

Another example of the full agreement is presented under (6), where all the six annotators analyze the sentence in the same way:

(6) (In the previous context, Edvard Beneš was mentioned as a theme of a recent TV- discussion.)

Edvard Beneš *byl tématem natolik kontroverzním, že přivedl do varu i nejserióznější historiky.*

[lit.: **Edvard Beneš** was theme in_so_far controversial, that he_upset even the_most_ respectable historians.]

[Edvard Beneš was such a controversial theme, that he upset even the most respectable historians.]

In some cases, there are more options in the choice of the test question, and subsequently the solutions differ with individual annotators. The sentence (7) presents an extreme example of disagreement among three annotators, where all the interpretations respect the basic guidelines of the annotation.

(7a) (There is no previous context, the text starts with this sentence.)

Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru *vynesly bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu, titul Fotbalista měsíce dubna v anketě týdeníku Kicker.*

(This analysis corresponds with the question: *What can we say about the following qualities of a football player?)*

[lit.: **Seven goals within nine matches, immense game dynamism and desire to win, as well as readiness quickly to_clear language barrier** have_brought to_the_former player _of_Slavia Pavel Kuka, now acting in German Kaiserlautern, title Footballer of_the_Month April in inquiry of_the_weekly_magazine Kicker.]

[Seven goals within nine matches, the immense dynamism in game and the desire to win, as well as the readiness to clear the language barrier quickly have brought the title The Footballer of the Month April in the inquiry of the weekly magazine Kicker to the former player of Slavia Club Pavel Kuka (now acting in German Kaiserlautern).]

(7b) *Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru vynesly* ***bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu,*** *titul Fotbalista měsíce dubna v anketě týdeníku Kicker.*

(Question: What can we say about the player Pavel Kuka?)

(7c) *Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru vynesly bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu,* ***titul Fotbalista měsíce dubna v anketě týdeníku Kicker.***

(Question: What can we say about the title The Footballer of the Month April?)

The control subcorpus manually annotated in the way described in our paper is now being compared with the output of the automatic assignment of Topic and Focus to the same subcorpus of texts (annotated on the tectogrammatical level of the Prague Dependency Treebank). The automatic procedure is based on the hypothesis that the bipartition of the sentence into its Topic and Focus can be derived from the values of contextual boundness, while the manual annotation reflects directly the bipartition.

As the results show, the variability of manual solutions must be taken into account in further steps. We should be aware that while we get only a single, unambiguous result from the automatic annotation, more ways of interpretation could be correct. This will be of a great importance in the phase of the comparison between the automatic and manual annotation and its evaluation – it must be reasonably determinated which type of an agreement between automatic and manual annotation is significant and which is not. Also, to achieve a deeper insight into the issue of the position of the boundary between Topic and Focus, it is necessary for the analysis of the cases of disagreement between annotators to take into account the appurtenance of the relevant items to different word classes and the structure of sentences generally (cf. the ambiguous position of nominal groups with rhematizers, of the predicate verb or adverbials in some Czech sentences; see Zikánová, 2006). The evaluation of the results of this comparison will be a useful test of the hypothesis and will enrich our knowledge of the information structure.

Acknowledgements

The research reported in this contribution has been carried out under the grant project of the Ministry of Education, Youth and Sports (Czech Republic) MSM-0021620838 "Modern methods, systems and structures of informatics".

References

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication.* Cambridge University Press, Cambridge.

Hajič, Jan et al. 2006. *Prague Dependency Treebank 2.0.* Linguistic Data Consortium, Philadelphia.

Hajičová, Eva and Petr Sgall. 2004. Degrees of Contrast and the Topic-Focus Articulation. In: Steube, Anita (Ed.), *Information Structure: Theoretical and Empirical Aspects.* de Gruyter, Berlin – New York, pp. 1–13.

Hajičová, Eva and Petr Sgall. *Corpus Annotation As a Test of a Linguistic Theory.* In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, ELRA, Paris, pp. 879–884.

Hajičová, Eva, Jiří Havelka and Kateřina Veselá. 2005. Corpus Evidence of Contextual Boundness and Focus. In: *Proceedings from The Corpus Linguistics Conference Series,* vol. 1, no. 1, 9 pp. Birmingham, ISSN 1747-9398, 2007.04.02 under http://www.corpus.bham.ac.uk/PCLC/birmingham-tex-def-def.doc

Mathesius, Vilém. 1947. O tak zvaném aktuálním členění větném. In Mathesius, Vilém. *Čeština a obecný jazykozpyt.* Melantrich, Prague, pp. 234–242. [About so called functional sentence perspective.]

Mathesius, Vilém. 1982. Aktuální členění větné a sloh. In Mathesius, Vilém. Řeč a sloh, quotation according to Mathesius, Vilém; Macek, Emanuel and Josef Vachek (Eds.): *Jazyk, kultura a slovesnost.* Odeon, Prague, pp. 124–128. [Functional sentence perspective and text composition.]

Sgall, Petr and Eva Hajičová. 2005. The Position of Information Structure in the Core of Language. In: Carlson, Gregory N. and Francis Jeffry Pelletier (Eds.): *Referency and Quantification: The Partee Effect.* CSLI Publications, Stanford (California), pp. 289–302.

Sgall, Petr, Eva Hajičová and Eva Buráňová. 1980. *Aktuální členění věty v češtině.* Academia, Prague. *[Topic-Focus Articulation of Czech Sentences.]*

Sgall, Petr, Eva Hajičová and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspect.* Reidel Publishing Company, Dordrecht and Academia, Prague.

Sgall, Petr. 2002. Moravská a pražská (malostranská) koncepce aktuálního členění. In: Hladká, Zdeňka and Petr Karlík (Eds.), *Čeština – univerzália a specifika, 4.* Nakladatelství Lidové noviny, Prague, pp. 51–58. [Moravian and Praguian (Lesser Town) conception of Topic-Focus Articulation.]

Zikánová, Šárka. 2006. Problematické syntaktické struktury: k rozborům aktuálního členění v Pražském závislostním korpusu. In: *Proceedings of the conference VII. mezinárodní setkání mladých lingvistů,* 15.–17. 5. 2006. Olomouc. [In print since 2006]. [Problematic syntactic structures: to the studies of the topic-focus articulation in the Prague Dependency Treebank.]

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 71–86

A Note on the Prague School

Jun Qian

Abstract

The 80th anniversary of the Prague Linguistic Circle offers an occasion to think about how to document the Prague School related events, how to keep whatever related to the Prague School, and how to make the Prague School resources easily accessible. In the following I will first chronologically list some Prague School related events in the past ten years (1996–2006). Then I will refer to several personal communications as related to certain aspects of the Prague School theory. Finally I will propose when faced with this age of globalization and digitization what can be done so as to maximally utilize the Prague School resources.

The following chronological list of Prague School related events are highly selective. Under the heading of the year are listed the events that occurred in that year.

1996

(1) From March 28 to 30, 1996, an international conference was held in Prague to celebrate the 70th anniversary of the Prague Linguistic Circle and to commemorate the centenary of the birth of Roman Jakobson. Some of the papers presented at this conference are included in *Prague Linguistic Circle Papers Volume 3* (1999).[1]

(2) Professor Josef Vachek (1909–1996) passed away on March 31. He was probably the last of the pre-war Prague School members. The international linguistic community's knowledge of the Prague School is largely due to his persistent effort (e.g. Vachek 1960, 1964a-b, 1966, 1968, 1983; Mathesius 1975). These efforts should be viewed in relation to the long-term unfavorable or hostile climate against the Prague School, in relation to the post-war behavior of some of the pre-war Prague School members such as Jan Mukařovský (1891–1975) and František Trávníček (1888–1961; cf. Toman 1995: Chapter 12; Firbas 1997), and in relation to the fact that further volumes of *Travaux du Cercle Linguistique de Prague* (*TCLP*, 1929–1939, 8 volumes) and *Travaux Linguistiques de Prague* (*TLP*, 1964–1971, 4 volumes) "were strangled by political authorities" (Vachek, foreword to *Prague Linguistic Circle Papers Volume 1*).

Please cite this article as: Jun Qian, A Note on the Prague School. The Prague Bulletin of Mathematical Linguistics No. 87, 2007, 71–86.

(3) *Prague Linguistic Circle Papers Volume 2* was published. The first volume was published in 1995. Both volumes were reviewed by Qian (1997).

1997

Professor Oldřich Leška (1927–1997) passed away on August 9. He succeeded Miloš Dokulil as chair of the Circle in 1996 (Eva Hajičová was the chair between 1997 and 2006). Leška was co-editor of the first three volumes of *Prague Linguistic Circle Papers* (*PLCP*; *PLCP* 1, 3, and 4 include his papers).

1999

Prague Linguistic Circle Papers Volume 3 was published. It was reviewed by Qian (2000) and Salzmann (2001).

2000

Professor Jan Firbas (1921–2000) passed away on May 5. He was best known for his work on functional sentence perspective (FSP, cf. Firbas 1992; Chamonikolasová 2001; Qian 2001a; Svoboda 2003). Of the post-war Prague School's work on FSP, Firbas is noted for his concept of communicative dynamism (CD), František Daneš (b.1919) for his concept of thematic progression (TP, Daneš 1974), and Petr Sgall (b.1926) for his study of topic-focus articulation (TFA, cf. Sgall 2006:227–301). Their work is representative of the post-war Prague School approach to syntax.

A collection, which was originally intended to celebrate Firbas's 80th birthday, came as a commemorative volume in 2003, i.e. *Language and Function: To the memory of Jan Firbas* (ed. by Josef Hladký, preface by Eva Hajičová and Petr Sgall). The book was reviewed by Kirtchuk-Halevi (2003) and Salzmann (2005).

2002

Prague Linguistic Circle Papers Volume 4 was published. It was reviewed by Qian (2002a), Webb (2002), and Salzmann (2004).

2003

(1) Josef Vachek's (1960) *Dictionnaire de linguistique de l'École de Prague* (avec collaboration de Josef Dubský) was translated into English, entitled *Dictionary of the Prague School of Linguistics* (edited by Libuše Dušková). It was reviewed by Qian (2004), Verleyen (2004), and Holes (2005).

(2) Stephen Rudy, professor of Russian and Slavic languages at New York University "died of head injuries after an accidental fall at home on Aug. 11." (OBITUARY, http://www.thevillager.

.com/villager_19/stephenrudy.html) He was only 54. Rudy did a lot of work to preserve Jakobson's linguistic legacy (Rudy 1990; Jakobson 1985, 1987, 1988; Waugh and Rudy 1991). His untimely death put an end to his plan to publish all that are not included in Jakobson's eight-volume *Selected Writings* (1962–1988) as Volumes 9 and 10. Since Volume 8 is Completion Volume I, Volumes 9 and 10 would be Completion Volumes II and III.

2006

(1) *Language in its Multifarious Aspects* (556pp.), Petr Sgall's collection of twenty-six papers, edited by Eva Hajičová and Jarmila Panevová, was published. It came on the occasion of Sgall's 80th birthday. The articles are selected from among his 1956–2003 publications and comprehensively reflect his views on and research achievements in various linguistic fields.

(2) Professor Patrick Sériot and Margarita Schönenberger from Université de Lausanne in Switzerland translated *Trubetzkoy's Letters and Notes* (ed. by Jakobson, 1975) from Russian into French (573pp). For information on Sériot and his colleagues' work one can visit their website http://www2.unil.ch/slav/ling.

3. The above description focuses on the scene in Europe. In this section I refer briefly to my work (1998, 2001b) with a focus on some personal communications (e-mails) produced during the period when I was writing a Chinese introduction to *Praguiana: 1945–1990.* These communications are replies to my inquiries about various aspects of the Prague School theory and are of value from the perspective of linguistic historiography (see Toman 1994 and Newmeyer 2001 for the use of personal communications).

In 1998 *Structural-Functional Linguistics: The Prague School* (in Chinese, 70+427pp.) was published. The monograph begins with three introductions by Petr Sgall, Catherine Chvany, and Edward Stankiewicz respectively.

In 2001 *A Roman Jakobson Anthology* (XLII+373 pp.) was published. It consists of 23 papers. Except *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates* (1952), which was translated by the late Professor Wang Li (1900–1986) and published in a Chinese journal, the rest 22 papers were translated and annotated by me and probably for the first time became available in Chinese.

In 2004 *Praguiana: 1945–1990* (ed. by Luelsdorff, Philip A., Jarmila Panevová, and Petr Sgall) was reprinted by Peking University Press, together with a 42-page long Chinese introduction of mine. Unfortunately, the editor changed my title from *The Prague School in its Post-Classical Period* to *Introduction*, and she deleted my footnote of acknowledgement, which runs as follows:

"This research was supported by a Peking University grant. The author is grateful to Professors Catherine Chvany, Edward Stankiewicz, Jarmila Panevová, and Petr Sgall for their help."

As is known, Catherine Chvany (b.1927) and Edward Stankiewicz (b.1920) were Roman Jakobson's students at Harvard, both being distinguished Slavists and versed in the Prague School theory (e.g. Chvany 1996; Stankiewicz 1976, 1977, 1983, 1987, 1991, 1999), while Jarmila Panevová and Petr Sgall are eminent present-day Prague School linguists (e.g. Sgall et al. 1986). *Praguiana: 1945–1990* covers a variety of diversified subjects and it was largely

through their unfailing help that the challenging task of writing an introduction in Chinese was accomplished. Perhaps I should have specified their help by quoting their e-mails to me in my footnotes. To illustrate their help, some personal communications are quoted as follows.

The following three e-mails are Stankiewicz's answers to my inquiries about Jakobosn's binarism, the term 'morpheme', and Leška's concept of transposition:

[1] "You know very well that I have criticized Jakobson for his exaggerated and misleading binarism. Specifically in my Prague paper of 1999 I point out that even in mathematics that speak of symmetry, asymmetry and dissymmetry, i.e., my reference to complementarity and "neutral terms" (e. g. the third person which refers to a third person or absence of a person, as in impersonal verbal constructions). Jakobson pushed the concept of economy too hard, and thus he misinterpreted certain relations (not only in morphology but also in phonology and in syntax). The relations of complementarity I illustrated on multiple examples." (Edward Stankiewicz, personal communication, January 30, 2004)

[2] "Now, as for your questions. The terms phoneme and morpheme have both been coined by Baudouin de Courtenay. He also defined the phoneme as a "bundle" (this was Jak.'s English translation of the term) of distinctive features. By definition both terms define the ultimate units of the two levels of language. Properties can always be added or reduced giving rise to new phonemes or morphemes. American descriptivists (as well as Kuryłowicz) have defined the phoneme and morpheme as sums of positional variants (allophones and allomorphs), an approach which denied the abstract character of the linguistic units, and created altogether a mess. (e.g., the past t. of *sleep*: *slept* was analyzed as past t. morpheme *-t* and the allomorph *slep-* of *sleep*). That's why they never developed the theory of morphophonemics. (The fault was largely Baudouin's). Why don't you come for a while to Yale to think over and discuss all these problems? Warmest greetings, Edward." (Edward Stankiewicz, personal communication, February 1, 2004)

[3] "Within the next few days I hope to send you the obituary I wrote about Oldrich Leška, who had spent some time in my dept. at the University of Chicago. I had a whole team working with me on the structural description of the Russian dialects. In 1970 (a year before I moved to Yale) I submitted 3 volumes of our results to The Office of Education which had sponsored the project. I never published them in a book form because certain problems were left unresolved (especially in morphology). I am aware of Leška's theory of transposition (which was in part influenced by Karcevskij) and I refer to it in my obituary, but I do not like it (though I do not criticize it in my obituary). The drastic transformation of functions referred to in Leška's "transposition" undercuts the theory of the invariant in which I follow both K. Bühler and Jakobson. The Russian forms *poshli*! *pojexali*! are indeed forms of the past tense. When used as commands they are still forms of the past that present an expected action as completed; hence they are used only in the perfective (we don't say *exali*! *xodili*!) and the expressions carry the connotation of a command only in a given colloquial and semi-metaphorical context. For our friend Leška the imperative *znaj*! ("know") is a mere variant of the unreal modal *znaj ja* "had I but known", a view that denies the strict correlation between form and meaning (ignoring, of course, the existence of homonyms). But *znaj ja* changes the form (the pronoun follows the verb) and the form is no longer an imperative, but a modal (historically there might have

indeed been a semantic connection since the imperative, like the modal, expresses the expectation of an event). But the attempt to reduce these different forms and meanings to the status of variants plays havoc with our basic understanding of language as a system of signs endowed with invariant and distinctive meanings. I hope I have answered your question to your satisfaction. With warmest greetings. Edward." (Edward Stankiewicz, personal communication, February 11, 2004)

The following four e-mails are Chvany's replies to my inquiries about Jakobosn's (1936) idea of the general meaning of a case, the concept of opposition, the possible division of morphemes, and the invariant meaning:

[1] "Hi Jun, Happy New Year (both European/American style and Chinese Year of the Monkey). You have my Selected Essays, don't you? Ch. 13 is on Jakobson's cube I think I also sent you my article from *Case in Slavic* (Brecht & Levine eds 1986). I'm not sure Jakobson's 1936 idea was fully accepted but it was certainly required reading among American Slavists and their graduate students, as was the 1958 version. The cube was very popular but Jakobson's inclusion of G2 and L2 and making them LESS marked than the much less restricted G1 and L1 was not widely accepted, i.e., the 1936 scheme was preferred (see in my References, one to an article by D. S.Worth of the 1980s: something like "G2 L2 revisited" – one might say it was accepted esp. by those who didn't actually work on case. At the same time, Kuryłowicz's distinction between syntactic and adverbial cases was an important rival, and I think corresponds better to current work in syntax. There was also an important critique by Timberlake in IJSLP 1987 (of Waugh & Halle eds 1984, Roman Jakobson, Russian and Slavic Grammar) where several problems with Jakobson's analysis are discussed.

Even now when it is no longer embedded in a Slavistic canon, any Slavist would be expected to know about Jakobson's features, and might have to defend views that disagree with one or the other Jakobsonian version (I found that some Slavists were quite careless in their reading and didn't really know the difference between the two versions – and Jakobson certainly did not provide helpful footnotes that said "this version of 1958 differs from my theory of 1936 in such-and-such ways". That's why I liken his work to an "objet d'art" and the l984 collection to a retrospective show by an artist, who might show different/successive versions of the same motif, without annotations. (I think I also sent you my 1987 review article in RLJ "Two Jakobson retrospectives and a research agenda", right?)

Obviously, the accusative does NOT have a "general meaning" (Gesamtbedeutung) of directionality, that is perhaps the "Hauptbedeutung" – the meaning it has in directional phrases opposed to locative phrases, and also in the cardinal transitive sentences with active agent and affected object, but certainly not in stative sentences (*dejstvie imeet mesto v Sochi* 'the action takes place in Sochi') – also RJ's features oppose A to D, not A to L. There is also an important article by Knorina (also among my References) on the functional load of the cases, showing that some pairs almost never contrast (if I recall correctly, G and L which RJ has differing only by "marginality". (Catherine Chvany, personal communication, January 23, 2004)

In one of my e-mails in January 2004, I asked Chvany about the concept of opposition: "it seems that Jakobson reduced ALL oppositions to binary and privative. If so, I can understand why binary (a logical operation, as Jakobson believed), but am uncertain about privative op-

position (marked/unmarked). If confined to just a single language, say, Russian or English, could ALL morphological oppositions be reduced to privative? Furthermore, does Jakobson's concept of morphological correlations (semantically defined) have a general or even universal validity?" Chvany answered:

[2] "Dear Jun,

Yes, Jakobson was a "reductionist". But one problem that arises is WHAT is included in "ALL OPPOSITIONS" – it may be fairly clear that bound morphemes (inflectional affixes) opposed to their absence (Signe zero: *l'opposition de quelque chose avec rien*) which trivializes 'binary' i.e. as I have written in several places, it is the same as calling the meanings discrete, i.e. as sets of one, opposed to the null set which, by convention, is a member of every set. And in that sense, that grammatical meanings are decomposable into discrete features in [sic]
But oppositions among 2 overt morphemes are not privative but equipollent (e.g. *the/a* are actually *the/0* and *a/0*). Once one leaves the bound morphology and gets into analytic forms, like passives, or causatives, there is no "opposition", in the sense that an active form does not signal "non-passive" the way "present" signals "non-past". If you still have my Peirce Seminar Paper of 1999, I detail these matters there. Also on how oppositions work, see Ch 15 of my SEofCVC book, and in more details in l988 American Contributions to the International Congress of Slavists – on how to account for the multiple meanings of Bulgarian forms in spite of relative poverty of morphology – some meanings (marked) are stable, denoted by certain morphs, while other meanings – the "unmarked ones" which mean the opposite by virtue of opposition with the marked one – are less stable, may be removed by context.
One thing that is quite amazing for intellectual historians is how Jakobson managed to have even very short little papers and casual remarks take on such authority!" (Catherine Chvany, private correspondence, January 29, 2004)

Shortly afterwards, I raised the question of morpheme: "Dear Catherine, thanks a lot for your instruction on the notion of opposition. Another question. You know the Prague School used to think that phoneme could not be further divided. Then they changed this position and defined phoneme as a bundle of "distinctive features" (simultaneous co-existence of properties). By analogy, morpheme can be likewise further divided. If yes, what is the term for those smaller units (components)? What might be such an example in English?" And Chvany answered as follows:

[3] "I think there are two possibilities, for instance in Russian oblique plural cases D *-am*, I *-ami* one could call the *-am* a "submorpheme" signifying obliqueness or marginality (but then what about L *-ax*, maybe it would be *-a* plus the low-tonality feature that is shared by all 3?). But that has not been a very productive approach (English has also some not-quite-morphemes called synesthemes, like initial GL and FL and SL, where GL is used for groups of words having something to do with light and also with stickiness (*gleam, glow, glint, glisten* ... and *glue*...). About these and many more see thesis by Margaret Magnus in Trondheim Norway (she is American, however, and has a web site)

Most of Jakobson's work on morphology involves distinctive SEMANTIC features, so that

he claims that "the ending of the instrumental case" (having different forms in various declensions, singular and plural) has the features +marginal and - for the 2 others, or the form of the past tense has the features +finite, +past, also whatever aspect features it may have. The agreement features of a verb are "portmanteau" morphemes, carrying (redundant) information about person, gender and/or number (copied from those of the subject, if any), even if no segment or phonetic feature of the morpheme can be identified with one of those meanings: that is, in *pisala* 'I /you/she wrote' the *-a* stands for both singular and feminine, and for any of the 3 persons if the subject is a woman.
Hope this is what you were asking about.
Best,
Catherine" (Catherine Chvany, private correspondence, February 1, 2004)[2]

Another question I asked Chvany is on invariant meaning: "In relation to his idea on functional transposition (e.g. a past form expresses an action in the future, as in *nu, mne pora, ja poshel.*), how could the invariant meaning of the past form or rather of any sign be determined? Is the so-called invariant meaning equal to general meaning?" And Chvany replies:

[4] "Dear Jun,

This is a problem linguists create for themselves: they posit an invariant meaning and then any exceptions force them to modify the theory, when the problem is they have incorrectly translated the "invariant meaning". For instance, the past tense affix (-L- in Russian) does not invariantly mean "past time" for it appears not only in "future" uses like *ja poshel* 'I'm off' 'I'm out of here!' but in hypothetical or contrary-to-fact conditions (usually the latter with the extra modal marker *by*). The invariant meaning is "distance in time or reality" from the speaker's position (the latter word is not too good, but I can't do better at the moment). In Jakobsonian terms "past time/tense" is the Hauptbedeutung of the past tense morpheme (the most common meaning), not the invariant general meaning (Gesamtbedeutung).

Other problems linguists create involve trying to translate a morpheme that represents several semantic features with one word when they should decompose it, e.g., a "tense" morpheme may be decomposed into "+deixis" +/-distancing +/-proximate (or +/-distal/remote), perhaps also have some aspect feature (like the French imperfect)...

In Bulgarian the -X- morpheme of the aorist supposedly the "witnessed past" is also used as a "future" for imminent disasters, more actively than Russian *ja poshel* constructions: *umrjax*!

not *I died but "I've had it/I'm as good as dead/I'm about to die" That same morpheme is also used in imperfective conditions: *Ako bjax*.. If I were ... (but I'm not...). But once linguists call -X- forms "witnessed pasts" and L-forms with zero 3rd person "unwitnessed" then what do they do with those examples, or with -L-forms in expressions of surprise at witnessed events: they create another "L" form borrowed from Albanian, the "admirative".

In my Ch. 14 I believe and in the Xth Congress paper I have another account. Distancing in time OR reality, for both X and L, but X is deictic (most often a shifter) while L is tactic (Jakobson's tense vs. taxis). The English D-preterite is also "distanced in time or reality" but it is not used for imminent events, as it is felt to be "remote" in opposition to the "proximate" present

perfect. In Bulgarian, however, the X-form is felt as "close, proximate, often as witnessed" – but not invariantly, but as a result of opposition with the more "remote" L-forms with 0-auxiliary (marked taxis with L and 0-auxiliary for suspension of speaker's responsibility for the statement).

As I recall, Richard Brecht wrote something about the *ja poshel* items. In Russian that usage is quite restricted, I believe because Russian, unlike Bulgarian, has no "remote" constructions (like the Bulgarian "renarrated" or "non-evidential" forms) to oppose it to. In English such use of the preterite is impossible, and "future" expressions, many of them slangy, use forms of the perfect (as in example translations above) (The English "present perfect's" meaning is "proximate": *I have been happy here* means you still are or at least the feeling is still relevant today, while *I was happy here* means you no longer are). (Just as THIS is +proximate, in opposition to -proximate THAT)

As you point out, resorting to "transpositions" undermines the concept of invariance. In the Waugh and Rudy volume that I reviewed in WORD 1996, there was a polemic between Waugh (who believes in markedness shifts) and her ex-husband Sangster (who doesn't). Shapiro uses shifts all the time, as did Jakobson, but as I point out if you allow for equipollent oppositions you don't need shifts and you can decompose "grammemes" into features that are stipulated as invariant. If you find homonymy or contradictions, then you haven't found the correct "interpretant" of the invariant. (That is, instead of one +/- universal aspect and shifting values – e.g., Perf is M in Russian but in English Progressive is M (as Friedrich and others have claimed) – UG menu offers several options, and some languages use both oppositions, namely Bulgarian,

which has +/- Perfective in lexical stems and +/- continuative carried by an affix, potentially on a P stem, though ++ and – combinations are rarer than the +- and -+ ones and involve some literary device.

Hope this very condensed version is helpful (you have my book, look at my ch. on Bulgarian oppositions).

Best regards,

Catherine" (Catherine Chvany, personal communication, February 9, 2004)

Personal communications as above contain points worthy of further consideration.

From my own experience, these personal communications are illuminating and of considerable help. The issue is how personal communications as above could be reasonably shared as part of academic resources so as to facilitate the study of the Prague School and benefit a wider circle of students. There are hundreds of personal communications kept in Roman Jakobson Papers (Manuscript Collection __ MC72. Institute Archives and Special Collections. The Libraries. MIT). However, unless they go digital and are put on a website, their use is extremely limited.

Perhaps one way out is to include these personal communications in the footnotes. Unfortunately, not everyone understands the value of such footnotes. My experience of writing a Chinese introduction to an upcoming reprint of Jan Firbas 1992 showed that editors sometimes are more concerned with the convenience of type-setting and could tolerate neither footnotes

nor endnotes. For example, I quoted as a footnote a personal communication from Svoboda to explain the term diatheme and the Latin phrase *in medias res*:

"1. Dia- in diatheme has the meaning of Greek "dia" = "through". The history of this term goes back to 1978 when I examined the thematic elements in detail and wished to differentiate the two basic types of theme. My original suggestion was "the point of departure", but Prof. Firbas regarded this term as too much used in a general way that it might mislead the reader, and he asked me to find some other name for this unit. I based my solution on the fact that one of the three main functions of this kind of theme is to mediate the rhematic information into the thematic sphere of the following discourse in a gradual way. For example: "Once upon a time there was a king. The king had three daughters. He thought that he loved all of them in the same way, but he ..." Rheme proper "a king" is followed by a (dia)thematic element "the king", and only after that it is referred to with the minimum language means "he". It is THROUGH (dia) the thematic element "the king" that the item is gradually established in the themtic sphere and can be later used in its minimum form. The course of events may follow some other direction but if the item "king" is to be introduced again, it is re-introduced, not as rheme proper or as theme proper, but as a non-minimum thematic unit – diatheme. Of course, the sequence need not always be "rheme proper – diatheme – theme proper", but it seems to be most frequent especially in texts that are hearer/reader-friendly and do not force the addressee to exert too much effort to follow the speaker's/writer's line of thought."

2. As to the title of the first chapter of Jan Firbas' introduction, you understand it perfectly well. He tries to introduce the reader into the middle of things, into the middle of FSP problems, which are to be dealt with later on. There is another, slightly different meaning used by Firbas in his book and a number of papers: When dealing with the Presentation Scale and the Quality Scale of dynamic semantic functions, he speaks of the way of starting the discourse with the Quality Scale instead of the Presentation Scale by using the stylistic device "in medias res". For example: If the above fairy-tale started with "A king had three daughters", it would skip the Presentation Scale (somewhere lived a king), and directly (in medias res) introduce "a king" as if it had already been introduced (but it was not). (By the way, "a king" is here a diatheme through which "king" is introduced into the discourse.)" (Aleš Svoboda, personal communication, February 20, 2007)

Footnotes or endnotes like this are of remarkable help to readers, but they had to be deleted to meet the editor's requirement.

4. Faced with this age of globalization and digitization, what can be done so as to maximally utilize the Prague School resources?

First, it is desirable that the Prague School writings in the past published in languages other than English (Czech, French, German) should be translated into English (e.g. Mathesius 1907, 1929, 1941, 1942; Trávníček 1937, 1939, 1961), and the Prague School writings coming up in the future should be in English.

When the whole world becomes a global village, the role of English as an internationally accepted academic language has become an established fact. True, articles in *TCLP* are multilingual, but the Prague Linguistic Circle in those days probably did not expect its writings to be read, taught, and studied beyond Europe and North America (in China, for example). The

fact that the Prague School writings are translated, one after another, into English points to such a need (e.g. Mathesius 1961/1975; Vachek 1960/2003). Furthermore, the fact that Chinese presses are increasingly interested in reprinting linguistics books in English also indicates the market value of English.

Second, the high-tech offers an unprecedented opportunity to digitize and store the Prague School writings on a website, or to live-broadcast conferences and talks (e.g. Vilém Mathesius Center Lecture Series), accessible to the international linguistic community. In my view all the articles in *TCLP* and in *TLP* as well as photographs of the Prague School members should be put on a Prague-based website so as to facilitate the study of the Prague School (please visit http://digitalcollections.harvard.edu/ to see how photographs can be digitized and viewed as visual resources). Technically there is no problem to do so. Considering the problems like the tight budget of many linguistics programs, to go digital will not only help remedy the unfavorable situation but also help the Prague School to keep going international. In this aspect *Brno Studies in English* (*BSE*), a Masaryk University-based journal, sets a good example. It has an on-line version at http://www.phil.muni.cz/angl/bse/bse.htm, from which one can have access to articles published in *BSE* from 1959 to 2003, e.g. articles by Josef Vachek, Jan Firbas and their colleagues. In contrast *The Prague Bulletin of Mathematical Linguistics* (*PBML*, at http://ufal.mff.cuni.cz/pbml.html), a Charles University-based journal, offers tables of contents for the volumes 71–85 (1999–2006) and index to the volumes 61–70, only several articles of which are downloadable. Apparently much remains to be done to facilitate the study of Petr Sgall and his colleagues' work

5. To conclude: the past ten years witnessed the activeness of the revived Circle and continued interest in the Prague School. The various responses to the Prague School indicate the impact and richness of the Prague School's legacy. Meanwhile, the changing age characterized by globalization and digitization calls for further effort on the side of the Prague School to facilitate the access to the Prague School resources.

Notes

1 In Luelsdorff (1994), important older contributions by Dokulil (on word formation), Skalička and Sgall (on the types of languages) and others were published. The papers by Hajičová (on information structure), Panevová (on valency) and Sgall (on the underlying sentence structures and semantic interpretation) included in this volume characterize to what degree the classical Prague School approaches offer starting points for a formal description based on syntactic dependency and integrating information structure into the underlying representations (see also Hajičová, Partee and Sgall 1998). More recently, especially the project of Prague Dependency Treebank has been useful, see e.g. Böhmová and Hajičová (1999).

2 Skalička considered morpheme to be characterized by "cumulation of functions" (thus constituting a bundle of "semes", e.g. the case morphemes of nouns in Czech corresponding to semes of case, number and gender), see Skalička and Sgall in Luelsdorff (1994).

Editor's Note

The full texts of the PBML contributions will be available on the web site `http://ufal.mff.cuni.cz/pbml.html` from this volume. As for the publications of the Prague School scholars, financial resources are being searched for the scanning of both pre-war Travaux and post-war series of Travaux, in order to make these publications available also in an electronic form. Full texts of some of the publications of Petr Sgall and his collaborators can be found on the same web page.

References

Böhmová, Alena and Hajičová, Eva. 1999. The Prague Dependency Tree Bank I: How much of the underlying syntactic structure can be tagged automatically? *The Prague Bulletin of Mathematical Linguistics* 71, 5–12.

Chamonikolasová, Jana. 2001. In memory of Jan Firbas. *Brno Studies in English* 27, 7–9.

Chvany, Catherine V. 1996. *Selected Essays of Catherine V. Chvany*. Edited by Olga T. Yokoyama and Emily Klenin. Columbus, Ohio: Slavica Publishers.

Daneš, František. 1974. Functional sentence perspective and the organization of the text. In Daneš, František (ed.), *Papers on Functional Sentence Perspective*. Prague: Academia.1974, 106–128.

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge: Cambridge University Press.

——. 1997. A tribute to Professor Josef Vachek. *Brno Studies in English* 23, 9–14.

Hajičová, Eva, Partee, Barbara H., and Sgall, Petr. 1998. *Topic-focus articulation, tripartite structures, andsemantic content*. Dordrecht:Kluwer.

Hladký, Josef (ed.). 2003. *Language and function: To the memory of Jan Firbas*. Amsterdam: John Benjamins.

Holes, Jan. 2005. Review of *Dictionary of the Prague School of Linguistics*. *Language* 81, Number 2, 521–522

Jakobson, Roman. 1936. Beitrag zur allgemeinen Kasuslehre. *TCLP* 6, 240–288. Reprinted in *SW* II, 23–71. English translation in Jakobson, *Russian and Slavic Grammar Studies, 1931–1981*. Ed. by Linda R. Waugh and Morris Halle. Berlin: Mouton de Gruyter. 1984, 59–103.

——. (ed.). 1975. *Trubetzkoy's Letters and Notes*. Berlin: Mouton.

——. 1985. *Verbal art, verbal sign, verbal time*. Ed. by Krystyna Pomorska

and Stephen Rudy. Oxford: Basil Blackwell.

——. 1987. *Language in literature.* Ed. by Krystyna Pomorska and Stephen Rudy. Cambridge, Mass.: Belknap Press.

——. 1988. *Selected Writings VIII. Major Works, 1976–1980.* (= Completion Volume I.) The Hague: Mouton.

Kirtchuk-Halevi, Pablo I. 2003. Review of Hladký, Josef (ed.). 2003. *Language and function.* http://linguistlist.org/issues/14/14-2977.html#1

Leška, Oldřich. 1995. Prague School teachings of the classical period and beyond.

PLCP 1,3–22.

——. 1999. Prague School Linguistics: Unity in Diversity. *PLCP* 3, 3–14.

——. 2002. Anton Marty's philosophy of language. *PLCP* 4, 83–99.

Luelsdorff, Philip, ed. 1994. *The Prague School of Structural and Functional Linguistics — A Short Introduction.* Amsterdam/Philadelphia:Benjamins. LLSEE, vol. 41.

Luelsdorff, Philip A., Jarmila Panevová, and Petr Sgall (eds.). 1994. *Praguiana: 1945–1990.* Amsterdam: John Benjamins.

Mathesius, Vilém. 1907. Studie k dějinám anglického slovosledu (Studies in the history of the English word order). *Věstíik České Akademie* 16, 261–275.

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

——. 1941. Základní funkce pořádku slov v češtině (The basic function of word order in Czech). *Slovo a slovesnost* 7, 169–180.

——. 1942. Ze srovnávacích studií slovosledných (From comparative word order studies). *Časopis pro moderní filologii* 28, 181–190, 302–307.

——. 1975. *A Functional analysis of present-day English on a general linguistic basis.* Edited by Josef Vachek; translated by Libuše Dušková. The Hague: Mouton; Prague: Academia.

Newmeyer, Frederick J. 2001. The Prague School and North American functionalist approaches to syntax. *Journal of Linguistics* 37, 101–126.

OBITUARY: Stephen Rudy, N.Y.U. professor of Russian and Slavic languages. *The Villager* Volume 73, Number 18 | September 3 - 9, 2003. http://www.thevillager.com/villager_19/ /stephenrudy.html

Qian, Jun. 1997. Review of *Prague Linguistic Circle Papers, Volumes 1 & 2. Prague Bulletin of Mathematical Linguistics* 68,77–79.

——. 1998. *Jiegou gongneng yuyanxue: bulage xuepai* (Structural-Functional Linguistics: The Prague School). Changchun: Jilin Education Press.

——. 2000. Review of *Prague Linguistic Circle Papers, Volume 3. Prague Bulletin of Mathematical Linguistics* 73–74, 77–84.

——. 2001a. Towards a history of linguistic ideas: A Note on Jan Firbas and the Prague School, *Prague Bulletin of Mathematical Linguistics* 76, 5–12.

——. 2001b. (ed., and trans.) *Yakebuxun wenji* (A Roman Jakobson Anthology). Changsha: Hunan Education Press.

——. 2002. Review of *Prague Linguistic Circle Papers, Volume 4. Prague Bulletin of Mathematical Linguistics* 78, 119–127.

——. 2004. Review of *Dictionary of the Prague School of Linguistics. Prague Bulletin of Mathematical Linguistics* 81, 77–81.

——. 2004. Daodu (Introduction to *Praguiana: 1945–1990*, edited by Philip A. Luelsdorff, Jarmila Panevova, and Petr Sgall. Amsterdam: John Benjamins, 1994.), in *Praguiana: 1945–1990*, 1–42. Beijing: Peking University Press.

Rudy, Stephen. (ed.) 1990. *Roman Jakobson, 1896–1982: a complete bibliography of his writings*. Berlin; New York: Mouton de Gruyter.

Salzmann, Zdenek. 2001. Review of *Prague Linguistic Circle Papers. New Series, vol. 3. Language* 77, Number 1, 181–182.

——. 2004. Review of *Prague Linguistic Circle Papers. New Series, vol. 4.Language* 80, Number 3, 623–623.

——. 2005. Review of *Language and function: To the memory of Jan Firbas. Language* 81, Number 2, 528–528.

Sériot, Patrick and Margarita Schönenberger. 2006. *N. S. Troubetzkoy: Correspondance avec Roman Jakobson et autres écrits.* (Edition établie par Patrick Sériot. Traduit du russe par Patrick Sériot et Margarita Schönenberger) Editions Payot Lausanne.

Sgall, Petr. 2006. *Language in its multifarious aspects.* Charles University in Prague: The Karolinum Press.

——, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: D. Reidel.

Stankiewicz, Edward. 1976. Prague School Morphophonemics. In Matejka, Ladislav (ed.), *Sound, sign and meaning: Quinquagenary of the Prague Linguistic Circle.* Ann Arbor: University of Michigan Press. 1976, 101–118.

——. 1977. Roman Jakobson's work on the history of linguistics. In Armstrong, Daniel and C.H. van Schooneveld (eds.), *Roman Jakobson: Echoes of his scholarship*. Lisse: Peter de Ridder Press. 1977, 435–451.

——. 1983. Roman Jakobson: A commemorative essay. *Semiotica* 44, 1–20.

——. 1987. The major moments of Jakobson's linguistics. In Pomorska, Krystyna et al (eds.), *Language, poetry and poetics.* Berlin: Mouton de Gruyter. 1987, 81–94

——. 1991. The concept of structure in contemporary linguistics. In Waugh, Linda, and Stephen Rudy (eds.), *New Vistas in Grammar: Invariance and Variation.* Amsterdam: John Benjamins. 1991, 11–32

——. 1999. Grammatical categories and their formal patterns. *PLCP* 3, 71–89.

Svoboda, Aleš. 2003. Jan Firbas — An outstanding personality of European linguistics. In Hladký (2003), 1–7.

Toman, Jindřich (ed.). 1994. *Letters and other materials from the Moscow and Prague Linguistic Circles, 1912–1945.* Ann Arbor: Michigan Slavic Publications.

——. 1995. *The magic of a common language: Jakobson, Mathesius, Trubetzkoy, and the Prague Linguistic Circle.* Cambridge, Mass.: MIT Press.

Trávníček, František. 1937. Základy československého slovosledu (The foundations of Czechoslovak word order). *Slovo a slovesnost* 3, 78–86.

——. 1939. Slovosled při důrazu (Emphatic word order). *Slovo a slovesnost* 5, 131–144.

——. 1961. O takzvaném aktuálnim členění větném (On so-called functional sentence perspective). *Slovo a slovesnost* 22.163–171.

Vachek, Josef (ed.) 1960. *Dictionnaire de linguistique de l'École de Prague* (avec collaboration de Josef Dubský). Utrecht: Spectrum éditeurs.

——. (ed.) 1964a. *A Prague School reader in linguistics.* Bloomington: Indiana University Press.

——. 1964b. On some basic principles of "classical" phonology. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung.* Reprinted in Makkai, Valerie (ed.), *Phonological theory: evolution and current practice.* New York: Holt, Rinehart and Winston, Inc. 1972, 424–441.

——. 1966. *The Linguistic School of Prague.* Bloomington, Indiana: Indiana University Press.

——. 1968. *Dutch linguists and the Prague linguistic school.* Leiden: Universitaire Pers Leiden.

——. (ed.) 1983. *Praguiana: Some basic and less known aspects of the Prague Linguistic School.* Prague: Academia.

——. (ed.) 2003. *Dictionary of the Prague School of Linguistics.* In collaboration with Josef Dubsky; translated from the French, German and Czech original sources by Aleš Klégr, Pavlina Saldová, Markéta Malá, Jan Cermák, Libuše Dušková; edited by Libuše Dušková. Amsterdam: John Benjamins.

Verleyen, Stijn. 2004. Review of *Dictionary of the Prague School of Linguistics*.

LINGUIST List: Vol-15–654. http://linguistlist.org/issues/15/15-654.html

Waugh, Linda R. and Stephen Rudy (eds.). 1991. *New vistas in grammar: Invariance and variation*. Amsterdam: John Benjamins.

Webb, Eric Russell. 2002. Review of *Prague Linguistic Circle Papers. New Series, vol. 4)*. LINGUIST List 13.3034. http://www.linguistlist.org/issues/13/13-3034.html

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 87–90

REVIEWS

Saussure Studies in China.

Zhao Ronghui (ed.)

Beijing: Commercial Press, 2005. v+495 pp.
ISBN 7-100-04347-6/H·1087

Reviewed by Jun Qian, Peking University

This book (Chinese title *Suoxuer yanjiu zai zhongguo*) is a collection of thirty-three papers, selected from among three hundred papers published by the end of 2003. The collection is headed by the editor's introduction, followed by papers divided into four sections, Introduction to Saussure (2 papers), Bibliographical Studies of Saussure's Works (4 papers), The Roots of Saussure's Ideas (3 papers), and Studies of Saussure's Theories (23 papers).

"Saussure and Saussure Studies in China" (1-48), the editor's introduction, introduces Saussure (1-7) and Saussure studies in China (8-48). As regards the latter theme it covers a variety of issues, such as the translation of Saussure's works, bibliographical studies of *Course*, the roots of Saussure's ideas (Saussure and historical-comparative linguistics, Saussure and sociology, Saussure and psychology, Saussure and William Dwight Whitney [1827-1894], Saussure and economics), studies of Saussure's ideas (langue and parole, synchrony and diachrony, semiotic conception of language, the arbitrariness of linguistic sign, value of the sign, Saussure's philosophy of language). Under each subheading relevant studies are succinctly introduced.

"Saussure in the World and in China" (49-65, Qi Yucun 1996) focuses on the dissemination of and introduction to the Saussure doctrine in different countries (Switzerland, France, Germany, former Soviet Union, the USA, Britain, Czechia, Denmark, and China). "The Myth of a Linguistic Master and Saussure as a Man" (66-76, Li Baojia 2001) attempts to answer questions like why Saussure was deeply dissatisfied with the diachronic approach, and why Saussure did not publish his theories in the form of papers or a monograph in his lifetime.

The four papers in the section Bibliographical Studies of Saussure's Works, i.e. "Two Books on Saussure" (77-113, Xu Guozhang 1983), "On the Russian Edition of Saussure's *Notes on General Linguistics*" (114-133, Xin Delin 1993), "Comparison and Explanation of Two Editions of

Saussure's *Course*" (134-150, Zhang Shaojie & Wang Kefei 1997), and "Keizaburo Maruyama's Study of Saussure's Manuscripts" (151-165, Wei Yulin 1999), deal respectively with Robert Godel (1957) and Tullio de Mauro (1972), N.A. Sljusareva (1990), *Course* (1916) and *Course* (1993), and Keizaburo Maruyama (1981 *Saussure's Ideas*, 1983 *How to Understand Saussure*, both in Japanese).

The section devoted to the exploration of the roots of Saussure's ideas consists of three papers. "The German Roots of Saussure's Linguistics Theory" (166-177, Yao Xiaoping 1993) compares Saussure with his German sources of influence in the areas of the conception of social psychology (Saussure and Hermann Paul [1846-1921]), the conception of system (Saussure and Wilhelm von Humboldt [1767-1835]), language and thought (Saussure and Humboldt), synchrony and diachrony (Saussure and Paul), syntagmatic relations and associative relations (Saussure and Paul). "Emile Durkheim's Sociology and Saussure's Linguistic Theory (178-186, Fang Guangtao 1997) believes that Saussure's linguistic theory and Durkheim's (1858-1917) sociology are identical or similar in some basic aspects. "The Economics Background of Saussure's Linguistic Theory" (187-199, Xiang Mingyou 2000) maintains that Saussure's ideas of synchrony and diachrony, the value of language, syntagmatic relations and associative relations are related with the then economics in varied degrees.

The last section Studies of Saussure's Theories tackles various aspects of the Saussure doctrine, such as langue and parole, synchrony and diachrony, semiotic conception of language, linguistic value, syntagmatic relations and associative relations, the nature of sign, the conception of society, the conception of time and space, the relation between language and writing, the relation between sound and writing, poetics, and Saussure's impact on modern Western thinking. It would be a daunting task to review all these twenty-three papers of such diversity and I would just translate all the titles so that those who do not understand Chinese could have some general idea of Chinese scholars' research interest. "Language, Speech, and Discourse" (200-210, Fan Xiao 1994, it offers the author's own conceptions of these notions rather than discuss Saussure.), "Saussure's Langue and Parole" (211-219, Yang Xinzhang 1996), "More Thoughts on Langue and Parole, Linguistics of Langue and Linguistics of Parole" (220-228, Cen Yunqiang 1996), "On Synchrony and Diachrony of Language" (229-245, Xu Siyi 1980), "Saussure's Semiotic Conception of Language" (246-255, Yue Meiyun 1994), "Reevaluation of Saussure's Semiotic Conception" (256-276, Lu Deping 2001), "Saussure's Theory of Linguistic Value" (277-289, Suo Zhenyu 1983), "Language: Grammatical System, Syntagmatic Relations and Associative Relations" (290-303, Nie Zhiping 1990), "Basic Principles of Saussurean Linguistics" (304-321, Pi Hongming 1994), "The Homogeneity Tendency in the Twentieth-Century Linguistic Research: Saussure's Linguistic Conception and Franz Boas' Methodology" (322-336, Chen Baoya 1997), "Linguistic Theory as a Universal Scientific Theory: Saussure's Linguistic Theory Viewed from a Phenomenological Perspective" (337-351, Xu Haiming 1998), "Arbitrary Sign System and Natural Sign System: Exploration of Saussure's and Halliday's Philosophy of Language" (352-361, Zhang Shaojie 2003), "The Philosophical Significance and Aesthetic Change of Saussure's Linguistic Model" (362-375, Yu Kailiang 2002), "System Study and Binary Thinking: On Saussure's Philosophy of Language" (376-388, Ju Yumei & Cao Chunchun 2002), "On Saussure's Conception of Society in His Linguistic Theory" (389-398, Zhao Ronghui

2000), "Saussure's Linguistic Theory in a Systemic Perspective "(399-409, Zhao Rixin 1996), "A Theme Eclipsed for about One Hundred Years: Conceptions of Time and Space in Saussure's Linguistic Theory" (410-423, Pei Wen 2002), "On Saussure's Remarks about Language-Writing Relation" (424-436, Zhang Pengpeng 1994), "Sound and Writing: Ferdinand de Saussure on Saturnian Verse Form" (437-445, Tu Youxiang 2003), "Restoring Saussurean-Jakobsonian Poetics to its Complexity" (446-455, Lan Luyi 1998), "Linguistic Foundation of Structuralism" (456-462, Gong Xiaobin 2002), "The Impact of Saussurean Linguistics on Modern Western Thinking" (463-472, Chen Benyi 2001), "The Life of the Saussure Doctrine" (473-474, Wang Xijie 2002).

There are two appendices at the end. Appendix One is a Chinese-foreign language contrastive list of foreign names. It would be of much more help if it were a name index. Appendix Two is a chronological list of Chinese publications on Saussure. There is no master bibliography or master references. The references originally at the end of each paper are re-arranged into the relevant places within the paper. As many of these papers are originally published in journals, they should have an abstract, key words, and an English title in line with academic conventions, and yet all these are absent in the collection.

The editor's effort is a laudable attempt and her merit lies in making available papers that are otherwise inaccessible as well as making an objective representation of Saussure studies in China. The collection could serve as a basis upon which one can observe and compare Saussure studies internationally (e.g. Skalička 1948; Jakobosn 1959, 1971; Percival 1981; Čermák 1996; Garcia 1997; Joseph 2002:133-155; Koerner 2002:63-74, 131-150).

References

Čermák, František. 1996. Ferdinand de Saussure and the Prague School of Linguistics. In Eva Hajičová, Oldřich Leška, Petr Sgall, Zdena Skoumalová (eds), *Prague Linguistic Circle Papers*, Volume 2. Amsterdam: John Benjamins, 1996, 59-72.

Garcia, Silvia B. 1997. *Zum Arbitraritätsbegriff bei F. de Saussure.* Münster: Nodus Publikationen.

Jakobson, Roman. 1959. Sign and System of Language: a reassessment of Saussure's doctrine. *Selected Writings II. Word and Language*. The Hague: Mouton, 1971, 272-279.

Jakobson, Roman. 1971. Saussure's Unpublished Reflections on Phonemes. *Selected Writings I. Phonological Studies.* The Hague: Mouton, 2nd extended edition, 1971, 743-750.

Joseph, John E. 2002. *From Whitney to Chomsky: Essays in the history of American linguistics*. Amsterdam/Philadelphia: John Benjamins.

Koerner, E. F. K. 2002. *Toward a History of American Linguistics.* London/New York: Routledge.

Percival, W. Keith. 1981. The Saussurean Paradigm: Fact or Fantasy? *Semiotica* 36, 33-49.

Skalička, Vladimír. 1948. The need for a linguistics of "la parole". *Recueil Linguistique de Bratislava* 1, 21-38. Reprinted in Josef Vachek (ed.), *A Prague School Reader in Linguistics*. Bloomington: Indiana University Press, 1964, 375-390.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 91–92

BOOK NOTICES

Issues in the Left Periphery: A Typological Approach to Topic and Focus Constructions

Sonja Ermisch

European University Studies, Series XXI, Linguistics, Vol. 304, 2007 Peter Lang, Frankfurt am Main, Berlin, ISBN 978-3-631-55432-6, x+182 pp.

Notice by Eva Hajičová

The conceptual basis for the study is Rizzi's 'cartographic' approach (the Split-C hypothesis) developed within the Principles and Parameters theory of Noam Chomsky and concerning the CP (complementizer phrase), i.e. the left-periphery position and the functional projection of C(omplementizer). The main objective of the book is to explore in detail the syntactic and functional characteristics of this position. It is claimed that the functional projections in the left periphery have two tasks, namely to establish a connection between the general discourse and the propositional content of the clause, and to provide a possibility how to account for the information structure (i.e. the topic and the focus of the sentence). The author recognizes several 'topic types' (e.g. true topics, Left Dislocation, 'Hanging topic') and (following up É. Kiss' analysis) he distinguishes between information focus and identificational focus. The main emphasis in the study is on the properties of topic and focus constructions in two African languages, namely Akan and Ewe, in which topicalized and focused elements are overtly marked.

Linguistics in the Twenty First Century

Eloína Miyares Bermúdez and Leonel Ruiz Miyares (eds.)

Newcastle: Cambridge Scholars Press in cooperation with Centro de Lingüística Aplicada, Santiago de Cuba, 2006, ISBN 978-1904303862, x+422 pp.

Notice by Petr Sgall

A selection of papers presented at the 9th International Symposium on Social Communication (Santiago de Cuba) contains 34 papers, among which 8 concern general linguistics, 2 corpus linguistics, 9

natural language processing, and the other belong to most differet areas. A. Nijholt (Enschede) writes on interaction in the virtuality continuum, N. Calzolari (Pisa) on content interoperability in handling language resources, S. Cardey and P. Greenfield (Franche-Comté) on subtlety and flexibility of linguistic systems and language technology. Among other contributions there are those dealing with translation quality standards, with sign languages, with lexical access and with writing as a process and a product. Thus, the miscellany brings a rich selection of contributions ranging over a very large domain of studies in linguistics and communication.

Higher-order Perl: Transforming programs with programs

Mark Jason Dominus

Morgan Kaufmann Publishers, San Francisco, 2005, ISBN 978-1558607019, x+600 pp.

Notice by Pavel Straňák

This book explains many concepts of functional programming on common everyday tasks and shows even experienced programmers that there might be a better way to use Perl than what they are used to. For anyone who comes from C or who is not deeply acquainted with functional programming, this book can be quite important. The author himself says about the book's intent:

> Lisp programmers go around making funny noises like 'cons' and 'cooder,' and they talk about things like the PC loser-ing the problem, whatever that is. They believe that Lisp is better than other programming languages, and they say so, which is irritating. But now it is okay, because now you do not have to listen to the Lisp folks. You can listen to me instead. I will make soothing noises about hashes and stashes and globs, and talk about the familiar and comforting soft reference and variable suicide problems instead of telling you how wonderful Lisp is, I will tel you how wonderful Perl is, and at the end you will not have to know any Lisp, but you will know much more about Perl.
>
> Then you can stop writing C programs in Perl. … Perl is much better at being Perl than it is at being a slow version of C. You will be surprised at what you can get done when you write Perl programs instead of C.

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007

LIST OF AUTHORS

Eva Hajičová
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
hajicova@ufal.mff.cuni.cz

Jiří Havelka
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
havelka@ufal.mff.cuni.cz

Markéta Lopatková
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
lopatkova@ufal.mff.cuni.cz

Marie Mikulová
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
mikulova@ufal.mff.cuni.cz

Jarmila Panevová
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
panevova@ufal.mff.cuni.cz

Martin Plátek
Department of Theoretical Computer Science
and Mathematical Logic
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
platek@ksi.mff.cuni.cz

Jun Qian
English Department
Peking University
Beijing 100871, P.R. China
junqian@pku.edu.cn

Petr Sgall
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
sgall@ufal.mff.cuni.cz

Pavel Straňák
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
stranak@ufal.mff.cuni.cz

Miroslav Týnovský
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
tynovsky@ufal.mff.cuni.cz

Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic
zabokrtsky@ufal.mff.cuni.cz

Šárka Zikánová
Institute of Czech Language and Theory of Communication
Charles University
náměstí Jana Palacha 2
116 38 Praha 1, Czech Republic
sarka.zikanova@ff.cuni.cz

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007

INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

Authors of any contributions receive two copies of the relevant issue of the PBML together with 10 offprints of their article.

The guidelines for the technical shape of the contributions are found on the web site `http://ufal.mff.cuni.cz/pbml.html`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.