1. Introduction

In the era of computers, language processing has gained a form different from what was known before. Vast amounts of data are available and computers can process them in a reasonably short time, but they need adequate tools for their work. Beside grammar rules they also need lexicons which they can understand.

In this work, an electronic lexicon of Czech verbs is presented. The use of the lexicon in Natural Language Processing (NLP) makes special demands on it. It differs from "human" lexicons in that all information must be explicit or deducible by exactly formulated rules of derivation.

While sketching the format of the dictionary, interesting theoretical problems were encountered, which are discussed in this work. Though the lexicon should not depend heavily on a particular theory, so that it can remain usable in another theoretical frame, it is impossible to make it totally theory-free. It is possible, however, to design the dictionary in such a manner that it will not be difficult to adapt it for a particular theoretical frame. The possibility to reuse our lexicon in other frameworks will be discussed at the end of the work.

The lexicon contains valency frames of ca 15,000 Czech verbs, and its purpose is to enrich information contained in other electronic dictionaries. The trend of recent years is to make large-scale reusable sources which can be combined with other sources. This work shows how the lexicon cooperates with an existing morphological lexicon and how it can be used in various NLP systems.

Chapter 2 discusses several theoretical approaches in comparison with Functional Generative Description (FGD), which is used for the dictionary. The explication concentrates especially on the structure of lexicons in single theories. A lexicon usually conforms certain preconditions resulting from using a given theoretical framework and we will explore the possibility of creating a lexicon which would be transferable to another theoretical framework.

Chapter 3 discusses the possibility of using existing sources, with respect to the desired result and the theoretical framework adopted for the work. There were already several Czech syntactic lexicons created in the past, but unfortunately their reuse would be rather difficult. This chapter mentions several such attempts, and describes in detail a lexicon which was used.

Chapter 4 describes the verb frame. In **Section 4.1** we will describe the format of

the lexical entry. In **Section 4.2** we will discuss various types of reflexive constructions in Czech, and their encoding in the lexicon. In **Section 4.3**, possible diatheses of the basic (active) frame are shown, and it is also discussed which of these diatheses can be added to the dictionary on a regular basis and which have to be treated as exceptions. **Section 4.4** describes so called equi and raising verbs.

In **Chapter 5** we will show the procedure of automatic conversion of the source dictionary to the proposed format. For this conversion, an algorithm was created which assigns the functors (semantic roles) to single members of a frame. The output of this procedure will serve as an input for an editor. We will discuss what amount of the source data can be completed by this procedure and what amount needs post-editing. We will also show how the resulting lexicon can be used in NLP systems.

Chapter 6 sums up. In **Section 6.1**, verbs are sorted into groups according their frames, and the results are compared with results of other researchers. In **Section 6.2**, perspectives of the language processing based on symbolic methods are discussed, and the possible usage of the lexicon in corpus linguistics.

1.1. Terminological remarks

Various authors differ their in understanding of the term *subject*. We will consider a subject only such a member of a frame which is in Nominative and with which the main verb agrees. Our criterion is the question for a subject: *kdo, co* (who$_{Nom}$, what$_{Nom}$). This means that we will not take Genitive in such constructions as *vody ubývá* (water$_{Gen}$diminishes) as subject. On the other hand, a clause or infinitive can be subjects, as we can ask the above question; in such a case the verb shows ageement with neuter singular.

In the text, we will use the terms *actants* and *inner participants* as synonymous. *Actant* is Tesnière's term, while FGD uses *inner participant*, but their meaning is so close that they are often interchanged.

We will also use the terms *animate* and *animacy*. For purposes of this work we will divide nouns into two groups: personal and non-personal. The former can be represented by the pronoun *kdo* (who), the latter by the pronoun *co* (what). Sometimes we will refer to personal nouns as to animate ones and to non-personal as to inanimate.

2. Theoretical background

When describing the role of verbs in the language, all authors agree on the necessity to describe syntactic properties of verbs in the dictionary. But they differ in the understanding of what sort of information should be included. Dictionaries for practical usage (language dictionaries for human readers, or machine dictionaries for grammar checking or shallow parsing) contain usually only the surface information.

Dictionaries that serve more sophisticated purposes must contain also information on the argument structure, and the relations between the two layers of linguistic description. The two views of the dictionary differ also in their understanding of what belongs to the verb frame. The classical lexicologists collect all *typical* complementations while the theoreticians discriminate between the *arguments* and *adjuncts*. The arguments are listed in subcat lists and grammar rules check whether all of them are present in a sentence. The adjuncts, on the other hand, are not obligatory, can occur more than once in a sentence, and they are not listed in the dictionary entries.

The dictionary described in this work is meant to provide for the automatic processing of the Czech language. The algorithms for the language processing do not necessarily have to be based on a linguistic theory, but we believe that with a theory we can develop algorithms that are efficient and elegant because they are linguistically adequate. The results of these algorithms, on the other hand, can serve to a linguistic theory as a feedback which helps to improve the theoretical description.

For this work we decided to utilize the Functional Generative Description (FGD) developed by Sgall, Hajičová and Panevová (Sgall et al., 1986), and especially the part dealing with the verb frames (Panevová, 4 75, 1980). We will show later that this does not prevent the dictionary from being used in other theoretical frameworks.

2.1. An overview of FGD

In FGD, several levels of language description are distinguished. For purposes of this work, we will only work with two of them—the *tectogrammatical* level and the *morphemic* level. To be able to express certain relations we will also need the notion of *subject*.

Each level has its own units, basic and compound. The compound units are formed from the basic ones with the help of *C-relations*. The translation between two neigh-

bouring levels is provided by *R-relations*. The basic units on tectogrammatical level are *semantemes* (lexical units), *functors* (syntactic units) and *grammatemes*. The compound units are *propositions*. The functors also serve as the C-relations with the help of which the propositions are constructed (see Sgall, 1967).

There are two types of functors—*inner participants* (Tesnière's *actants*) and *free modifications*. A verb frame denotes which functors are required by a certain semanteme (verb lemma). A frame can contain up to five inner participants (*Actor*, *Patient*, *Addressee*, *Origin* and *Effect*) and any number of free modifications. Some of the inner participants can be *optional* (also called *facultative*), which means that they do not need to be present in the sentence—neither on the tectogrammatical nor morphemic level. Other participants are always *obligatory*. However, they can be realized as *general*—the structure on the tectogramatical level then contains a general participant, which is not realized on the morphemic level. Whether a participant is optional or obligatory, and whether an obligatory participant can be realized as general can be tested by a question test (Panevová, 1980, pp.29-30). Let us imagine the following dialogue:

(1) *Petr čte.* *Co?* *Nevím.*
Petr is reading. What? I don't know.

The answer 'I don't know' is acceptable, as the the speaker does not need to know what Petr's reading is, but it must be something which is usually read (a newspaper, a book, etc). This shows that Patient in the frame of the verb *číst* (read) can be general. On the other hand, in dialogue (2), the answer 'I don't know' is nonsensical. This shows that Actor is an obligatory participant in the frame of the verb *přijít* (come).[1]

(2) *Už přišel.* *Kdo?* **Nevím.*
(He) has already come. Who? I don't know.

In example (3) the sentence is actually ungrammatical, if the participant is omitted—this is clear evidence that the participant is obligatory.

(3) **Petr daroval.*
Petr donated.

Free modifications normally are not members of a frame, but they can become members as *obligatory free modifications*:

(4) a. *Jan se choval jako blázen.*
Jan behaved like a fool.

[1]The fact that the surface realization of Actor in this sentence is omitted is caused by another phenomenon: Czech is a so called pro-drop language and thus a personal pronoun in the position of a subject can be omitted. Morphological markers of the person and number (in the past tense also of the gender) are present also in the verb form and thus the personal pronoun is redundant (see Karlík, 2000).

b. **Jan se choval.*
Jan behaved.

In some cases, when the modification is known from the context, it can be omitted on the surface; such free modification is called *obligatory and deletable free modification.* For testing whether a free modification is an obligatory member of a frame the question test can be used again. In the sentence in (5) the question test proves that the direction is an obligatory and deletable free modification of the verb *přijít* (come, arrive).

(5) *Petr přišel. Kam? *Nevím.*
Petr arrived. Where? I don't know.

In other theoretical models (Daneš et al., 1987a; Grepl and Karlík, 1989; Karlík et al., 1995), the repertory of participants is wider: instead of Actor the authors speak about Agent, Causer, Experiencer, etc. Patient is more or less a synonym of the direct object and Recipient a synonym of the indirect object. In FGD, Actor and Patient are determined by syntactic criteria rather than by semantic ones (cf. Tesnière, 1959), and other participants are determined semantically:

(6)
1. If the verb frame contains only one participant, this participant is Actor.
2. If the frame contains two participants, one of them is Actor and the other is Patient. In most cases, Actor is the subject of the active construction, but there are some exceptions to this rule, which will be discussed later.
3. If the verb frame has more than two participants, the roles of Actor and Patient must be occupied, and the other participants occupy the roles of Addressee, Effect or Origin. The decision about which participant bears which role is based on the semantics of the participants.

The basic units on the morphemic level are *semata*, and the compound units are *morphemes* and *formemes*—units which combine prepositions with morphological cases.

The lexicon in FGD contains semantemes, their functors and grammatemes. In our informal example, parantheses denote whether a functor is obligatory, obligatory deletable or optional:

(7)

spát	Act		
pojídat	Act	Pat	
těšit_se	Act	Pat	Gram:{Refl[se]}
darovat	Act	Pat	(Addr)

Beside it, the lexicon should also define the R-relation which translates every functor and grammateme to the morphemic level. After this addition, the lexicon will have the following format:

(8)

spát	Act[Noun+Nom]		
pojídat	Act[Noun+Nom]	Pat[Noun+Acc]	
těšit_se	Act[Noun+Nom]	Pat[Noun+Acc+na]	Gram:{Refl[se]}
darovat	Act[Noun+Nom]	Pat[Noun+Acc]	(Addr[Noun+Dat])

2.2. Comparing FGD with other theories

In this section, a short comparison of the main contemporary linguistic theories is provided and the possibility of interchange of a common dictionary is discussed.

2.2.1. Government-Binding Theory

In Government-Binding Theory, the lexicon contains words with subcat lists and lists of θ-roles. The match between arguments of a verb and θ-roles is taken care of by θ-Criterion:

> Each argument bears one and only one θ-role, and each θ-role is assigned to one and only one argument.

The match between categories in a subcat list and θ-roles is called θ-marking:

> If α subcategorizes the position occupied by β, then α θ-marks β.

The lexicon in GB then contains the *word*, its *category*, *subcat list* and list of θ roles:

(9) sneeze, V, (Agent)
devour, V, <NP>, (Agent, Theme)
donate, V, <NP, PP>, (Agent, Theme, Goal)

where the θ-roles are results of θ-marking.

Subjects do not occur in subcat lists, as it is presupposed that every verb has a subject. The theoretical explanation is that the subject is an *external* argument.

Passivization in GB is provided by movement rules. Active sentences are transformed to other constructions and ungrammatical structures are then ruled out by various principles that exploit θ-roles assignment to single complementations.

2.2.2. Lexical-Functional Grammar

In Lexical-Functional Grammar, the dictionary plays the central role (as the name suggests). The theory works with grammatical categories (as NP, S′, XCOMP, etc.) and grammatical functions (as Subject, Object, etc.). Categories are used for constructing c-structures, while functions are used for f-structures. The lexicon in LFG has the following format:

(10) *sneeze* V (↑PRED) = 'sneeze<(↑SUBJ)>'
devour V (↑PRED) = 'devour<(↑SUBJ), (↑OBJ)>'
donate V (↑PRED) = 'donate<(↑SUBJ), (↑OBJ), (↑OBL_{GO})>'

The theory does not work with *arguments* directly, but it supposes some sort of linking between θ-roles and grammatical functions called *Predicate-Argument Structure*. An enhancement of the theory is the *semantic structure* which, however, works with concepts as ARG_1, ARG_2, rather than with θ-roles.

Frames for passive sentences are created with the help of lexical rules which may have the following form:

(11) (↑SUBJ) ↦ NULL
(↑OBJ) ↦ (↑SUBJ)
˜(↑TENSE)
(↑PARTICLE)=$_c$ PASS

These rules erase the original subject from the frame, move the object to its place, and they add two constraints on the verb form—it must not be a finite form and the value of the attribute PARTICLE must be PASS.

2.2.3. Head-Driven Phrase Structure Grammar

HPSG works with *signs* which are in fact *well typed* attribute value matrices (AVM).[2] The whole grammar is based on combining AVM's together with the help of unification. A lexical entry has a form of an AVM, too:

(12) *walks*

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & {}^{verb}\big[\mathit{fin}\big] \\ \text{SUBCAT} & \big\langle \text{NP}\big[\mathit{nom}\big]_{\boxed{1}[3rd,sing]} \big\rangle \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{RELN} & \mathit{walk} \\ \text{WALKER} & \boxed{1} \end{bmatrix} \end{bmatrix}$$

[2]The term *well typed AVM* means that what attributes can appear in an AVM is determined by its type.

sees

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \overset{verb}{\left[fin\right]} \\ \text{SUBCAT} & \left\langle \text{NP}[nom]_{\boxed{1}[3rd,sing]},\ \text{NP}[acc]_{\boxed{2}} \right\rangle \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{RELN} & see \\ \text{SEER} & \boxed{1} \\ \text{SEEN} & \boxed{2} \end{bmatrix} \end{bmatrix}$$

gives

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \overset{verb}{\left[fin\right]} \\ \text{SUBCAT} & \left\langle \text{NP}[nom]_{\boxed{1}[3rd,sing]},\ \text{NP}[acc]_{\boxed{2}},\ \text{NP}[acc]_{\boxed{3}} \right\rangle \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{RELN} & give \\ \text{GIVER} & \boxed{1} \\ \text{GIVEN} & \boxed{2} \\ \text{GIFT} & \boxed{3} \end{bmatrix} \end{bmatrix}$$

The valency frame is contained in the attribute SUBCAT. A mapping between the subcat list and CONTENT is provided by the indices (1, 2, 3, etc). The attributes in CONTENT are not marked as ARG_1, ARG_2, etc., as one would expect but their names are derived from the verb lemma. For linking the arguments with θ-roles, so called *linking theory* is used.

Passive frames are created with the help of lexical rules. They change the chracteristics of the verb form and cyclically permute subcat lists, as shown in (13):

(13) SUBCAT $<\text{NP}_1,\text{NP}_2> \mapsto$ SUBCAT $<\text{NP}_2,\text{PP}[by]_1>$
SUBCAT $<\text{NP}_1,\text{NP}_2,\text{NP}_3> \mapsto$ SUBCAT $<\text{NP}_2,\text{NP}_3,\text{PP}[by]_1>$
SUBCAT $<\text{NP}_1,\text{NP}_2,\text{PP}[to]_3> \mapsto$ SUBCAT $<\text{NP}_2,\text{PP}[to]_3,\text{PP}[by]_1>$
...

The resulting lexical entries then look as shown in (14):

(14) *seen*

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \overset{verb}{\left[pass\right]} \\ \text{SUBCAT} & \left\langle \text{NP}[nom]_{\boxed{2}[3rd,sing]},\ \text{PP}[by]_{\boxed{1}} \right\rangle \end{bmatrix} \\ \text{CONTENT} & \begin{bmatrix} \text{RELN} & see \\ \text{SEER} & \boxed{1} \\ \text{SEEN} & \boxed{2} \end{bmatrix} \end{bmatrix}$$

$$given \quad \left[\begin{array}{ll} \text{CAT} & \left[\begin{array}{ll} \text{HEAD} & {\scriptstyle verb}\left[pass\right] \\ \text{SUBCAT} & \left\langle \text{NP}[nom]_{\boxed{2}[3rd,sing]},\ \text{NP}[acc]_{\boxed{3}},\ \text{PP}[by]_{\boxed{1}} \right\rangle \end{array}\right] \\ \text{CONTENT} & \left[\begin{array}{ll} \text{RELN} & give \\ \text{GIVER} & \boxed{1} \\ \text{GIVEN} & \boxed{2} \\ \text{GIFT} & \boxed{3} \end{array}\right] \end{array}\right]$$

Some authors argue (Oliva, 1994; Kathol, 1994) that lexical rules are not necessary, as the desired effect could be achived by applying constraints on the hierarchy of types, but we will not go to details here.

2.2.4. Comparison with FGD

In all the above mentioned theories, some sort of mapping between surface forms and θ-roles is supposed, whether it is called θ-marking, predicate-argument structure, or linking theory. The common feature is that subcat lists are viewed as primary syntactic structure attached to lexical entries and the θ-roles are mapped onto the subcat list by some sort of mapping function.

In FGD the opposite assumption is made: the tectogrammatical functors form a primary syntactic structure of a verb and the surface forms are their counterparts on the morphemic level which are translated by R-relation from the functors.

Beside this, the θ-roles differ from the repertory of participants in FGD. Not only are their names different, but also their distributions to single verbs. An *Actor* in FGD can be marked as *Agent* or *Bearer* or *Experiencer* in other theories, etc.

If we use FGD as the background theory of a dictionary, we will be unable to transfer the lexicon to another theoretical framework 'as is'; it should not be difficult, however, to extract the subcat lists. It will be shown in Chapter 6 that this is possible and feasible. For utilizing the tectogrammatical information, we would have to find a mapping function which would have to take into consideration also the semantics of single verbs, which will be the subject of further research.