3. Using existing sources

When we try to create a new electronic dictionary, it is of course possible to start from scratch, but it is more efficient to use existing sources. Printed dictionaries usually contain syntactic information, but unfortunately this information is meant for human readers, and very often it is assumed that the reader knows the rules that apply in usual cases, and only exceptions are listed. Beside this, the information is not encoded in a formal way which could be understandable to a machine.

There exists a Czech dictionary of verbs (see Svozilová et al., 1997) which contains the verb frames encoded in a formal way. But its size is quite limited (ca 600 verbs) and the information concerns only the surface frames. Nevertheless, this dictionary can serve as an aid to creators of an electronic dictionary.

One of the first attempts at making an electronic dictionary of verb frames was made in the project RUSLAN (see Oliva, 1989). This project was focused on machine translation from Czech to Russian and the format of the lexicon was adapted for this purpose; it contained the Czech word stem and its Russian translation, Czech and Russian morphological information, the Czech surface frame and its translation to the Russian surface frame. The domain of the translated texts were programming manuals, which affected the coverage of the lexicon. Another drawback (caused by limited computational resources) was the small size of the lexicon—it contained ca 10,000 entries (including all word classes). The work invested in this project was useful for gaining experience with natural language processing rather than for creating working software.

Another small lexicon was created for the purposes of the project LaTeSlav (see Avgustinova et al., 1995). This was a project for creating grammar-checkers for two Slavic languages (Czech and Bulgarian). In fact, there were two lexicons for Czech, as the project split into two branches. Both the lexicons (see Oliva, 1996; Skoumalová, 1994) contained a small number of entries which had very rich syntactic information, but unfortunately they were "hardwired" in the software and it would not be easy to extract them for other purposes.

The most promising source of valency frames is a dictionary created at Masaryk University by Karel Pala and his team (see Pala and Ševeček, 1997; Horák, 1998b). This dictionary was compiled from several printed dictionaries, and the valency frames were taken mainly from SSJČ (1989). We used this dictionary as a source of surface frames and enhanced them with information at the tectogrammatical level.

3.1. Source data

The dictionary contains ca 15,000 verbs with surface frames. The original format called BRIEF contains lemma, starting delimiter of the list of frames (<v>) and the list itself (see example in 15a). (15b) translates this notation to a readable form.

(15) a. `agitovat <v>hPc4,hPc3-hPc4,hPTc4r{pro},hPTc3r{proti}`

b. **agitovat** (to agitate) *koho* (*komu*), pro *koho*, proti *komu*

In BRIEF format, frames are separated by commas, and single members of a frame are separated by dashes. The obligatoriness is not marked, but a frame can be repeated several times, with and without the optional, deletable or generalizable members. In example (15) this is the case of the frame *koho* (*komu*).

BRIEF encoding is described in Horák (1998a,b). Here, we only provide a short overview of attributes and values used in the dictionary. Every member of a frame is described by a list of attributes and their values. We can understand these attributes and their values as grammatemes occurring on the tectogrammatical level.

3.1.1. The attributes used in the lexicon and their values

h — 'Semantic' feature. This attribute has rather heterogeneous values. Single values are only applicable for certain word classes and thus they include implicit information on the part of speech as well. The values are:

- P — Person (only applicable for nouns and pronouns); this value actually stands for 'case questions' *kdo* (who), *koho*, etc.
- T — Thing (only nouns an pronouns); it stands for 'case questions' *co* (what), *čemu*, etc. The values P and T can be grouped together.
- R — Long reflexive pronoun *sebe*, *sobě*, etc.
- Q — Quality (only adjectives).
- M — Amount (only numbers).
- L — Location (only adverbs).
- A — Direction where (only adverbs).
- F — Direction from (only adverbs).
- D — Which way (only adverbs).
- W — When (only adverbs).

c — Morphological case. This attribute is only applicable for nominal word classes, and so it only occurs if the h attribute has one of the values P, T, R, or Q. The values are 1, 2, 3, 4, 6 and 7.

`r` — Preposition. This attribute can only occur after a morphological case. The value is the preposition itself closed in curly brackets: `r{na}`, `r{o}`, `r{vzhledem k}`, etc.

`s` — Clause or infinitive. The values are:

- `I` — Infinitive.
- `C` — Clause attached by the conjunction *až* (when).
- `D` — Clause attached by the conjunction *že* (that).
- `F` — Clause attached by the conjunction *jestli*, *zda* (if, whether).
- `P` — Clause attached by the conjunction *ať* (let).
- `R` — Clause attached by a relative expression *co* (what), *který* (which), *kdo* (who), *kolik* (how many), etc.
- `U` — Clause attached by the conjunction *aby* (so that).
- `Z` — Clause attached by the conjunction *jak* (how).

`e` — Negation (in a clause). The values are `A` (affirmative) and `N` (negative). The affirmative value is the default and it is not marked in the lexicon.[1]

`i` — Idiom. The value is a string closed in curly brackets. The string contains words forming the idiom and a case marker for the variable part. If there are possible variants in the fixed part, they are put in parentheses and separated by commas, or they are separated by a vertical bar. The variants in the variable part are separated by a vertical bar. Examples:

```
brát <v>i{pod ochranu|do ochrany <koho>}
                                        (place sb under protection)
dávat <v>i{koncert|hru|film}          (put concert, play, movie on)
házet <v>i{přes palubu <koho|co>}              (throw sb over board)
chovat <v>i{(přátelství, zášť, nenávist) <ke komu>}
                                          (feel friendship, hatred)
```

`v` — Constraint applied for a single valency frame. The constraint is an attribute with a required value, or an attribute with a prohibited value, preceded by `^`. Currently, only `v{eN}` is used, for verbs whose negated forms have different valency frames:

```
hledět <v>hPTc4r{na},hPc3,hTc2r{do},hPc3-hTc2r{do},v{eN}hTc3r{k}
                                                  (not to look at st)
```

[1] This attribute is mainly used together with a clause attached by the conjunction *aby* (so that)—`sUeN`, e.g. *bát se* (fear), *varovat* (warn), etc. Though this is a typical usage, the affirmative clause cannot be excluded. After a simple query in the Czech National Corpus (Kocek et al., 2000) we found eight affirmative clauses (out of ca 230 occurrences of the verb *bát se* with the conjunction *aby*), e.g. *Po volbách se úředníci bojí, aby přežili ... změnu dnešního ministra ...* (After elections, clerks are afraid whether they will survive the change of the current minister ...).

```
páchnout <v>hTc6r{po},hTc7,v{eN}hTc2r{do} (not to set foot on st)
znát <v>hPTc4,v{eN}hTc2z{jen se záporem},hTc4-hPTc6r{na}
                                (not to know--Genitive of negation)
```

`z` — Comment in curly brackets (see the example above).

The frames do not contain subjects as the printed dictionaries usually do not list them. For an automatic processing of language, however, this information is necessary. We can make a simple assumption that the subject will be a noun in Nominative but there are exceptions to this rule. We will discuss this in more detail in Chapter 5.