Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS

Karel Vandas

Automatické určování sémantických preferencí pro slovesná valenční doplnění

Automatic Identification of Semantic Preferences for Valency Complementations of Verbs

Institute of Formal and Applied Linguistics (ÚFAL)

Supervisor of the master thesis: doc. RNDr. Markéta Lopatková, PhD.

Study programme: I3, Mathematical Linguistics

Prague 2012

Dedication. I dedicate this work to all abandoned ellipses and unrelated coreferences. First of all I am grateful to my supervisor doc. RNDr. Markéta Lopatková, Ph.D. for valuable comments, suggestions, help and time my supervisor has devoted to this thesis. Sometimes I had a feeling that my supervisor works fulltime on it despite the fact how many duties she has as a head of a department. I also want to thank very much to RNDr. Ondřej Bojar, Ph.D. and RNDr. Daniel Zeman, Ph.D. for their valuable help in a war with unsuccessfull parsing and other morphology fights. I also want to thank to my brother Jakub for feeding me in time of deprivation and hunger, to my motivating friends Maria, Radek, Marjanka and Ioanna, mum Zuzana, sister Lucie and father Karel for their support. Without you this thesis would not happen. **Thank you!**

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In .... date .... signature of the author

Název práce: Automatické určování sémantických preferencí pro slovesná valenční doplnění

Autor: Karel Vandas

Ústav: Ústav formální a aplikované lingvistiky (ÚFAL)

Vedoucí magisterské práce: doc. RNDr. Markéta Lopatková, PhD., ÚFAL MFF UK

Abstrakt: Slovesná valence hraje důležitou úlohu v popisu chování sloves a propojuje povrchovou realizaci jazyka s jeho sémantikou. Sloveso samotné může být použito ve více významech. Slovesná valenční doplnění pak pomáhají identifikovat správné čtení slovesa. Dosud byla většinou slovesná valenční doplnění studována zejména z morfologického a syntaktického hlediska. Účelem této práce je vyhodnotit možnosti automatického určení sémantických preferencí pro valenční slovesná doplnění. Práce taktéž porovnává úspěšnost systému s různými úrovněmi dostupné informace o valenci ve spojení se shlukovou analýzou. Práce je zakončena evaluací dostupných metod a jejich vzájemným srovnáním.

Klíčová slova: slovesná valence, slovesná valenční doplnění, sémantika, shluková analýza

Title: Automatic identification of semantic preferences for valency complementations of verbs

Author: Karel Vandas

Department: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Markéta Lopatková, PhD., ÚFAL MFF UK

Abstract: Verb valency plays an important role in the description of behaviour of verbs and connects surface realisation of language with its semantics. Verb itself usually encodes several readings. Complementations of a verb help to identify correct reading of the verb. So far valency verb complementations are mostly studied from morphological and syntactical point of view. The purpose of this thesis is to examine possibilities of automatic identification of semantic preferences for valency complementations of verbs. The thesis discusses performance of system with different levels of available verb valency information in connection with cluster analysis. The thesis contains an evaluation section that compares available methods and their comparision.

Keywords: verb valency, verb valency complementations, semantics, cluster analysis

Contents

Chapter 1

Introduction

1.1 Introduction to the Thesis

Natural language processing (NLP) is a field of science concerned with processing of human language by computers. Throughout the time there has been development resulting in lots of different disciplines such as machine translation, speech recognition, discourse analysis, natural language generation, parsing but also lots of different attitudes varying from pure statistical, through using hand-made grammars and logics, to a pure semantical processing.

Recently dependency grammars and research concerning a verb as a centre of a sentence become a standard.[1] Not only syntactical aspects of sentences play role in research. Also semantical aspects of sentences and other disciplines as discourse or coreference gain on importance.

Verb became a centre of research and this thesis tries to enhance a verb description with respect to the selectional preferences of its valency complementations and describe verb interactions with its complementations. Certain verb readings can be only distinguished by complementation selections, therefore we focus on verb and its complementations, their type and we try to find a way to predict their abstractions.

For the thesis we focus on the Czech language. There is a deep description of the Czech language up to tectogramatical layer, see **section 2.2** (Verb Valency and the FGD Formalism on page 4). The tools capable of processing the raw data are available as well. Also valency lexicons and ontology for the Czech language are available. On the other hand, attitudes described in the thesis are general enough to be used with any language with the same level of description.

1.2 Structure of the Thesis

In chapter **Verb Valency** on page 3 we discuss the basics of verb valency, we show examples of usage and difficulties of its description. General introduction serves as a context reading and might be skipped by readers with verb valency experience.

Chapter **Cluster Analysis** on page 11 introduces the basics of theory of clustering and similarity analysis of texts.

[1] Let's leave this statement without a proof only on reader's intuition.

Chapter **Data** on page 21 offers an overview of the data resources such as Vallex, Valeval, CzEng or WordNet. For further data descriptions please see **chapter B** (Really Technical Details on page 73). Readers more interested in application of the selected approaches and not really interested in data resources details might skip this chapter.

Chapter **Experiments** on page 25 is the core part of the thesis. Experiments show how the system and presented methods perform.

Chapter **Evaluation** on page 33 tries to propose methods of evaluation for the system.

Chapter **Discussion** on page 53 gives an insight into the results discovered by the system.

Chapter **Conclusion** on page 55 presents outcomes of experiments and discuss the final benefits of the thesis.

List of figures, tables and bibliography follows.

Chapter **The User Guide** on page 63 is a user documentation of a created tool giving more detailed insight into its functions, installation and control.

Chapter **Really Technical Details** on page 73 takes you behind the scene.

Chapter 2

Verb Valency

"Everything should be made as simple as possible, but not simpler," Albert Einstein said. Natural language as we know it today is a complex system. It has been developing for generations with different types of influence on it. It is sufficient to use, for human it is easy to obtain but it's still very hard to describe and process automatically by machines.

Language itself lacks explicit mathematical models that could be used as patterns for its usage. Models that exists nowadays are just a simple approximation of how language is being used. Most of those models are still heavily based on surface form of a language, although spoken version might differ.[1]

Linguists and people interested in machine processing of a language are trying to find approximations, models or explanations for individual behaviour of the language. There are many theories that describe particular features of languages, but they never capture a language as a whole system.

We believe that the description of a language can't be simply based on its surface form. It can't even be only based on its analytical structure as it was demonstrated a long time ago, as shown in [21].

Besides the quite well defined syntactic structure a language contains a lot of constructs that are acquired by speakers of the language either historically or because of language economy. One of these constructs is an ability of verbs to connect with certain words in a sentence and to fill in different semantical roles needed to express the correct meaning of the verb.

Let's define the concept of relation of a verb and its complementations in a sentence.

Definition. "Verb valency is the range of syntactic elements either required or specifically permitted by a verb or other lexical unit," according to [13].

2.1 Verb Valency as a Step to Meaning

Example. The example "Colorless green ideas sleep furiously," from [21], has a perfect analytical structure. All the constituence fill sentence structure according to a well-defined pattern. The meaning of this sentence is not making any sense. Complementations of verbs are not properly chosen that is why we feel the sentence has no meaning.

[1] As an example let's look at the Czech language formal and informal usage.

Verb valency introduces particular patterns in a sentence's deep structure. The concept of verb valency follows the path of disambiguation of the meaning of verbs and the valency itself represents a step from the analytical layer to the deep structure of the sentence capturing both - the sentence structure as it is constructed and its semantics where all the verb complementations take over particular roles.

The problem of verb valency lies in the need of a manual identification of senses of a word. Some of verb senses do not occur even in a large corpus. Also annotators need to be skilled and even then the inter-annotator agreement might score only around 75%, as shown in [7]. Another problem lies in a verb valency identification and its description in valency lexicons - lexicons are descriptive and its usage when disambiguating verb senses is still quite limited when applying on real data.

2.2 Verb Valency and the FGD Formalism

Functional Generative Description (FGD) is a system for the language description. "FGD is a (dependency-based) stratificational approach, i.e. it decomposes the description of language into a system of levels.," according to [6]. The levels, or layers, are tectogramatical, surface-syntactic, morphological, morphonological and phonetic, as shown in [22]. "Valency theory ... is one of the core components of FGD, especially of its tectogrammatical level," according to [6].

Following key parts are mentioned in [6]:

> "
>
> - verbal complementations (dependents) can be classified either as inner participants (actants,arguments) or as free modifications (adjuncts),
> - the relation between the governor and its dependent is labeled with a functor; five functors for actants are distinguished: actor, patient, addressee, origin, effect; functors also distinguish between various types of temporal, locational, causative and other free modifications,
> - both actants and free modifiers can be either obligatory, or optional for the given verb; the so called dialogue test was introduced as a criterion for distinguishing obligatory and optional dependents,
> - a valency frame (in the narrow sense) contains only actants and those free modifiers which are obligatory for the given verb,
> - a verb's valency in the wider sense concerns also all of its optional adjuncts; the present thesis is not concerned with this aspect,
> - the concept of shifting of cognitive roles is used when assigning functors to an actant: if a verb has one actant, it is always actor; if there are two, one is always actor and the other is patient, no matter what its cognitive role with respect to the verb is; only if there are three or more actants, semantic criteria come into play.

"

2.3 Valency Lexicons

Valency lexicons provides information on the valency structure of verbs in their particular senses. Each of senses is identified by a valency frame - a description of a particular meaning accompanied with a functor definitions and types.

- Vallex [23] is a typical example of a valency lexicon. It describes the structure of verb complementations and gives various examples of its usage. The content of this lexicon has been created to cover most frequent verbs in Czech language. The lexicon is manually annotated and various examples of valency frames are used. The information about the semantics of a verb is captured implicitly (as a valency frame entry).

  Uses of verbs are captured in an analytical structure and distinguished just by different types of verb complementations (so called functors) and their different expressions. At some cases the correct meaning of a verb in a sentence can be only distinguished by a verb complementation's surface realisation.

  Vallex 2.5 describes 2730 verb lexemes containing about 6460 lexical units typically corresponding to one sense, as shown in [23]. Vallex is available as a web based lexicon, or as a XML file.

- Another lexicon is PDT-Vallex [1], that is built based on real occurence of verbs in the Prague Dependency Treebank. "The valency lexicon PDT-Vallex has been built in close connection with the annotation of the Prague Dependency Treebank project (PDT) and its successors (mainly the Prague Czech-English Dependency Treebank project, PCEDT). It contains over 11000 valency frames for more than 7000 verbs which occurred in the PDT or PCEDT. It is available in electronically processable format (XML) together with the aforementioned treebanks (to be viewed and edited by TrEd, the PDT/PCEDT main annotation tool) , and also in more human readable form. The main feature of the lexicon is its linking to the annotated corpora - each occurrence of each verb is linked to the appropriate valency frame with additional (generalized) information about its usage and surface morphosyntactic form alternatives," according to [1].

- The Verbalex valency lexicon [4] is a project of the Centre of Natural Language Processing of Masaryk University. It contains 6256 the Czech verb synsets, 21032 literals, 10469 verb lemmas and 19247 valency frames. This valency lexicon is not publicly available.

- Let's mention some English Valency lexicons, such as PropBank or FrameNet.

  Let's continue with descriptions as given in [6]:

  > " The main goal of the Proposition Bank project is to add a level of semantic annotation into the phrase-structure Penn Treebank trees. The Berkeley FrameNet project is aimed at creating an online lexical resource for English, based on frame semantics. Its

> goal is to document the range of semantic and syntactic combinatoric possibilities of each word (especially verbs and *frame-bearing* nouns) in each of its senses. ”

In Czech there are many examples of verb complementations of the same verb that fit into the same analytical structure but differ in its meaning.

Example. The word *uspořádat* is a Czech verb with three[2] different valency frames. If we consider example *uspořádat výlet/hostinu/svatbu* the meaning tells us about making an event as a trip, reception or a wedding. On the other hand, *uspořádat složky/noty* tells us about sorting out the files or note sheets. The meaning in both cases is different although the analytical structure of the sentence is the same.

Example. Another example might be the verb *chovat* which has also two semantically different readings. The first reading reffers to *cradling a baby* (*chovat dítě v náruči*). The other reading reffers to *breed an animal* (*chovat koně*). Also this example has the same syntactic markers.

Example. Let's consider the word *zavřít.* This verb can be connected with object reffering to criminals as *zavřít darebáka* (*lock the criminal in the jail*) or *zavřít zločince.* It also might reffer to close physically doors or a window *zavřít dvěře, okno.* Another reading more reffers to a phrasal use of the word, i.e., *be quiet* (*zavřít pusu*).

Describing verbs by their valency frames is a challenging task as shown in given examples.

2.4 Advantages of Usage of the Verb Valency and Complementation Abstraction

- Particular verb valency complementations typically share common semantic characteristics.

  Example. One would have hard time trying to describe a word *zapojit* (*to connect, to plug or to use*). You can plug a usb stick and this meaning is used in a majority of cases. One can also point out that a person should think more about the problem (*zapoj hlavu* (*use the head*)). Making an abstraction over both of meanings make no sense therefore a description of a verb by its valency frame helps in this case.

- The detailed description of a verb and its complementations can help create links for machine translation systems to transfer the meaning from one language to another.

  Example. There are also other examples of English verbs with the same lemma being translated differently to Czech and vice versa. Let's consider *make a bed* and *make a tea.* In both cases we use verb 'make' but the complementation of the verb differs. There is a possibility to translate both as *udělat postel* or *udělat čaj*, but the translation that would be better accepted by speakers of the Czech language would be *ustlat postel* and *uvařit čaj* respectivelly.

[2] according to VALLEX 2.5

2.5 Difficulties of the Verb Valency

- **Manual annotations.** Verb valency is a feature of semantics that can't be recognized fully automatically nowadays, therefore manually annotated lexicons are used. Manual annotation is hard because the border of the meaning between two valency frames can be quite blurred and hard to recognize even for professional annotators, as shown in [7]. Manual annotation also costs a lot of time of professionals and it is very expensive in the environment of a university research.

- **Automatic identification based on manual annotations.** Even with a manually annotated training corpus it is hard to train a system capable of correctly distinguishing valency frames. Valency lexicons are therefore a useful source of knowledge but still not widely used in language applications.

2.6 Further Description of Verbs

Valency lexicons so far are an overview of lexical items for each lexeme. They capture the verb, its complementation types and morphological form with several examples. Valency frame doesn't contain the full semantic information (i.e., semantic features, types) about all the verb complementations that might take over the role of a particular frame member.

2.7 Possible Improvements of Valency Lexicons

Example. Let's consider a situation when we know a semantics of complementations of a verb. In more detail, we might know that a verb drink usually connects with a liquid as its complement, i.e., *drink water*. If we then find in a corpus *drink a stone* we might consider it as a mistake in the corpus annotation because we might know that all the suitable complementation types are covered by a liquid type[3]. In case of the verb *drink* the situation is quite straightforward, but we could also return back to the verb *to make* with *make a bed* and *make a tea* and based on the complement type (furniture or a liquid) we might be able to distinguish the correct meaning of the verb and use the correct valency frame. With a correct valency frame we could map one language's valency frame to other language's valency frame and improve the translation tasks.

Disambiguation of semantics of a verb. Complementations of a verb and their types and abstractions might help to identify a correct valency frame of the verb usage. So far, however, there has been unresolved discussion about using the abstraction of verb complementations to discover differences in the semantics of verb valency frames, see **section 2.8.3** (Former Work on page 8). This thesis focuses on an automatic identification of abstractions of verb complementations of verbs to bring insight in the verb valency frames differentiation.

[3] We should still consider the exploitation.

2.8 Verb Valency and This Thesis

2.8.1 Goal Specifications

The goal of this thesis is to introduce a system capable of automatic identification of verb complementations in an unsupervised way. Following steps need to be taken:

- Morphology and analytical analysis of a sentence.
- Identification of verb complementations and their types.
- Identification of a set of valency frames; for each of them a set of semantic types of individual verb complementations must be determined.
- Obtaining an abstraction for each valency frame member.

In another words, we aim at a system that for a given verb identify usages of this verb and put them into semantically homogenous groups based on corpus data. As a prerequisit for this analysis we need tools for morphology analysis, syntactic analysis. For a more advanced system, we might need also deep structure analysis.

After identification of similar sentence examples we try to find the abstraction of verb complementations based on a chosen ontology.

We choose a web-application form to efficiently serve to interested users.

2.8.2 Benefits of the Thesis

The thesis itself investigate cluster analysis usage for verb complementation's semantic preferences abstraction.

The thesis also serves as an investigation for ranking of benefits of manually annotated data. The tool provided with the thesis is capable of various settings for searching similar sentence examples.

As a benefit (next to the tool itself) researchers get a system for testing valency frames disambiguation based on clustering methods. The tool is compatible with CzEng 1.0 data format.

2.8.3 Former Work

In this thesis, we reffer to Jiří Semecký's PhD. thesis [7]. Semecký's thesis introduced unsupervised methods for verb disambiguation.

We also take experience from Karel Vandas' bachelor thesis [9]. The thesis investigated the CIDR [10] and MEAD [11] methods for text summarization. A system is capable of collecting documents with similar topics and their summarization compared to manually created extracts.

For further word sense disambiguation interested reader might look into Eduard Bejček's master thesis [8]. "Two approaches to distinguish word senses have been examined. The first method is derived from synsets from the Czech WordNet, the second one uses valency frames from the PDT-VALLEX dictionary. The goal was to assign appropriate senses (synsets/frames) to ambiguous lemmas.

Machine learning methods were employed, the most notable of which are the decision trees," according to [8].

Further research concerning FrameNet and verb valency might be found in the research of Václava Kettnerová, such as [5].

Chapter 3

Cluster Analysis

3.1 Basics of Clustering

Let's have a set of observations. We would like to discover whether there is any regularity among set members. Let's take all observations, assign a representation (i.e., a vector, where each dimension represents an aspect of observation's behaviour). Let's put all vectors together into a black box called cluster analysis. As an output, we get subsets of observations that share some common features, see Figure 3.1 on page 12.

"Objective of clustering is to put objects (persons, households, transactions, ...) into a number of groups in such a way that objects within the same group are similar, but the groups are dissimilar," according to [16]. "The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering," according to [15]. "Individual groups are called clusters and each of them marks similar behaviour of items that the cluster contains. Items might be placed in clusters exclusively (each item can be placed in exactly one cluster). There are many situations in which an item could reasonably be placed in more than one cluster, and these situations are better addressed by non-exclusive clustering. In the most general case, an overlapping or non-exclusive clustering is used to reflect the fact that an object can simultaneously belong to more than one group," according to [15].

Each of clusters has potentionally a centroid[1] - a point that represents the cluster as a whole. We can define a centroid as a real element included in the cluster or we can reffer to an artifficial point (such as an average over all the cluster dimensions).

Example. Let's have a set of documents. We would like to know whether there is any relation among them. We can take each document, create its vector representation, see **section 3.3** (Language Data Representation for Clustering Methods on page 15) for details, and use this representation as an input for a cluster analysis. As a result, we get documents that share common feature, i.e., documents with the same lexicon that is being used. The resulting set depends on a feature descriptions, on feature weights and also on the setting of cluster method being used.

In clustering, identifying a set of describing features and a number of resulting

[1] **Reminder.** "For data with continuous attributes, the prototype of a cluster is often a centroid, i.e., the average (mean) of all the points in the cluster," according to [15].

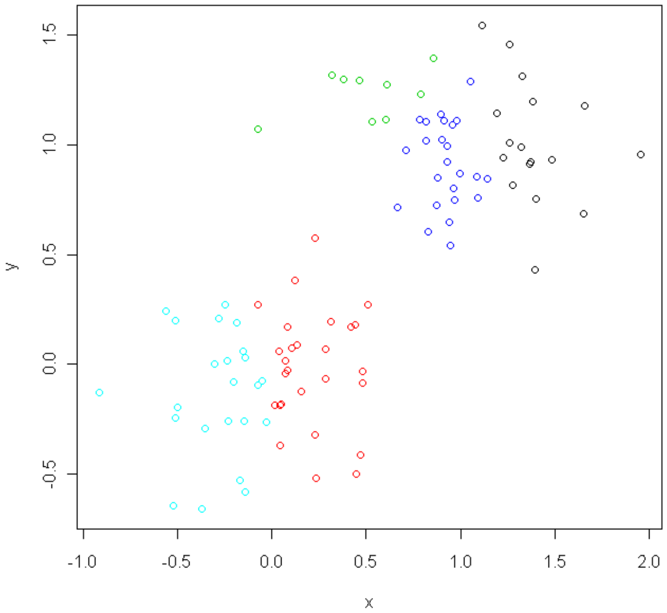


Figure 3.1: Sample diagram showing different group behaviour among data items. To find this regularity and put these items together is an objective of clustering. Here we can see the result of K-means algorithm in statistical tool R.

clusters is essential. For a small data set, setting a relatively high number of resulting clusters can bring as poor results as describing a document by a set of features that is not distinguishing documents at all. One should also consider that an output of a clustering method is just an implicitly expressed similarity of individual elements of a set. Only further analysis of obtained data might reveal their regularities.

3.1.1 Cluster Analysis and the Thesis

In this thesis, we decided to make use of two standard algorithms - Agglomerative Hierarchical Clustering and K-means. We chose these two approaches because of their straightforward implementation and our familiarity and previous experience with them. The thesis implementation includes Agglomerative Hierarchical Clustering, for comparision we use K-means algorithm used with the tool R.

Our motivation is to examine whether these clustering algorithms are suitable for processing small chunks of text with characteristics of verb valency features.

We use as input data sets of sentences - each of them forms a vector. We can view each sentence as a short document as well. For further description how to make a vector representation of the sentence see **section 3.3** (Language Data Representation for Clustering Methods on page 15).

3.2 Cluster Analysis Algorithms

Agglomerative Hierarchical Clustering represents a cluster analysis that builds clusters by merging the most similar items, starting from clusters that contains just one item. K-means on the other hand tries to find by trial the best positions to place cluster centres and iterativelly centres better reflecting the data variety.

These two methods well reflect two different attitudes to clustering and their comparision can be a source for further research. Let's introduce both algorithms.

3.2.1 Agglomerative Hierarchical Clustering

"This approach refers to a collection of closely related clustering techniques that produce a hierarchical clustering by starting with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single, all-encompassing cluster remains," according to [15]. Graphical representation of the process is called dendrogram[2], see at Figure 3.3 on page 14.

The algorithm is summarized below, see the algorithm at Figure 3.2.

```
Basic agglomerative hierarchical clustering algorithm
Compute the proximity matrix, if necessary.
repeat
    Merge the closest two clusters.
    Update the proximity matrix to reflect the proximity between
    the new cluster and the original clusters.
until Only one cluster remains.
```

Figure 3.2: Basic agglomerative hierarchical clustering algorithm, as shown in [15].

The key to the correct cluster analysis, apart from identifying a good set of features, is also a properly chosen metric used for calculating distances of points or for vector sizes.

Standard metric for a vector representation is the euclidean metric. There are several more commonly used metrics described in the Table 3.1.

Standard vector metrics overview	
Euclidean distance	$d_2(a,b) = \sqrt{\sum_i (a_i - b_i)^2}$
Squared euclidean distance	$d_2^2(a,b) = \sum_i (a_i - b_i)^2$
Manhattan distance	$d_1(a,b) = \sum_i \lvert a_i - b_i \rvert$
Maximum distance	$d_\infty(a,b) = \max_i \lvert a_i - b_i \rvert$

Table 3.1: Table shows various metrics for vector representation of items.

For the vector representation we use metrics described in Table 3.1. For the cluster similarity calculation we use metrics summarized in Table 3.2. This metric is used for cluster items as a way to calculate the distance between two clusters. A cluster item in our case is a vector representation of a sentence in a cluster, see **section 3.4** (The tf-idf and the Document Representation on page 16).

[2] A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.

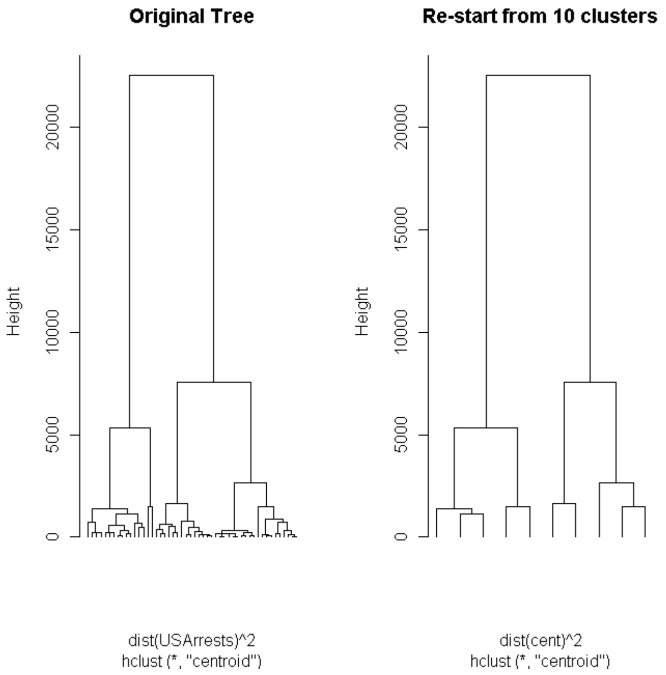


Figure 3.3: Sample dendrogram representing hierarchical process of hierarchical cluster analysis.

Measures between clusters	
Single linkage (nearest neighbour)	$d_{min}(A, B) = min\{d(a, b)\|a \in A, b \in B\}$
Complete linkage (furthest neighbour)	$d_{max}(A, B) = max\{d(a, b)\|a \in A, b \in B\}$

Table 3.2: Table shows various cluster distance measures., as shown in [16]. The function d(a,b) refers to a chosen metric, see Table 3.1.

We also experimented with cosine similarity measure, see [15],

$$\cos\theta = \frac{a * b}{\parallel a \parallel * \parallel b \parallel} .$$

3.2.2 More about cluster analysis

K-means

“This technique attempts to find a user-specified number of clusters (K), which are represented by their centroids,” according to [15].

There are two different methods, K-means and K-medoid. Former one represents cluster by a centroid calculated as a mean of all included vectors, latter looks for the most representative member of a cluster and it is just an approximation of a real centre point. The difference can be described on example of average and median of the number collection.

For further details of K-means algorithm see [19].

Evaluation

Interested readers might read more about clustering evaluation in [15].

3.3 Language Data Representation for Clustering Methods

In this section we introduce attitudes for document clustering techniques. In the thesis we represent each separate sentence as a document. There has been a research about word clustering using engine results based on co-occurence of words, see [20].

A standard way of translating language data into number representation is a $tf - idf$ measure [17]. Let's first define needed prerequisities.

"We would like to compute a score between a query term t and a document d, based on the weight of t in d. The simplest approach is to assign the weight to be equal to the number of occurrences of term t in document d. This weighting scheme is referred to as **term frequency** and is denoted $tf_{t,d}$, with the subscripts the term and the document in order," according to [18].

There are various proposals how to enumerate a term frequency feature. In this thesis, we use natural, logaritm and boolean definition, see Table 3.3.

Term frequency definitions	
n (natural)	$tf_{t,d}$
l (logarithm)	$1 + log(tf_{t,d})$
b (boolean)	$tf_{t,d} = \begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$

Table 3.3: Various term frequency definitions. Table shows proposed natural, logarithm and boolean definition, as shown in [18].

"Denoting as usual the total number of documents in a collection by N, we define the **inverse document frequency** (*idf*) of a term t as follows:

$$idf_t = log\frac{N}{df_t}$$

where we define df_t to be the number of documents in the collection that contain term t," according to [18].

Inverse document frequency value has also more than one specification. Apart from "ignored" one (that ignores the weight of the term across documents and do not penalize frequent ones) we have a natural definition that has been used in this thesis, see Table 3.4. Inverse document frequency penalizes frequently used terms across data collection.

Inverse document frequency definitions	
ignored	1
n (natural)	$log\frac{N}{df_t}$, where N is a size of a collection

Table 3.4: Various inverse document frequency definitions. Table shows proposed "ignored" definition and a natural one.

Example. Let's consider the following example taken from Valeval, set of examples related to Vallex: *Nechranická vodní nádrž byla postavena v letech 1961 až 1968* (*Water dam Nechranice has been built between years 1961 and 1968*).

All lemmas in the sentence are unique, therefore the term frequency (in natural definition) for each lemma is 1. If one of those lemmas would be twice in the sentence, the term frequency of such a lemma would be 2.

Example. Let's consider the same example again. The inverse document frequency for term t is a number of documents in the collection divided by the number of documents containing term t, taking the *log* of the fraction. In this case we have

$$idf_t = log\frac{1}{1} = 0.$$

In case we have N documents in a set, containing term t in just one of them, the definition of *idf* gives us

$$idf_t = log\frac{N}{1}.$$

The highest *idf* of the term t is when the term t is included in just one document in a set.

A combination of the term frequency and the inverse document frequency gives us a *tf-idf* definition.

Definition. The **tf-idf** weighting scheme assigns to term t a weight in document d given by

$$tf\text{-}idf_{t,d} = tf_{t,d} * idf_t.$$

Authors in [18] describe that the *tf-idf* of a term grows with a high frequency of a term in a small subset of considered documents.

3.4 The tf-idf and the Document Representation

The tf-idf measure tries to describe the weight of each dimension of a vector. In case of a document we have a dimension of a vector usually represented by a word. Simple word count accross document set is an irrelevant value. There might be documents with a high-frequency terms, but on the other hand, we can still count terms like conjunctions or punctuation.

The tf-idf measure is the way how to measure quality of words and their importance in a set of documents.

Example. Let's have a hundred of documents. 99 of them are related to music. Just one of them is related to some distant topic, such as a health care. Let's now compare word frequencies of the word *singer* and the word *patient.* Let's assume both words occur in each of documents equal number of times, let's say 5 times. That means we have term frequencies (in the natural definition) $t_{singer} = 5 * 99 = 495$ and $t_{patient} = 5$ over the document set. If we would be using just the term frequency measure, music documents would beat the health care document.

Example. Let's consider the same example with documents. In case of $tf - idf$ measure we take into account the *idf* measure. For music documents, given our previous words, we have (in the natural definition) $t_{singer} = log\frac{100}{99} = 0.01$ and $t_{patient} = log\frac{100}{1} = 4.6$ respectivelly.

If we now take the $tf\text{-}idf$, $tf\text{-}idf_{singer} = 495 * 0.01 = 4.95$ and $tf\text{-}idf_{patient} = 5*4.6 = 23$ respectivelly, we can see the quality measure of a word over a document set. Therefore this measure is highly related to the uniqueness of a word. Unique words usually characterize documents, therefore they should have higher weight in a vector representation.

3.5 Language Features and Clustering

Vector dimensions extraction for sentence or document description serves as an input for further processing and enumerating tf-idf value, see **section 3.4** (The tf-idf and the Document Representation on page 16). We introduce features available in the thesis.

3.5.1 Lemmatization of a Sentence

As a first idea for representation of a sentence might be a representation of its words as dimensions. The benefit of this representation is mainly its simplicity.

On the other hand, in small or not homogeneous data sets this approach can lead to the data sparseness and problems when identifying closest clusters because of uniqueness in a description of each of them. Another difficulty arises from the identification that higher layers of description perform better, as shown in [7].

Motivation. Words are basic elements of a language and sentences with the same verb might also share the vocabulary.

Example. Let's consider our favourite sentence *Nechranická vodní nádrž byla postavena v letech 1961 až 1968* (*Water dam Nechranice has been built between years 1961 and 1968*). Extracting full word information including lemma, form and morphological tag would lead to data sparseness and uniqueness of a vector that would be created. Therefore, for the representation of word layer, we use lemmas only, Table 3.5.

Word	Word representation
Nechranická	nechranický
vodní	vodní
nádrž	nádrž
byla	být
postavena	postavit
v	v
letech	rok
1961	1961
až	až
1968	1968
.	.

Table 3.5: Word layer extraction. All words are lemmatized.

3.5.2 Morphological Analysis

The second idea might be to extract just morphological information of individual words. To prevent data sparseness we decided to use just first positions of part of a speech tag, see **section 5.2.5** (Tag Positions Used on page 28) for setting up the parameter.

Motivation. Word layer granularity might be too high, therefore a usage of less granular description might lead to higher similarity of inspected documents.

Example. Again, please remind our favourite sentence *Nechranická vodní nádrž byla postavena v letech 1961 až 1968* (*Water dam Nechranice has been built between years 1961 and 1968*). Extracting full part of speech tag would lead to data sparseness. Therefore we extracted for this example just first two positions of the morphological tag, see Table 3.6.

Word	Morphology layer representation
Nechranická	AA
vodní	AA
nádrž	NN
byla	Vp
postavena	Vs
v	RR
letech	NN
1961	C=
až	J^
1968	C=
.	Z:

Table 3.6: Morphological tag simplification.

3.5.3 Analytical Layer of a Sentence

The objective of this thesis is verb valency that is a tectogramatical layer feature. Therefore we do not use this layer of description of a sentence.

3.5.4 Tectogramatical Layer of a Sentence

The next idea is to use the deep structure layer. The deep structure captures well the dependencies of words as well as it omits words that are not related to the semantics of a verb.

Motivation. Tectogramatical layer captures semantics as well as verb valency.

Example. Our example again, *Nechranická vodní nádrž byla postavena v letech 1961 až 1968* (*Water dam Nechranice has been built between years 1961 and 1968*). We extract lemmas from verb valency frames only, see Table 3.7.

3.5.5 Tectogramatical Layer of a Sentence - Functor extraction

The next idea is to use the deep structure layer and extract functors from it. The deep structure captures well the dependencies of words as well as it omits words that are not related to the semantics of a verb.

Word	Tectogramatical layer representation
#PersPron	#GEN
Nechranická	-
vodní	-
nádrž	nádrž
byla	-
postavena	postavit
v	-
letech	rok
1961	-
až	-
1968	-
.	-

Table 3.7: Tectogramatical layer extraction.

Word	Tectogramatical layer representation with functors
#PersPron	ACT
Nechranická	-
vodní	-
nádrž	PAT
byla	-
postavena	PRED
v	-
letech	TWHEN
1961	-
až	-
1968	-
.	-

Table 3.8: Tectogramatical layer extraction of functors.

Motivation. Tectogramatical layer captures semantics as well as verb valency.

Example. Our example again, *Nechranická vodní nádrž byla postavena v letech 1961 až 1968* (*Water dam Nechranice has been built between years 1961 and 1968*). We extract lemmas from verb valency frames only, see Table 3.8.

3.5.6 Hyperonymy Extraction

The last idea is to use the lemmas of the sentence and try to find the concepts and abstractions that match using word relation called hyperonymy, the super-subordinate relation.

Motivation. Abstractions or concepts can help identify similar sentences.

Example. Our example again, *Nechranická vodní nádrž byla postavena v letech 1961 až 1968* (*Water dam Nechranice has been built between years 1961 and 1968*). We take words, look for all the senses they have and try to identify all the hypernymy they provide, see Table 3.9.

Word	Czech WordNet Lemmas	Hyperonymy
#PersPron	#GEN	-
Nechranická	-	-
vodní	-	-
nádrž	nádrž-1	putna
byla	-	-
postavena	postavit-1	budovat
v	-	-
letech	rok-1, rok-2	lhůta (rok-1), shromáždění (rok-2)
1961	-	-
až	-	-
1968	-	-
.	-	-

Table 3.9: Hyperonymy extraction. At the case of the word *nádrž* and the word *postavit* is a hyperonymy relation straightforward. The interesting part occurs in the case of the word *rok* with two different senses.

3.5.7 Further Modifications of a Sentence Representation

As a last step of representation might be various modifications of vectors. We might use some stop lists of words that we do not want to put in the vector representation.

Word Stop List

User can define a word stop list to remove words out of considerations of further processing, see **section 5.2.2** (Stop POS List on page 28)

Chapter 4

Data

This chapter introduces data formats we used for the project.

For further experiments we used modified versions of CzEng, Valeval and Czech WordNet, data format examples are shown in **section B.3** (Data Format Examples on page 74).

4.1 CzEng

CzEng is a parallel corpus of the Czech and the English language [24]. It has been used as a resource of the Czech part of it. Introduction of Czeng describes the resource in [24]:

> “CzEng 1.0 contains 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers, see **section 2.2** (Verb Valency and the FGD Formalism on page 4), of a representation.
> ”

Recent article describes more in [14]:

> “CzEng 1.0 is shuffled at the level of “blocks”, sequences of not more than 15 consecutive sentences from one source. The original documents thus cannot be reconstructed but some information about cross-sentence phenomena is preserved. Specifically, CzEng includes Czech and English grammatical and textual co-reference links that do span sentence boundaries. Each “block” comes from one of the text domains (EU Legislation, Fiction, Movie Subtitles, Parallel Web Pages, Technical Documentation, News, Navajo Project) and the domain is indicated in the sentence ID.”

For the purpose of this project we make use only of the Czech part of data and we convert it into an alternative simplified data format capturing only sentence definitions, lemmas with forms, positional tags and valency constituents, are defined and used, see **section 4.4.1** (Tool Data Format on page 23).

4.2 Vallex, Valeval

We used the Vallex valency lexicon and Valeval set of examples at the latest version 2.6, see [23].

Description of Vallex can be found at [23]:

> "The main goal of the Vallex project is to create consistent electronic dictionary rendering underlying structure of Czech verbs and additional syntactico-semantic information useful for the analysis and synthesis of Czech texts as well as other applied tasks in NLP. The Valency Lexicon of Czech Verbs is a collection of linguistically annotated data and documentation, resulting from an attempt at formal description of valency frames of Czech verbs. ... The lexicon provides valency frames with basic syntactico-semantic characterization of the most frequent verbs in their particular senses (number of verb complementations, their morphological forms and obligatoriness), glosses, examples, additional characteristics such as idioms, control, reflexivity, reciprocity, syntactico-semantic class. The lexicon is available in three formats - html version for comfortable browsing and sorting according various criteria, pdf version for printing and xml data for further applications."

Vallex contains a special part - Valeval. Valeval contains over eight thousand manually annotated corpus sentences with respect to the verb frames. Each verb valency frame is accompanied with a context of three sentences. Valeval identifiers (frame identificators) are consistent with Vallex.

This thesis uses Valeval examples as an input to provide an insight into the performance of system with partial valency frame information when making verb complementation abstraction analysis. Valeval data are processed by CzEng workflow and exported into the new data format, see **section 4.4.1** (Tool Data Format on page 23).

4.3 WordNet

As stated in [3]:

> "WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. ... WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity."

Author closes in [3]:

> "The most frequently encoded relation among synsets is the super-subordinate relation (also called **hyperonymy**, hyponymy or ISA relation). It links more general synsets like {furniture, piece_of_furniture} to increasingly specific ones like {bed} and {bunkbed}. Thus, WordNet states that the category furniture includes bed, which in turn includes bunkbed; conversely, concepts like bed and bunkbed make up the category furniture. All noun hierarchies ultimately go up the root node entity. Hyponymy relation is transitive: if an armchair is a kind of chair, and if a chair is a kind of furniture, then an armchair is a kind of furniture."

4.3.1 Czech WordNet

The Czech Wordnet, see [2], origins as a translation of English WordNet. It captures hyperonymy relation as it is introduced in **section 4.3** (WordNet on page 22).

"The Czech WordNet has been developed at the Faculty of Informatics of Masaryk University since 1998. The initial effort was made within the second phase of the EuroWordNet project and the database was further developed when Masaryk University participated as a partner in the BalkaNet project," according to [12].

"The Princeton WordNet contains 152,059 literals organized in 115,424 synsets for a total of 203,145 word-sense pairs. Out of this, 79,689 (69 %) are noun synsets, 13,508 (12 %) are verb synsets, 18,563 (16 %) are adjective synsets and 3664 (3 %) are adverb synsets. ... The Czech WordNet comprises 34,026 literals organized in 28,478 synsets for a total of 47,542 word-sense pairs. Out of this, 21,018 (74 %) are noun synsets, 5162 (18 %) are verb synsets, 2129 (7 %) are adjective synsets and 166 (1 %) are adverb synsets," according to [12].

Problems with Czech WordNet

Original data violated several dogmas of XML well defined documents. The root element of all the document is missing, some of elements did not contain just atomic values but also a mixture of text and markup.

On the top of violation of XML dogmas, some entries contained infinite loop of dependencies, i.e., entries 'stav' and 'podmínka'. Some of entries also contained hyperonymy relation to itself.

For more information how we resolved issues with Czech WordNet see **section B.3.4** (Czech WordNet Issues and Solutions on page 75).

4.4 Data Formats Used for This Thesis

4.4.1 Tool Data Format

For this thesis we decided to use simplified versions of data formats as described in **section 4.1** (CzEng on page 21), **section 4.2** (Vallex, Valeval on page 22), and **section 4.3.1** (Czech WordNet on page 23). For more descriptive information about data format used as input for the tool related to this thesis please consult **section B.3** (Data Format Examples on page 74).

For an example of the tool data format, please see **section B.3.5** (The Tool Data Format. The CzEng Data Format Origin on page 75).

4.4.2 Simplified CzEng Data Format

Original CzEng data format contains much more information than we needed as an input. Therefore we extracted only a surface sentence represetantion, individual words from sentence together with forms, lemmas and tags and also a deep structure tree capturing relations among valency frame members.

For an example of the original CzEng data format, please see **section B.3.1** (The CzEng 1.0 Data Format on page 74). For an example of the tool data format that origins from the CzEng data format, please see **section B.3.5** (The Tool Data Format. The CzEng Data Format Origin on page 75).

4.4.3 Simplified Valeval Data Format

Original Valeval data are not as much descriptive as CzEng data format. Each sentence of Valeval set captures the frame id, context and identify the verb occurence. We decided to process the Valeval data set with the CzEng workflow to get the same input data format and then we simplified this data format as described in **section 4.4.1** (Tool Data Format on page 23).

For an example of the Valeval Data Format, please see **section B.3.3** (The Valeval Data Format on page 75). For an example of the tool data format used for Valeval, please see **section B.3.5** (The Tool Data Format. The CzEng Data Format Origin on page 75).

4.4.4 Simplified WordNet Data Format

Also Czech WordNet data format has been simplified. We extracted only needed hyperonymy relation to make the data format simpler and faster to load.

For an example of the WordNet Data Format, please see **section B.3.4** (The WordNet Data Format on page 75). For an example of the tool data format that origins from the WordNet Data Format, please see **section B.3.5** (The Tool Data Format. The WordNet Data Format Origin on page 75).

Chapter 5

Experiments

5.1 Overview of the Steps Taken

5.1.1 Step One. Verb Selection

Our intention is to choose low frequency verbs, high frequency verbs and verbs with a lots of different valency frames and verbs with just a few valency frames entries. Obvious limitations of data resources caused that infrequent words had just a few entries in the whole data set, so we decided to abandon the idea of low frequency words. On the other hand, frequent words tended to be modal words or words with a huge diversity in valency frames members. That is why we decided to only consider the aspect of a number of different valency frames. For more details, see **section 6.1** (Data Preparation on page 33).

5.1.2 Step Two. Automatic Analysis

We used CzEng 1.0 data release, see [24], as an input for our experiments. A set of these data is converted into the new data format, see **section 4.4.1** (Tool Data Format on page 23) and **chapter B** (Really Technical Details on page 73). As described in data section, CzEng data format contains morphological, analytical and tectogramatical layers of description, see for details in [22]. We extracted the information for sentence descriptions - each sentence is described as a n-dimensional vector represented by a combination of features introduced in **section 3.5** (Language Features and Clustering on page 17). As an input we use the CzEng data format, see **section B.3.1** (The CzEng 1.0 Data Format on page 74), we convert it into the Tool data format, see **section 4.4.1** (Tool Data Format on page 23) and as an output we get vectors of documents (formed from sentences), see **section 3.4** (The tf-idf and the Document Representation on page 16).

5.1.3 Step Three. Clustering of Verb Examples

For cluster analysis of various data inputs we use different settings, see **section 5.3** (Experimental Settings on page 30) for more details.

Clustering of Valeval Examples

Clustering of Valeval examples individually is straightforward - we simply place all the sentences from a discovered frame entry into a cluster.

Clustering of Valeval Examples with Mixed Frames

Motivation. Comparision of Mixed frames and separated frames of a valency lexicon can tell whether the manual annotation helps to identify extra information. We are interested in a question whether the verb valency frames can be distinguished if we mix them and try to cluster them back together. For this purpose all the sentences from Valeval are put together and then processed by the same procedure as in **section 5.1.3** (Clustering of Raw CzEng Data on page 26).

Clustering of Raw CzEng Data

As an unsupervised method we try to cluster all the sentences that contain the chosen verb. In this case we have no information about frames and their relations.

5.1.4 Step Four. Abstraction

We make use of WordNet, see [3], to identify hyperonymy relations among words. For each word we found a branch of words that describes the word in a hyperonymy relation tree and then by merging branches together for each functor individually we got final result - an abstraction of a complementation, see the algorithm at Figure 5.1.

Algorithm discussion

The algorithm is quite straightforward. The Czech WordNet data definition forms multiple trees, therefore looking for a particular hypernymy relation branch is just traversing the tree to the root. The only hard part is to identify the synset of the word and it is set externally.

The algorithm has constant memory complexity and it runs in linear time (in the length of the branch from the word to the root of the tree).

```
Abstraction algorithm
Select words at position of certain functor.
repeat for each one of words
    Identify desired synset of the word
    Extract hyperonymy branch starting from word
    going up the ontology tree.
    Add the resulted branch of words into the list of sets
of generalised words.
until No word has left.
Count occurences of words in the list of sets.
Normalize the count by size of words array.
```

Figure 5.1: Use case of the abstraction algorithm.

Difficulties

There are three problems that are connected with this attitude applied to WordNet.

The first difficulty corresponds to the data definition. Some of entries of the WordNet do not correspond to the entries of analysed words as we get it from CzEng, i.e., the word *atrakce* with sense “1” from the WordNet might not equal to *atrakce* (*show*) with sense “1” from the morphological analyser. This is solved by the user who choose the proper meaning of a word itself by selecting in the tool.

Another difficulty lies in no proposed desired granularity or ranking of abstraction level. Therefore there is no automatic way how to decide whether the proposed abstraction fits the purpose. And it is again up to the user to select the proper abstraction suitable for their need.

The last difficulty points to the fact that sometimes there are non-corresponding entries put together in the WordNet. This difficulty is caused by the translation from English to Czech.

Verb Complementations Types Level of Abstraction

There are two extremes that might be introduced by generalisation of verb complementations.

The first extreme is to make the generalisation on the level of individual words - the granularity of such generation is too high and the use is rather difficult. The other extreme is to make a generalisation to such a general semantic preference that it fits any word.

We tried to move in between these two borders and introduced a visualisation of results in a way that anyone can decide whether it is preffered to make higher or lower generalisation of a semantic preference of a complementation.

Example. Let’s abstract the word *atrakce* (*show*). We get four abstraction levels all with the same value. But, in interaction with other words some of those abstractions rank higher or lower, that means that some of abstractions present better solution than others, i.e., abstraction *vztah* (*a relationship*) ranks higher (it is more often to meet a word with this abstraction, such as *love*, *hate*) than with abstraction *zábavný pořad* (*a tv show*).

5.2 Parameters to be Set and Their Effects on Data Analysis

We have decided to run several experiments. Each of them is characterised by unique parameters settings. These settings can be found at **chapter 6** (Evaluation on page 33).

5.2.1 Use the Whole Sentence

The parameter sets whether to use the whole available sentence or just valency frame members.

5.2.2 Stop POS List

The parameter sets part of speeches removed from considerations for further processing.

Motivation. Some of certain types of part of speech might add noise to data when processing them. The setting allow user to choose from basic part of speech types, punctuation and unknown type (morphologically not recognized).

5.2.3 Depth of the Node under the Verb

The parameter sets what is the desired maximal depth for considered valency frame members starting from the verb.

5.2.4 Sentence Features

The parameter defines the information extracted from the sentence when creating a sentence representation.

Motivation. Let's remind various features described in **section 3.3** (Language Data Representation for Clustering Methods on page 15). Setting feature extraction to word layer of a sentence, see **section 3.5.1** (Lemmatization of a Sentence on page 17), might cluster together sentences with similar vocabulary. Setting feature extraction to morphology layer of a sentence, see **section 3.5.2** (Morphological Analysis on page 17), might cluster together documents that share similar morphology structure. The setting of feature extraction to tectogramatical layer of a sentence, see **section 3.5.4** (Tectogramatical Layer of a Sentence on page 18), might cluster together only documents concerning similar topics and deep structure. Setting the functors identified to consider, see **section 3.5.5** (Tectogramatical Layer of a Sentence - Functor extraction on page 18), might cluster together the same deep layer structures. The last concern, looking for the matching abstraction definitions, see **section 3.5.6** (Hyperonymy Extraction on page 19), might help to identify and cluster together the same classes of verbs.

For sentence feature descriptions and examples, see **section 3.5** (Language Features and Clustering on page 17).

5.2.5 Tag Positions Used

The parameter sets the number of used positions of positional part of speech tag of words of a sentence.

Motivation. The more position taken from the tag, the more granular vector set is created.

Together with **Sentence features**, the tag positions taken from the part of speech tag influence the result of cluster analysis.

5.2.6 Remove Predicate

The parameter defines whether the predicate word is removed from the sentence before the vector creation.

Motivation. Some sentences might only cause the similarity of each other based on the predicate that share. Therefore user can choose to remove the predicate to prevent this situation, remind **section 3.3** (Language Data Representation for Clustering Methods on page 15).

5.2.7 Vector Metric

The parameter sets the metric in which the vector's attributes are calculated.

Motivation. Different vector metrics, see Table 3.1, influence the vector sizes needed for calculating vector similarities, see **section 3.2** (Cluster Analysis Algorithms on page 12).

5.2.8 Term Frequency

The parameter sets the definition of term frequency used for the processing.

Motivation. Let's remind various definitions of term frequency, see **section 3.3** (Language Data Representation for Clustering Methods on page 15). There are three definitions mentioned. The first definition (natural) only takes into account discovered term frequency in its original form. This form is prefered if the term frequency is balanced - there is the same number level of it. The second definition (logarithm) tries to make certain equality of exponentially different values.

The third definition (boolean) makes all the values of term frequency to be equal.

Term frequency refers to a type of value extracted from the input. This value is based on any textual data provided by Sentence features, see **section 5.2.4** (Sentence Features on page 28).

5.2.9 Cluster Creation Strategy

The parameter sets the strategy for clusters creation either to the maximal similarity or to the minimal similarity.

Motivation. The motivation of merging minimal similarity might be to leave maximum number of similar clusters and get rid of clusters with zero mutual similarity.

Similarity Treshold

The parameter sets the numerical border under which the cluster process is stopped (when **Merge below treshold** parameter is set to true).

Motivation. It makes sense to merge only clusters that enhance their inner information and do not cause noise in the data.

Merge Below Treshold

The parameter sets whether the **Similarity Treshold**, see **section 5.2.9** (Similarity Treshold on page 29), is used and cluster analysis is stopped when reaching the border similarity.

5.2.10 Number of Clusters

The parameter sets the number of clusters to be created. This parameter is ignored in case the **Similarity Treshold** parameter is reached and the **Merge below treshold** parameter is set to false.

Motivation. Definining certain number of clusters gives the user the power to stop the cluster analysis process and investigate discovered data sets.

5.3 Experimental Settings

We have experimented with many various settings of the tool to get in our opinion results that can be used for further investigation. The results can only be recieved by cluster analysis resulting in only several clusters containing majority of items. As an outcome we propose following settings of the tool that are further elaborated in **6 Evaluation** chapter on page 33.

5.3.1 Settings One

We have decided to use the default setting of a tool, see Table 5.1. The motivation for this setting is to identify similar sentences based on their hyperonymy features, nominative-accusative feature and analytical and deep feature. We aimed at the maximal deep structure information to distinguish valency frames.

Setting	Value
Use the Whole Sentence	true
Stop POST List	all except noun
Depth of the Node under the Verb	4
Sentence Features used	HNAD
Analytical Positions Used	2
Remove Predicate	true
Vector Metric	euclidean
Term Frequency	boolean
Cluster Creation Strategy	max
Similarity Treshold	0.0
Similarity Metric	cosine similarity
Merge Below Treshold	false
Number of clusters	5

Table 5.1: Experimental setting no. 1.

5.3.2 Settings Two

Setting two was chosen after applying the setting one. The idea is to make larger clusters.

Setting	Value
Use the Whole Sentence	true
Stop POST List	all except noun
Depth of the Node under the Verb	6
Sentence Features used	LA
Analytical Positions Used	5
Remove Predicate	true
Vector Metric	euclidean
Term Frequency	boolean
Cluster Creation Strategy	max
Similarity Treshold	0.0
Similarity Metric	euclidean distance
Merge Below Treshold	false
Number of clusters	5

Table 5.2: Experimental setting no. 2.

Chapter 6

Evaluation

6.1 Data Preparation

There are various aspects to consider when doing an evaluation. Data selection is definitelly one of them.

We decided to choose a verb *postavit* based on considerations given in **section 5.1.1** (Step One. Verb Selection on page 25).

Verb *postavit* has desired variety in meaning, it has enough verb valency frames included in Valeval and a sufficient number of examples. For the purpose of the thesis it also gives nice demonstrative examples.

6.2 Evaluation Procedure

For evaluation we chose several aspects to consider that touch all the steps of process of choosing semantic preferences for valency complementations of a verb.

6.2.1 Evaluation of CzEng Output for Distinguished Examples of Valeval

The first important aspect, in case of using separated frames of Valeval examples, is the performance of CzEng workflow applied on the selected data. For this concern we manually checked number of frames that are well recognized by CzEng workflow from manually annotated Valeval set. We took into account the fact that analysing a sentence on the tectogramatical layer is a hard task, therefore we only check for a presence of obligatory functors in valency frames of selected verbs.

The task asks whether an element in a set of Valeval examples is well recognized or not. We use *Recall*

$$Recall = \frac{|relevant \cap retrieved|}{|total\ relevant|}$$

, where as *relevant* we treat all the entries[1], as *retrieved* we treat entries that have been identified as well recognized. In our case the value *total relevant* is equal to the *relevant* value.

[1] We simply aim at the full set by the *retrieved* value.

6.2.2 Discussion of Results of Cluster Analysis

In this section of evaluation we mention settings that influence the results of the clustering, we mention possible reasons that conclude to numbers we get. We also compare results of Agglomerative Hierarchical Algorithm and K-means algorithm for analysis with known valency frames, known but mixed valency frames from Valeval and unknown valency frames from CzEng.

6.2.3 Evaluation of Abstraction Analysis

The Last evaluation section discusses results of abstraction analysis. We consider several aspects as data input, influence of CzEng analysis, cluster analysis. We also note the influence of input data from WordNet.

We use following formulas for abstraction ranking,

$$\textit{Abstraction level}: \text{ functor} \times \text{word } \rightarrow [0,1],$$

$$\textit{Abstraction level(functor, abstraction)} = \frac{\sum_{word \in words} \textit{Abstraction(word, abstraction)}}{|words|},$$

where $functor$ is one of functors investigated, $words$ are all tectogramatical lemmas with functor $functor$ in a current set of functors extracted from a desired set of sentences and $\mathit{Abstraction}(\mathit{word}, \mathit{abstraction})$ is a function,

$$\textit{Abstraction}: \text{ word} \times \text{word } \rightarrow \{0,1\},$$

$$\textit{Abstraction(word, abstraction)} = \begin{cases} 1 & \text{if } \mathit{abstraction} \text{ is a hyperonymy of } \mathit{word} \\ 0 & \text{otherwise} \end{cases}$$

We also define a baseline for this measure, function $\mathit{baseline}: \text{functors} \rightarrow [0,1]$,

$$\textit{baseline(functor)} = \begin{cases} \frac{1}{|words|} & \text{if } |words| \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

where *words* are all tectogramatical lemmas with functor *functor* in a current set of sentences.

We present both analysis results in percentage. As an abbreviation we use notation $A_{word}(words)$ for describing the level of abstraction for a set *words* by the word *word*.

6.3 Verb *postavit*

We have chosen the verb *postavit* because of the nice variety of data and also for its different readings.

We will inspect the CzEng data input accuracy with respect to the valency frame identification, we will inspect two various settings for cluster analysis and we will discuss results for usual functors **ACT** and **PAT**.

Feel free to inspect more verb complementations yourself.

6.3.1 Evaluation of the CzEng Output for *postavit*

The Valeval data set refers to examples of five valency frames of verb *postavit*[2]. The accuracy of correct identification of valency frame after the application of the CzEng workflow is shown in Table 6.1.

Frame Id	# Examples	# Recognized	%
1	37	25	48.08 %
2	2	2	3.85 %
5	3	0	0 %
6	6	1	1.92 %
7	4	4	7.69 %
Total	52	32	61.54 %

Table 6.1: Table shows the accuracy of verb valency frame identification by CzEng workflow (number of cases frame members were identified correctly). The comparision was made with the data taken from Valeval.

As Table 6.1 shows, accuracy of correct identification of valency frames varies. Some valency frame identification fails constantly; i.e., valency frame with id 5 is not recognized even once. With further analysis of results we discovered that instead of EFF functor DIR3 functor is used in all cases of automatic analysis.

The most of valency frame identification fails on too complex sentences, where the structure is incorrectly parsed.

6.3.2 Evaluation of Setting One with Vallex data source with a defined frame #1

In this case we have a defined valency frame therefore we do not run the cluster analysis. We chose on purpose the biggest cluster as we do in other analysis of Setting one.

ACT functor. The highest abstraction is a word *skupina*, the baseline is quite low because of just a few actors capable of abstraction.

If we look closer to the sets that are abstracted, abstraction *skupina* and *společenská skupina* are based on the same words. The same case sets are also a base for abstractions *forma života* and *jednotlivec*. In this case user is the one deciding which type of abstraction is more suitable.

PAT functor. The highest abstraction in this case reaches *objekt* with abstraction level of 53.19%. Then *výrobek*, *konstrukce* and *skupina* continues.

The baseline of the PAT functor is 2.127%, so the abstraction outperform the baseline significantly.

As the reason for such a high result we should mention that the valency frame examples in this case were manually annotated and even if the language analysis was processed by an automatic tool, we still get high portion of originally well-annotated information.

Let's now inspect sets that are the source for abstraction. For the highest result of PAT functor extra words *kazeta-1, silnice-1, voda-2, kus-1, nádrž-1* are used. The second abstraction by word *výrobek* does not include these extra words.

[2] Although there is twelve valency frames distinguished in Vallex.

Interesting abstraction occurs in case of C_{PAT4}, where instead of a subset of previous set we have an abstraction refering to *skupina*.

Therefore abstractions *objekt, výrobek* and *konstrukce* share the same valency frame and *skupina* belongs to a different one.

This means that the frame definition might not be correct and would need further specification.

Further information can be found at Table 6.2, Table 6.3.

Abstractions %	Baseline	#	Words
$A_{skupina}(C_{ACT1}) = 57.14\%$ $A_{společenská\ skupina}(C_{ACT2}) = 57.14\%$ $A_{forma\ života}(C_{ACT3}) = 42.85\%$ $A_{jednotlivec}(C_{ACT4}) = 42.85\%$ $A_{politické\ zřízení}(C_{ACT5}) = 28.57\%$	14.28 %	7	režim-1 , architekt-1 , muž-1 , společnost-1 , kancléř-1 , podnik-1 , vláda-1
$A_{objekt}(C_{PAT1}) = 53.19\%$ $A_{výrobek}(C_{PAT2}) = 42.55\%$ $A_{konstrukce}(C_{PAT3}) = 31.91\%$ $A_{skupina}(C_{PAT4}) = 12.76\%$ $A_{budova}(C_{PAT5}) = 10.63\%$ $A_{ubytování}(C_{PAT6}) = 10.63\%$ $A_{společenská\ skupina}(C_{PAT7}) = 8.510\%$ $A_{forma\ života}(C_{PAT8}) = 6.382\%$ $A_{změna}(C_{PAT9}) = 4.255\%$ $A_{čin}(C_{PAT10}) = 4.255\%$	2.127 %	47	středisko-1 , budova-1 , čistírna-1 , kazeta-1 , byt-1 , lanovka-1 , silnice-1 , hotel-1 , procento-1 , ubytovna-1 , účel-2 , přebytek-1 , školka-1 , prostředek-1 , člověk-1 , město-1 , srub-1 , stádo-1 , projekt-1 , řada-1 , kinematografie-1 , elektrárna-1 , nepřátelství-1 , dům-1 , smrt-1 , mlékárna-1 , čerpadlo-1 , továrna-1 , voda-2 , blok-1 , byt-1 , kus-1 , obchod-1 , tiskárna-1 , nádrž-2 , chrám-1 , kasárna-1 , hora-1 , narození-1 , dům-1 , krematorium-1 , návštěvník-1 , mládě-1 , mrakodrap-1 , podnik-1 , Evropa-1 , kříž-1

Table 6.2: Table showing results for known valency frame with Setting one. Synsets of words are mentioned to give a full description. The sets compared further in the thesis.

6.3.3 Evaluation of Setting One with Valeval mixed valency frames

Clustering in the case of setting 1, **section 5.3.1** (Settings One on page 30) gives quite interesting results. First of all, most of the sentences are not clustered together and they are forming separate clusters of size one. The total number of

Set Id	Words
C_{ACT1}	{ režim-1, společnost-1, podnik-1, vláda-1 }
C_{ACT2}	{ režim-1, společnost-1, podnik-1, vláda-1 }
C_{ACT3}	{ architekt-1, muž-1, kancléř-1 }
C_{ACT4}	{ architekt-1, muž-1, kancléř-1 }
C_{ACT5}	{ režim-1, vláda-1 }
C_{PAT1}	{ středisko-1, budova-1, čistírna-1, kazeta-1, byt-1, silnice-1, hotel-1, ubytovna-1, prostředek-1, srub-1, elektrárna-1, dům-1, čerpadlo-1, továrna-1, voda-2, byt-1, kus-1, tiskárna-1, nádrž-2, chrám-1, kasárna-1, dům-1, krematorium-1, mrakodrap-1, kříž-1 }
C_{PAT2}	{ středisko-1, budova-1, čistírna-1, byt-1, hotel-1, ubytovna-1, prostředek-1, srub-1, elektrárna-1, dům-1, čerpadlo-1, továrna-1, byt-1, tiskárna-1, chrám-1, kasárna-1, dům-1, krematorium-1, mrakodrap-1, kříž-1 }
C_{PAT3}	{ středisko-1, budova-1, čistírna-1, byt-1, hotel-1, ubytovna-1, srub-1, dům-1, továrna-1, byt-1, chrám-1, kasárna-1, dům-1, krematorium-1, mrakodrap-1 }
C_{PAT4}	{ lanovka-1, školka-1, stádo-1, řada-1, blok-1, podnik-1 }
C_{PAT5}	{ středisko-1, hotel-1, chrám-1, krematorium-1, mrakodrap-1 }
C_{PAT6}	{ byt-1, ubytovna-1, srub-1, byt-1, kasárna-1 }
C_{PAT7}	{ lanovka-1, školka-1, blok-1, podnik-1 }
C_{PAT8}	{ člověk-1, návštěvník-1, mládě-1 }
C_{PAT9}	{ smrt-1, narození-1 }
C_{PAT10}	{ kinematografie-1, obchod-1 }

Table 6.3: Table showing sets for the current experiment with verb postavit.

clusters is 22, except clusters with just one sentence there were clusters with 2 (3 occurences), 3 (2 occurences), 4, 6 and 16 (1 occurence) items. More clusters then 5 desired are found because of the setting not to merge clusters if the similarity of a merge reaches zero.

For the evaluation we chose clusters with size at least 2. The biggest cluster contains 16 sentences.

The output of basic K-means algorithm in R (when asking the 22 resulting clusters) on the same data set resulted in one big cluster containing 21 sentences and 21 other clusters containing just one sentence.

ACT functor. Interesting thing occured in the case of ACT functor - is seems abstractions are equally well to a defined valency frame, therefore the considered cluster contains similar sentences. On the other hand, the abstraction words are equal to the previous case.

PAT functor. The highest abstraction level reaches 61.9% with *objekt* abstraction. Next abstraction, reaching 52.38%, states that the PAT functor semantic preference is *výrobek*. In this case, words that abstracted to this level of abstraction are {*středisko, dům, byt, hotel, ubytovna, dům, srub, krematorium, prostředek, kasárna, byt*}. Apart from *konstrukce* with 47.61%, there is a high abstraction level for the word *ubytování* (23.8%).

Nine of sixteen of sentences is about constructing of buildings. The abstraction can be still improved by choosing the word *řada* in its 4th synset, then we reach

66.66% of coverage of input words. For details about this experiment you can inspect Table 6.4, Table 6.5.

The most interesting abstraction is with C_{PAT4} refering to *ubytování* that has skipped *skupina* and *budova*. This observation is probably based on a fact that clustering has identified more sentences with type *ubytování* together than in a valency frame defined case.

Baseline for **PAT** functor is 4.761%. Therefore proposed abstraction offers significant improvement.

Abstractions %	Baseline	#	Words
$A_{skupina}(C_{ACT1}) = 50.0\%$ $A_{forma\ života}(C_{ACT2}) = 50.0\%$ $A_{společenská\ skupina}(C_{ACT3}) = 50.0\%$ $A_{jednotlivec}(C_{ACT4}) = 50.0\%$	16.66 %	6	architekt-1 , režim-1 , muž-1 , kancléř-1 , společnost-1 , firma-1
$A_{objekt}(C_{PAT1}) = 61.90\%$ $A_{výrobek}(C_{PAT2}) = 52.38\%$ $A_{konstrukce}(C_{PAT3}) = 47.61\%$ $A_{ubytování}(C_{PAT4}) = 23.80\%$ $A_{skupina}(C_{PAT5}) = 14.28\%$ $A_{budova}(C_{PAT6}) = 14.28\%$ $A_{společenská\ skupina}(C_{PAT7}) = 9.523\%$ $A_{ubikace}(C_{PAT8}) = 9.523\%$	4.761 %	21	dům-1 , procento-1 , kasárna-1 , ubytovna-1 , středisko-1 , nepřátelství-1 , výrobce-1 , kinematografie-1 , prostředek-1 , kus-1 , blok-1 , nádrž-2 , projekt-1 , krematorium-1 , byt-1 , školka-1 , hotel-1 , řada-1 , dům-1 , byt-1 , srub-1

Table 6.4: Valeval mixed frame experiment with Setting One.

Set Id	Words
C_{ACT1}	{ režim-1, společnost-1, firma-1 }
C_{ACT2}	{ architekt-1, muž-1, kancléř-1 }
C_{ACT3}	{ režim-1, společnost-1, firma-1 }
C_{ACT4}	{ architekt-1, muž-1, kancléř-1 }
C_{PAT1}	{ dům-1, kasárna-1, ubytovna-1, středisko-1, prostředek-1, kus-1, nádrž-2, krematorium-1, byt-1, hotel-1, dům-1, byt-1, srub-1 }
C_{PAT2}	{ dům-1, kasárna-1, ubytovna-1, středisko-1, prostředek-1, krematorium-1, byt-1, hotel-1, dům-1, byt-1, srub-1 }
C_{PAT3}	{ dům-1, kasárna-1, ubytovna-1, středisko-1, krematorium-1, byt-1, hotel-1, dům-1, byt-1, srub-1 }
C_{PAT4}	{ kasárna-1, ubytovna-1, byt-1, byt-1, srub-1 }
C_{PAT5}	{ blok-1, školka-1, řada-1 }
C_{PAT6}	{ středisko-1, krematorium-1, hotel-1 }
C_{PAT7}	{ blok-1, školka-1 }
C_{PAT8}	{ kasárna-1, ubytovna-1 }

Table 6.5: Table showing sets for the current experiment with verb postavit.

6.3.4 Evaluation of Setting One with CzEng data source

Clustering in the case of setting 1, **section 5.3.1** (Settings One on page 30) gives different output. There are 32 resulting clusters, most of them with just one sentence. There are only 3 resulting clusters with 2 items and only 1 resulting cluster containing 4 items and one cluster containing 6 items.

We chose the cluster with the biggest size that gives the best abstraction result. In this case result of K-means resulted in one cluster of size 11, one cluster of size 2 and the rest were clusters with size one. Further inspection of results given by R might be interesting.

ACT functor. In this case we have found no abstraction better than a baseline. This is because of the fact that there is no common ontology member that would conclude to a subset of {*spousta-3 , impérium-1 , hotel-1*}.

PAT functor. The highest abstraction in this case reaches 40% with abstraction *vůz, letoun, plavidlo, dopravní letadlo.*

In this case the baseline, also because of the size of the cluster, is quite high, 20%. Therefore the level of abstraction does not gain much new information.

In case of different selection of synsets, abstraction in the case of choose synset *lod̆-3* results surprisingly in 60.0% with words { *hotel-1, lod̆-3, lod̆-3* } into abstraction *výrobek*. This means that sometimes better abstractions can be find if another synsets are investigated.

Detailed results are shown in Table 6.6, Table 6.7, Table 6.8, Table 6.9.

Abstractions %	Baseline	#	Words
	33.33 %	3	spousta-3 , impérium-1 , hotel-1
$A_{vůz}(C_{PAT1}) = 40.0\%$ $A_{letoun}(C_{PAT2}) = 40.0\%$ $A_{plavidlo}(C_{PAT3}) = 40.0\%$ $A_{dopravní\ letadlo}(C_{PAT4}) = 40.0\%$	20.0 %	5	hotel-1 , akademie-1 , oprava-1 , lod̆-1 , lod̆-1

Table 6.6: CzEng data output for the biggest cluster.

Set Id	Words
C_{PAT1}	{ lod̆-1, lod̆-1 }
C_{PAT2}	{ lod̆-1, lod̆-1 }
C_{PAT3}	{ lod̆-1, lod̆-1 }
C_{PAT4}	{ lod̆-1, lod̆-1 }

Table 6.7: Table showing sets for the current experiment with verb postavit.

6.3.5 Discussion of Setting One and verb *postavit*

The crucial observation of the comparision of **section 6.3.2** (Evaluation of Setting One with Vallex data source with a defined frame #1 on page 35), **section 6.3.3** (Evaluation of Setting One with Valeval mixed valency frames on page 36), **section 6.3.4** (Evaluation of Setting One with CzEng data source on page 39) is the influence of clustering technique.

Abstractions %	Baseline	#	Words
	33.33 %	3	impérium-1 , hotel-1 , spousta-3
$A_{výrobek}(C_{PAT1}) = 60.0\%$ $A_{objekt}(C_{PAT2}) = 60.0\%$ $A_{konstrukce}(C_{PAT3}) = 60.0\%$	20.0 %	5	oprava-1 , akademie-1 , hotel-1 , loď-3 , loď-3

Table 6.8: CzEng data output for the biggest cluster with a synset loď-3.

Set Id	Words
C_{PAT1}	{ hotel-1, loď-3, loď-3 }
C_{PAT2}	{ hotel-1, loď-3, loď-3 }
C_{PAT3}	{ hotel-1, loď-3, loď-3 }

Table 6.9: Table showing sets for the current experiment with verb postavit with a synset loď-3.

The size of resulting clusters influence the baseline value and therefore for clusters with many items baseline value decreases and the abstraction level gains on importance.

Even if the best abstraction levels were comparable, by the influence of baseline defined valency frame experiment in **section 6.3.2** (Evaluation of Setting One with Vallex data source with a defined frame #1 on page 35) beats the rest because the sample for the abstraction of a defined valency frame was much higher than in case of Valeval data and CzEng output.

In case of ACT functor we did not recieve interesting results and in one case we did not get any abstraction at all. Therefore we can't really state that ACT functor of the word postavit has any specialized abstraction.

In case of PAT functor we got abstraction that is interesting (if we leave out abstraction *objekt*). We can abstract PAT functor by words {*výrobek, konstrukce, ubytování*}.

Next setting motivation. The result of CzEng data should be evaluated when larger clusters are obtained.

6.3.6 Evaluation of Setting Two with Vallex data source with a defined frame #1

In this case we have a defined valency frame therefore we do not run the cluster analysis. We chose on purpose the biggest cluster as we do in other analysis of Setting two, see **section 5.3.2** (Settings Two on page 30).

The verb valency frame is defined, therefore cluster analysis is skipped and clusters are formed from given examples.

ACT functor. Identified abstractions are equal to the experiment with a setting one.

PAT functor. We have discovered that the abstractions are equally ordered to the previous Setting one, see **section 5.3.1** (Settings One on page 30), used with the same valency frame with a bit less abstraction level value.

The baseline in this case is 2%. Further information can be found at Table

6.10, Table 6.11.

Abstractions %	Baseline	#	Words
$A_{skupina}(C_{ACT1}) = 50.0\%$ $A_{společenská\ skupina}(C_{ACT2}) = 50.0\%$ $A_{forma\ života}(C_{ACT3}) = 37.5\%$ $A_{jednotlivec}(C_{ACT4}) = 37.5\%$ $A_{politické\ zřízení}(C_{ACT5}) = 25.0\%$	12.5 %	8	společnost-1 , muž-1 , architekt-1 , krev-1 , vláda-1 , podnik-1 , kancléř-1 , režim-1
$A_{objekt}(C_{PAT1}) = 50.0\%$ $A_{výrobek}(C_{PAT2}) = 40.0\%$ $A_{konstrukce}(C_{PAT3}) = 30.0\%$ $A_{skupina}(C_{PAT4}) = 16.0\%$ $A_{společenská\ skupina}(C_{PAT5}) = 10.0\%$ $A_{budova}(C_{PAT6}) = 10.0\%$ $A_{ubytování}(C_{PAT7}) = 10.0\%$ $A_{forma\ života}(C_{PAT8}) = 6.0\%$ $A_{atribut}(C_{PAT9}) = 6.0\%$ $A_{vlastnost}(C_{PAT10}) = 4.0\%$	2.0 %	50	město-1 , kříž-1 , narození-1 , elektrárna-1 , dům-1 , krematorium-1 , byt-1 , návštěvník-1 , národ-1 , kus-1 , hotel-1 , tiskárna-1 , lanovka-1 , silnice-1 , Evropa-1 , podnik-1 , projekt-1 , člověk-1 , mrakodrap-1 , čerpadlo-1 , mládě-1 , blok-1 , mlékárna-1 , chrám-1 , přebytek-1 , kasárna-1 , hora-1 , byt-1 , školka-1 , voda-2 , společenstvo-1 , prostředek-1 , budova-1 , kazeta-1 , obchod-1 , účel-2 , dům-1 , továrna-1 , srub-1 , nádrž-2 , středisko-1 , řada-1 , ubytovna-1 , kinematografie-1 , smrt-1 , stádo-1 , funkce-1 , čistírna-1 , nepřátelství-1 , procento-1

Table 6.10: Table showing results for known valency frame with Setting two.

6.3.7 Evaluation of Setting Two with Valeval mixed valency frames

We made the cluster analysis and as desired we got all the most similar items grouped together. The rest of items was placed one by one in the rest of clusters.

We compare results of this section to **section 6.3.3** (Evaluation of Setting One with Valeval mixed valency frames on page 36).

ACT functor. There is more words that are identified in an abstraction type as given in the previous experiment with *postavit*.

PAT functor. In this experiment the baseline has decreased to 1.639% and also the order of abstractions differ in 4th position (instead of *ubytování* we have *skupina*).

Set Id	Words
C_{ACT1}	{ společnost-1, vláda-1, podnik-1, režim-1 }
C_{ACT2}	{ společnost-1, vláda-1, podnik-1, režim-1 }
C_{ACT3}	{ muž-1, architekt-1, kancléř-1 }
C_{ACT4}	{ muž-1, architekt-1, kancléř-1 }
C_{ACT5}	{ vláda-1, režim-1 }
C_{PAT1}	{ kříž-1, elektrárna-1, dům-1, krematorium-1, byt-1, kus-1, hotel-1, tiskárna-1, silnice-1, mrakodrap-1, čerpadlo-1, chrám-1, kasárna-1, byt-1, voda-2, prostředek-1, budova-1, kazeta-1, dům-1, továrna-1, srub-1, nádrž-2, středisko-1, ubytovna-1, čistírna-1 }
C_{PAT2}	{ kříž-1, elektrárna-1, dům-1, krematorium-1, byt-1, hotel-1, tiskárna-1, mrakodrap-1, čerpadlo-1, chrám-1, kasárna-1, byt-1, prostředek-1, budova-1, dům-1, továrna-1, srub-1, středisko-1, ubytovna-1, čistírna-1 }
C_{PAT3}	{ dům-1, krematorium-1, byt-1, hotel-1, mrakodrap-1, chrám-1, kasárna-1, byt-1, budova-1, dům-1, továrna-1, srub-1, středisko-1, ubytovna-1, čistírna-1 }
C_{PAT4}	{ národ-1, lanovka-1, podnik-1, blok-1, školka-1, společenstvo-1, řada-1, stádo-1 }
C_{PAT5}	{ lanovka-1, podnik-1, blok-1, školka-1, společenstvo-1 }
C_{PAT6}	{ krematorium-1, hotel-1, mrakodrap-1, chrám-1, středisko-1 }
C_{PAT7}	{ byt-1, kasárna-1, byt-1, srub-1, ubytovna-1 }
C_{PAT8}	{ návštěvník-1, člověk-1, mládě-1 }
C_{PAT9}	{ přebytek-1, účel-2, funkce-1 }
C_{PAT10}	{ účel-2, funkce-1 }

Table 6.11: Table showing sets for the current experiment with verb postavit.

Also the type of merges is only restricted to analytical information of words. Details of this experiment is summarized in Table 6.12, Table 6.13.

6.3.8 Evaluation of Setting Two with CzEng data source

We make cluster analysis and all the items except four are clustered together again. This is what we desired because now we can inspect abstraction over all the dataset.

K-means algorithm in this case gives almost identical result - it clusters 38 items into one clusters and create a cluster with two items. The rest of clusters is one item by a cluster.

ACT functor. Interesting abstraction results in {*případ, jednání, čin*}. The result is pointing to a fact that positions of actors differ significantly from the rest of experiments in this section.

PAT functor. The result shows the usual first three positions - *objekt*, *výrobek*, *konstrukce.*

Baseline is 2.3%. Details see in Table 6.14, Table 6.15.

6.3.9 Discussion of Setting Two and verb *postavit*

The crucial observation of the comparision of **section 6.3.6** (Evaluation of Setting Two with Vallex data source with a defined frame #1 on page 40), **section 6.3.7** (Evaluation of Setting Two with Valeval mixed valency frames on page 41), **section 6.3.8** (Evaluation of Setting Two with CzEng data source on page 42) is that obviously clustering does not really involve the order of levels of abstraction, it only influence the result based on abstraction level we defined.

The overall observation is that clustering probably does not help much when processing abstraction.

6.4 Verb *odpovídat*

We have decided to evaluate one more word with, in our opinion, more successful and interesting setting from the perspective of clustering. We chose setting one, **section 5.3.1** (Settings One on page 30) and sample of Vallex known frame and Valeval data input. The word *odpovídat* has been chosen because of its interesting distinctions that can be observed with relativelly different valency frame complementations.

We demonstrate the Setting one on the defined valency frames and valency frames identified by clustering from Valeval. We also compare the verb to the previous verb *postavit* with setting one.

6.4.1 Evaluation of Setting One with Vallex data source with a defined frame #4

We chose a frame no. 4 with 60 entries (the biggest valency frame).

ACT functor. The ACT functor results in an abstraction *atribut*, followed by *objekt* and *vztah*.

Sets C_{ACT1}, C_{ACT2}, C_{ACT3} shows that the semantic preference is derived from different sets.

The baseline is 2.22%, therefore three significant ACT functor abstractions have been identified.

PAT functor. PAT functor differs - it abstracts into *atribut*, *vědění* and *čin*.

Also the PAT position shows various abstractions for various subsets of all functors considered.

Based on our observations, a valency frame might be further specialised based on identified types of different semantic preference.

See details in table Table 6.16, Table 6.17.

6.4.2 Evaluation of Setting One with Valeval mixed valency frames

Clustering in the case of setting 1, **section 5.3.1** (Settings One on page 30) does not cluster most of the sentences together. The total number of clusters is 28, except clusters with just one sentence there were clusters with 2 (2 occurences) and 68 (1 occurence) items. More clusters then 5 desired are found because of the setting not to merge clusters if the similarity of a merge reaches zero.

The biggest cluster contains 68 sentences.

ACT functor. Compared to Table 6.16, ACT results also in *jednotlivec*.

PAT functor Is very similar to the previous experiment. Notable difference is in the abstraction *vědění* that has been identified in the previous experiment.

We also demonstrate different functor positions that are identified.

For details see Table 6.18, Table 6.19.

6.4.3 Discussion of Setting One and verb *odpovídat*

The most interesting comparision of these two experiments are in the abstracted words that are used for the abstraction. Although semantic preferences are very similar, maybe investigating these words might help to identify distinctions in valency frames.

6.4.4 Discussion of Setting One and the verb *odpovídat* and the verb *postavit*

Verb *postavit* results in a better value of abstraction than the verb *odpovídat*. The reason probably is that the distinction of complementations of verb differ more in the case of a verb *postavit*. Also, the abstraction of a verb *postavit* is based on more words than word *odpovídat*.

From the semantic preference point of view, the word *postavit* seems that distinctions in semantics of the verb *postavit* are also more easily identified than in the case of the verb *odpovídat*.

Abstractions %	Baseline	#	Words
$A_{\textit{společenská skupina}}(C_{ACT1}) = 46.15\%$ $A_{\textit{skupina}}(C_{ACT2}) = 46.15\%$ $A_{\textit{jednotlivec}}(C_{ACT3}) = 38.46\%$ $A_{\textit{forma života}}(C_{ACT4}) = 38.46\%$ $A_{\textit{politické zřízení}}(C_{ACT5}) = 15.38\%$ $A_{\textit{předák}}(C_{ACT6}) = 15.38\%$ $A_{\textit{objekt}}(C_{ACT7}) = 15.38\%$	7.692 %	13	úřad-1 , podnik-1 , škola-2 , kancléř-1 , firma-1 , vláda-1 , režim-1 , muž-1 , trenér-1 , host-1 , architekt-1 , společnost-1 , krev-1
$A_{\textit{objekt}}(C_{PAT1}) = 37.70\%$ $A_{\textit{výrobek}}(C_{PAT2}) = 27.86\%$ $A_{\textit{konstrukce}}(C_{PAT3}) = 19.67\%$ $A_{\textit{skupina}}(C_{PAT4}) = 16.39\%$ $A_{\textit{forma života}}(C_{PAT5}) = 11.47\%$ $A_{\textit{společenská skupina}}(C_{PAT6}) = 11.47\%$ $A_{\textit{jednotlivec}}(C_{PAT7}) = 8.196\%$ $A_{\textit{čin}}(C_{PAT8}) = 8.196\%$ $A_{\textit{budova}}(C_{PAT9}) = 8.196\%$ $A_{\textit{ubytování}}(C_{PAT10}) = 6.557\%$	1.639 %	61	funkce-1 , kus-1 , rybník-1 , voda-2 , ubytovna-1 , mládě-1 , továrna-1 , nováček-1 , hráč-1 , novinka-1 , byt-1 , podnik-1 , krematorium-1 , duel-1 , byt-1 , kazeta-1 , putna-1 , kasárna-1 , lanovka-1 , produkce-1 , pravidlo-1 , čistírna-1 , procento-1 , kříž-1 , dům-1 , obchod-1 , přebytek-1 , čerpadlo-1 , chrám-1 , prostředek-1 , básník-1 , účel-2 , blok-1 , nepřátelství-1 , hotel-1 , hora-1 , výrobce-1 , činnost-1 , středisko-1 , město-1 , člověk-1 , hrábě-1 , projekt-1 , tiskárna-1 , návštěvník-1 , řada-1 , smrt-1 , společenstvo-1 , stádo-1 , kinematografie-1 , úřad-1 , školka-1 , narození-1 , elektrárna-1 , Evropa-1 , ministerstvo-1 , mlékárna-1 , národ-1 , soustava-1 , silnice-1 , mrakodrap-1

Table 6.12: Experiments with Setting two and the mixed frames from Valeval.

Set Id	Words
C_{ACT1}	{ úřad-1, podnik-1, firma-1, vláda-1, režim-1, společnost-1 }
C_{ACT2}	{ úřad-1, podnik-1, firma-1, vláda-1, režim-1, společnost-1 }
C_{ACT3}	{ kancléř-1, muž-1, trenér-1, host-1, architekt-1 }
C_{ACT4}	{ kancléř-1, muž-1, trenér-1, host-1, architekt-1 }
C_{ACT5}	{ vláda-1, režim-1 }
C_{ACT6}	{ kancléř-1, trenér-1 }
C_{ACT7}	{ škola-2, krev-1 }
C_{PAT1}	{ kus-1, voda-2, ubytovna-1, továrna-1, byt-1, krematorium-1, byt-1, kazeta-1, putna-1, kasárna-1, čistírna-1, kříž-1, dům-1, čerpadlo-1, chrám-1, prostředek-1, hotel-1, středisko-1, hrábě-1, tiskárna-1, elektrárna-1, silnice-1, mrakodrap-1 }
C_{PAT2}	{ ubytovna-1, továrna-1, byt-1, krematorium-1, byt-1, kasárna-1, čistírna-1, kříž-1, dům-1, čerpadlo-1, chrám-1, prostředek-1, hotel-1, středisko-1, tiskárna-1, elektrárna-1, mrakodrap-1 }
C_{PAT3}	{ ubytovna-1, továrna-1, byt-1, krematorium-1, byt-1, kasárna-1, čistírna-1, dům-1, chrám-1, hotel-1, středisko-1, mrakodrap-1 }
C_{PAT4}	{ podnik-1, lanovka-1, blok-1, řada-1, společenstvo-1, stádo-1, úřad-1, školka-1, ministerstvo-1, národ-1 }
C_{PAT5}	{ mládě-1, nováček-1, hráč-1, básník-1, výrobce-1, člověk-1, návštěvník-1 }
C_{PAT6}	{ podnik-1, lanovka-1, blok-1, společenstvo-1, úřad-1, školka-1, ministerstvo-1 }
C_{PAT7}	{ nováček-1, hráč-1, básník-1, výrobce-1, návštěvník-1 }
C_{PAT8}	{ duel-1, produkce-1, obchod-1, činnost-1, kinematografie-1 }
C_{PAT9}	{ krematorium-1, chrám-1, hotel-1, středisko-1, mrakodrap-1 }
C_{PAT10}	{ ubytovna-1, byt-1, byt-1, kasárna-1 }

Table 6.13: Table showing sets for the current experiment with verb postavit.

Abstractions %	Baseline	#	Words
$A_{případ}(C_{ACT1}) = 28.57\%$ $A_{jednání}(C_{ACT2}) = 28.57\%$ $A_{čin}(C_{ACT3}) = 28.57\%$ $A_{podnik}(C_{ACT4}) = 28.57\%$	14.28 %	7	nehoda-1 , spousta-3 , impérium-1 , uskutečnění-1 , porucha-2 , opatření-1 , hotel-1
$A_{objekt}(C_{PAT1}) = 46.51\%$ $A_{výrobek}(C_{PAT2}) = 32.55\%$ $A_{konstrukce}(C_{PAT3}) = 20.93\%$ $A_{čin}(C_{PAT4}) = 11.62\%$ $A_{atribut}(C_{PAT5}) = 11.62\%$ $A_{budova}(C_{PAT6}) = 9.302\%$ $A_{letoun}(C_{PAT7}) = 6.976\%$ $A_{skupina}(C_{PAT8}) = 6.976\%$ $A_{plavidlo}(C_{PAT9}) = 6.976\%$ $A_{dopravní\ letadlo}(C_{PAT10}) = 6.976\%$	2.325 %	43	služba-1 , otázka-1 , zisk-1 , známost-1 , nemocnice-1 , slovo-1 , předpoklad-1 , voda-2 , rampa-1 , dvůr-2 , odpověď-1 , město-1 , dráha-1 , vor-1 , zařízení-1 , akademie-1 , rychlost-1 , oprava-1 , počet-1 , dům-1 , většina-1 , úkryt-1 , dům-1 , opatření-1 , postel-1 , židle-1 , hotel-1 , loď-1 , věž-1 , náklad-1 , klec-1 , okno-1 , tlak-1 , nemocnice-1 , materiál-1 , loď-1 , pec-1 , hlava-1 , hranice-1 , otec-1 , obydlí-1 , škola-2 , rám-1

Table 6.14: Experiment with Setting two with CzEng data input.

Set Id	Words
C_{ACT1}	{ nehoda-1, porucha-2 }
C_{ACT2}	{ uskutečnění-1, opatření-1 }
C_{ACT3}	{ uskutečnění-1, opatření-1 }
C_{ACT4}	{ nehoda-1, porucha-2 }
C_{PAT1}	{ nemocnice-1, voda-2, rampa-1, dráha-1, zařízení-1, dům-1, dům-1, postel-1, židle-1, hotel-1, věž-1, náklad-1, klec-1, okno-1, nemocnice-1, materiál-1, pec-1, obydlí-1, škola-2, rám-1 }
C_{PAT2}	{ nemocnice-1, rampa-1, dráha-1, dům-1, dům-1, postel-1, židle-1, hotel-1, věž-1, klec-1, nemocnice-1, obydlí-1, škola-2, rám-1 }
C_{PAT3}	{ nemocnice-1, dům-1, dům-1, hotel-1, věž-1, nemocnice-1, obydlí-1, škola-2, rám-1 }
C_{PAT4}	{ služba-1, otázka-1, předpoklad-1, oprava-1, opatření-1 }
C_{PAT5}	{ zisk-1, známost-1, rychlost-1, počet-1, většina-1 }
C_{PAT6}	{ nemocnice-1, hotel-1, nemocnice-1, škola-2 }
C_{PAT7}	{ vor-1, loď-1, loď-1 }
C_{PAT8}	{ dvůr-2, akademie-1, hranice-1 }
C_{PAT9}	{ vor-1, loď-1, loď-1 }
C_{PAT10}	{ vor-1, loď-1, loď-1 }

Table 6.15: Table showing sets for the current experiment with verb postavit.

Abstractions %	Baseline	#	Words
$A_{atribut}(C_{ACT1}) = 22.22\%$ $A_{objekt}(C_{ACT2}) = 15.55\%$ $A_{vztah}(C_{ACT3}) = 11.11\%$ $A_{komunikace}(C_{ACT4}) = 11.11\%$ $A_{vlastnost}(C_{ACT5}) = 8.888\%$ $A_{\textit{forma života}}(C_{ACT6}) = 8.888\%$ $A_{\textit{výrobek}}(C_{ACT7}) = 8.888\%$ $A_{jednotlivec}(C_{ACT8}) = 6.666\%$ $A_{podnik}(C_{ACT9}) = 6.666\%$ $A_{\textit{psaný jazyk}}(C_{ACT10}) = 6.666\%$	2.222 %	45	cena-1 , proud-1 , bod-1 , tabulka-1 , host-1 , tvorba-1 , ředitel-1 , velikost-1 , představa-1 , dopis-1 , otvor-1 , popis-1 , napětí-1 , skladba-1 , vláda-1 , počet-1 , úhel-1 , věta-1 , funkce-1 , poloha-1 , kompetence-1 , průběh-3 , soustava-1 , sazba-1 , kurs-2 , růst-1 , teplota-1 , střelec-1 , člověk-1 , smlouva-1 , odchylka-1 , realita-1 , forma-1 , složka-1 , výsledek-1 , smlouva-1 , prostředí-1 , prostředek-1 , úroveň-1 , suma-1 , rozměr-1 , vzdálenost-1 , cena-1 , výrobek-1 , část-1
$A_{atribut}(C_{PAT1}) = 19.69\%$ $A_{\textit{vědění}}(C_{PAT2}) = 12.12\%$ $A_{\textit{čin}}(C_{PAT3}) = 10.60\%$ $A_{vztah}(C_{PAT4}) = 10.60\%$ $A_{objekt}(C_{PAT5}) = 10.60\%$ $A_{idea}(C_{PAT6}) = 9.090\%$ $A_{komunikace}(C_{PAT7}) = 9.090\%$ $A_{kvantita}(C_{PAT8}) = 7.575\%$ $A_{\textit{výrobek}}(C_{PAT9}) = 6.060\%$ $A_{\textit{úkaz}}(C_{PAT10}) = 6.060\%$	1.515 %	66	příjmení-1 , cena-1 , úroveň-1 , poslání-1 , stres-1 , množství-1 , dolar-1 , obraz-1 , vodič-2 , hlas-1 , situace-1 , inzerát-1 , mzda-1 , hlas-1 , hodnota-1 , charakter-1 , norma-1 , kategorie-2 , pojetí-1 , styl-1 , scéna-1 , význam-1 , bod-1 , den-1 , průměr-1 , vývoj-1 , princip-1 , úroveň-1 , vztah-1 , import-1 , norma-1 , osud-1 , teplota-1 , nadprodukce-1 , doba-1 , prostředí-1 , síla-1 , představa-1 , přídavek-1 , ekonomika-1 , předpis-1 , měřítko-1 , příznak-1 , velikost-1 , činnost-1 , služba-1 , objem-1 , rovina-1 , skutečnost-1 , růst-1 , výsledek-1 , požadavek-1 , středisko-1 , ...

Table 6.16: Current experiment with verb odpovídat.

Set Id	Words
C_{ACT1}	{ cena-1, velikost-1, počet-1, funkce-1, poloha-1, kompetence-1, teplota-1, úroveň-1, vzdálenost-1, cena-1 }
C_{ACT2}	{ otvor-1, skladba-1, složka-1, prostředek-1, rozměr-1, výrobek-1, část-1 }
C_{ACT3}	{ dopis-1, popis-1, věta-1, smlouva-1, smlouva-1 }
C_{ACT4}	{ dopis-1, popis-1, věta-1, smlouva-1, smlouva-1 }
C_{ACT5}	{ cena-1, funkce-1, kompetence-1, cena-1 }
C_{ACT6}	{ host-1, ředitel-1, střelec-1, člověk-1 }
C_{ACT7}	{ skladba-1, prostředek-1, rozměr-1, výrobek-1 }
C_{ACT8}	{ host-1, ředitel-1, střelec-1 }
C_{ACT9}	{ proud-1, odchylka-1, výsledek-1 }
C_{ACT10}	{ dopis-1, smlouva-1, smlouva-1 }
C_{PAT1}	{ cena-1, úroveň-1, hlas-1, hlas-1, hodnota-1, charakter-1, průměr-1, úroveň-1, teplota-1, velikost-1, poloha-1, poloha-1, přízeň-1 }
C_{PAT2}	{ pojetí-1, význam-1, bod-1, princip-1, představa-1, předpis-1, cíl-1, standard-1 }
C_{PAT3}	{ poslání-1, import-1, nadprodukce-1, činnost-1, služba-1, požadavek-1, příval-1 }
C_{PAT4}	{ příjmení-1, inzerát-1, norma-1, styl-1, norma-1, příznak-1, věta-1 }
C_{PAT5}	{ obraz-1, scéna-1, přídavek-1, měřítko-1, středisko-1, přehrávač-1, voda-2 }
C_{PAT6}	{ pojetí-1, význam-1, bod-1, princip-1, předpis-1, standard-1 }
C_{PAT7}	{ inzerát-1, norma-1, styl-1, norma-1, příznak-1, věta-1 }
C_{PAT8}	{ dolar-1, den-1, doba-1, objem-1, věk-1 }
C_{PAT9}	{ obraz-1, měřítko-1, středisko-1, přehrávač-1 }
C_{PAT10}	{ vývoj-1, síla-1, růst-1, růst-1 }

Table 6.17: Table showing sets for the current experiment with verb odpovídat.

Abstractions %	Baseline	#	Words
$A_{\text{forma života}}(C_{ACT1}) = 20.45\%$ $A_{\text{atribut}}(C_{ACT2}) = 20.45\%$ $A_{\text{jednotlivec}}(C_{ACT3}) = 18.18\%$ $A_{\text{objekt}}(C_{ACT4}) = 13.63\%$ $A_{\text{vlastnost}}(C_{ACT5}) = 9.090\%$ $A_{\text{čin}}(C_{ACT6}) = 9.090\%$ $A_{\text{výrobek}}(C_{ACT7}) = 9.090\%$ $A_{\text{vztah}}(C_{ACT8}) = 6.818\%$ $A_{\text{komunikace}}(C_{ACT9}) = 6.818\%$ $A_{\text{psaný jazyk}}(C_{ACT10}) = 6.818\%$	2.272 %	44	část-1 , dopis-1 , ředitel-1 , suma-1 , představa-1 , prostředek-1 , host-1 , prostředí-1 , výsledek-1 , růst-1 , velikost-1 , žena-1 , průběh-3 , muž-1 , ruka-1 , reprezentant-1 , cena-1 , výrobek-1 , realita-1 , odchylka-1 , teplota-1 , vláda-1 , zástupce-1 , bod-1 , kompetence-1 , tajemník-1 , smlouva-1 , podpora-1 , člověk-1 , soustava-1 , poloha-1 , tým-1 , úroveň-1 , skladba-1 , cena-1 , složka-1 , funkce-1 , průzkum-1 , tvorba-1 , smlouva-1 , rozměr-1 , zvědavec-1 , počet-1 , úhel-1
$A_{\text{atribut}}(C_{PAT1}) = 17.14\%$ $A_{\text{čin}}(C_{PAT2}) = 15.71\%$ $A_{\text{vztah}}(C_{PAT3}) = 12.85\%$ $A_{\text{komunikace}}(C_{PAT4}) = 12.85\%$ $A_{\text{vědění}}(C_{PAT5}) = 11.42\%$ $A_{\text{objekt}}(C_{PAT6}) = 10.0\%$ $A_{\text{idea}}(C_{PAT7}) = 7.142\%$ $A_{\text{vlastnost}}(C_{PAT8}) = 5.714\%$ $A_{\text{předmět}}(C_{PAT9}) = 5.714\%$ $A_{\text{kvantita}}(C_{PAT10}) = 5.714\%$	1.428 %	70	den-1 , norma-1 , příval-1 , množství-1 , pojetí-1 , část-1 , dílo-1 , přídavek-1 , služba-1 , požadavek-1 , úroveň-1 , dolar-1 , poloha-1 , obraz-1 , prostředí-1 , princip-1 , bolest-1 , podstata-2 , penízek-1 , výhrůžka-1 , kategorie-2 , hlas-1 , činnost-1 , majetek-1 , mzda-1 , voda-2 , poloha-1 , dialog-1 , ekonomika-1 , norma-1 , síla-1 , představa-1 , příměří-1 , růst-1 , problém-1 , výsledek-1 , cena-1 , návrh-1 , osud-1 , reportér-1 , úroveň-1 , doba-1 , hlas-1 , nadprodukce-1 , otázka-1 , příznak-1 , poslání-1 , růst-1 , cíl-1 , otázka-1 , předpis-1 , standard-1 ...

Table 6.18: Abstractions of ACT and PAT functors in Valeval mixed frame experiment of Setting one.

Set Id	Words
C_{ACT1}	{ ředitel-1, host-1, žena-1, muž-1, reprezentant-1, zástupce-1, tajemník-1, člověk-1, zvědavec-1 }
C_{ACT2}	{ velikost-1, cena-1, teplota-1, kompetence-1, poloha-1, úroveň-1, cena-1, funkce-1, počet-1 }
C_{ACT3}	{ ředitel-1, host-1, žena-1, muž-1, reprezentant-1, zástupce-1, tajemník-1, zvědavec-1 }
C_{ACT4}	{ část-1, prostředek-1, výrobek-1, skladba-1, složka-1, rozměr-1 }
C_{ACT5}	{ cena-1, kompetence-1, cena-1, funkce-1 }
C_{ACT6}	{ průběh-3, podpora-1, průzkum-1, tvorba-1 }
C_{ACT7}	{ prostředek-1, výrobek-1, skladba-1, rozměr-1 }
C_{ACT8}	{ dopis-1, smlouva-1, smlouva-1 }
C_{ACT9}	{ dopis-1, smlouva-1, smlouva-1 }
C_{ACT10}	{ dopis-1, smlouva-1, smlouva-1 }
C_{PAT1}	{ úroveň-1, poloha-1, podstata-2, hlas-1, poloha-1, cena-1, úroveň-1, hlas-1, teplota-1, přízeň-1, hodnota-1, velikost-1 }
C_{PAT2}	{ příval-1, služba-1, požadavek-1, činnost-1, nadprodukce-1, otázka-1, poslání-1, otázka-1, import-1, závazek-1, otázka-1 }
C_{PAT3}	{ norma-1, výhrůžka-1, dialog-1, norma-1, problém-1, návrh-1, příznak-1, problém-1, dopis-1 }
C_{PAT4}	{ norma-1, výhrůžka-1, dialog-1, norma-1, problém-1, návrh-1, příznak-1, problém-1, dopis-1 }
C_{PAT5}	{ pojetí-1, princip-1, bolest-1, představa-1, cíl-1, předpis-1, standard-1, bod-1 }
C_{PAT6}	{ část-1, dílo-1, přídavek-1, obraz-1, voda-2, středisko-1, přehrávač-1 }
C_{PAT7}	{ pojetí-1, princip-1, předpis-1, standard-1, bod-1 }
C_{PAT8}	{ podstata-2, cena-1, přízeň-1, hodnota-1 }
C_{PAT9}	{ výhrůžka-1, problém-1, návrh-1, problém-1 }
C_{PAT10}	{ den-1, dolar-1, doba-1, věk-1 }
C_{MEANS1}	{ povinnost-1, úkon-1 }
C_{MEANS2}	{ rozsah-1, měsíc-1 }
C_{MEANS3}	{ neochota-1, většina-1 }
C_{RSTR1}	{ známý-1, pan-1 }
C_{TWHEN1}	{ rok-1, věk-1 }
C_{LOC1}	{ postel-1, základ-1 }
C_{LOC2}	{ redakce-1, konference-1 }
C_{LOC5}	{ resort-1, zájem-1 }
C_{APP1}	{ příběh-1, řízení-1, ochrana-1, zápas-1, akce-2, rozdělení-1, poptávka-1 }
C_{APP2}	{ řízení-1, ochrana-1, akce-2, poptávka-1 }
C_{APP3}	{ novinář-1, žena-1, zaměstnanec-1 }
C_{APP5}	{ síla-1, plamen-1 }
C_{APP7}	{ kancelář-1, vybavení-1 }
C_{APP8}	{ nabídka-1, dokument-1 }

Table 6.19: Table showing sets for the current experiment with verb odpovídat. We also demonstrate inputs for another functors than just ACT and PAT.

Chapter 7

Discussion

7.1 Data Input

The CzEng toolchain performs quite better than expected. The expected problem was that the output of CzEng toolchain serves as an input of the thesis, therefore the low accuracy could influence final results significantly.

One of reasons why the overall accuracy for abstraction identification is that maybe the analysis of correct verb valency frame is not that important.

7.1.1 Cluster Analysis

Clustering as a method for valency frame discovery without gentle settings tends to fail. In our opinion, there is several reasons why this behaviour is observered.

- **Cluster analysis suffer from data sparseness.** This is probably the crucial observation. Even with several introduced features overall performance of clustering did not improve much.

- **Missing context of clustered data make clustering process too complex.** From the definition of the CzEng 1.0 data release, all the sentences are mixed. This aspect of data builds another barrier for using clustering as a method for the valency frame identification.

- **Data sparseness cause low similarity of merged clusters.** Forcing clusters to merge together bring often noise in the data and the capabilities of abstraction in the cluster is rather low. For merging words as pronouns or prepositions are used.

7.2 Abstraction identification

As the last step the abstraction depends on information provided from the input part as well as it depends on the clustering part. For abstraction another issues rises because of using selected Czech WordNet ontology:

- **Translation of Czech WordNet from English version often provides suspicious results.** Mapping of the correct meaning of the word from one language to another shows as problematic. As an example we can refer to the usage of word *atrakce*.

- **Insufficient coverage of corpus words.** Another considerable issue lies in discovery that there is relativelly low coverage of common words by Czech WordNet. Abstraction process is therefore heavilly influenced by low data input because of this consideration.

- **Missing abstraction granularity prevents from correct abstraction level identification.** Data of the Czech WordNet does not contain information about particular level of abstractions that might be used. Some levels of abstractions are simply too general to bring any sufficient infromation.

  Example. Let's refer to words such as *forma života* (*living being*) or *objekt* (*an object*). Sometimes the only abstraction for a word is very generic and therefore does not bring any new information. On the other hand, sometimes there is many hyperonymy words between the word and the most general abstraction of it. It is hard to decide which abstraction to use and not to bring too general and useless information.

- **Incompatibility of word synsets.** Another difficulty lies in incompatibility of word synsets. Numbering of synsets differ from the synset codes of ÚFAL's tools.

  Example. Let's demonstrate this aspect on the word *měsíc* - in case of ÚFAL's synset definition we have *month* (*měsíc*), in case of the Czech WordNet description we have *the Moon* (*měsíc*).

7.2.1 Aspect of Manual Categorisation of Data

As one of the results of the thesis is a statement that manually annotated information bring significant improvement of the abstraction task. In our case we mean only semantic categorisation into valency frames.

Abstraction process executed on manually separated verb valency frames outperform mixed version of verb valency frames as well as CzEng data input source.

Chapter 8

Conclusion

We studied behaviour of valency complementations of verbs and its automatic identification of semantic preferences of these complementations.

For that purpose we implemented a clustering method with different types of settings, see **chapter A** (The User Guide on page 63) and **chapter B** (Really Technical Details on page 73), studied the method and made experiments (see **chapter 5** (Experiments on page 25)) and made final thesis describing the theory (see **chapter 2** (Verb Valency on page 3), **chapter 3** (Cluster Analysis on page 11)) and the practise of experiments (again see **chapter 5** (Experiments on page 25)) including the evaluation (see **chapter 6** (Evaluation on page 33)).

We demonstrated the abstraction analysis on verb *postavit* with two different types of settings.

We found out, see **6 Evaluation** chapter on page 33 and **7 Discussion** chapter on page 53, certain regularities and limited capabilities of abstraction of valency complementation types. Clustering as a method for the purpose of automatic verb valency determination fails on sentences because of data sparseness and problems in syntactical and morphological analysis. It also seems that abstraction itself can skip the step of clustering itself without loosing its performance.

As a side-effect of the project we have experienced CzEng data format, data preparation, especially XSL transformations and XML processing by Java. We also worked in the environment of web framework Spring using Eclipse, SVN and Tex for writing final thesis. We have also developed a web-based tool capable of extending to using different methods for automatic semantic preferences identification.

Further perspectives. One of further perspectives might be research on abstraction without using clustering. Further research of semantic preferences of verbs might help to identify new sub-valency entries.

List of Figures

List of Tables

Bibliography

[1] Urešová, Z.; Štěpánek, J.; Hajič, J. (2007). *PDT-Vallex 2.0.*

[2] Pala, K.; Čapek, T.; Zajíčková, B.; Bartůšková, D.; Kulková, K.; Hlaváčková, D.; Hoffmannová, P.; Bejček, E.; Straňák, P.; Hajič, J. (2010). *Czech WordNet 1.9 PDT.*

[3] Fellbaum, C. (2005). *WordNet and wordnets.* In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

[4] Hlaváčková, D; Horák, A. (2006). *VerbaLex - New Comprehensive Lexicon of Verb Valencies for Czech.* In Computer Treatment of Slavic and East European Languages. Bratislava, Slovakia.

[5] Kettnerová, V.; Lopatková, M.; Bejček, E. (2012). *Mapping Semantic Information from FrameNet onto VALLEX.* The Prague Bulletin of Mathematical Linguistics 97, 23-41.

[6] Žabokrtský, Z. (2005). *Valency Lexicon of Czech Verbs.* Doctoral thesis, Faculty of Mathematics and Physics, Charles University in Prague.

[7] Semecký, J. (2007). *Verb Valency Frames Disambiguation.* PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague.

[8] Bejček E. (2006). *Automatické přiřazování významu - "Sense-tagging".* Master thesis, Faculty of Mathematics and Physics, Charles University in Prague.

[9] Vandas K. (2009). *Methods of Text Summarization.* Bachelor thesis, Faculty of Mathematics and Physics, Charles University in Prague.

[10] Radev, Dragomir R.; Hatzivassilouglou, V.; McKeown, K. R. (1999). *A Description of the CIDR System as Used for TDT-2.* Department of Computer Science, Columbia University.

[11] Radev, Dragomir R., et al. (2004). *Centroid-based Summarization of Multiple documents..* In Crestani Fabio (ed.). Information Processing and Management. ELSEVIER.

[12] Blahuš M. (2011). *Extending Czech WordNet Using a Bilingual Dictionary.* Master thesis, Faculty of Informatics, Masaryk University in Brno.

[13] Matthews, P. H. (1997). *Concise Oxford Dictionary of Linguistics.*

[14] Bojar, O.; Žabokrtský, Z., et al. (2012). *The Joy of Parallelism with CzEng 1.0. Proceedings of LREC2012.* ELRA. Istanbul, Turkey.

[15] Tan, P.-N.; Steinbach, M. and Kumar, V. (2005). *Introduction to Data Mining-Book.* Chapter 8 - Cluster Analysis:Basic Concepts and Algorithms, First Edition, Pearson-Addison Wesley Higher Education publishers, pages:532-568.

[16] Feelders, A. (2011). *Lectures of Advanced Data Mining.*

[17] Salton G; McGill MJ (1986). *Introduction to modern information retrieval.*

[18] Manning, Christopher D. (2008). *An introduction to information retrieval.* Cambridge University Press.

[19] Hartigan, J. A. and Wong, M. A. (1979). *A K-means clustering algorithm.* Applied Statistics 28, 100–108.

[20] Matsuo, Y.; Sakaki, T.; Uchiyama, K.; Ishizuka, M. (2006). *Graph-based Word Clustering using a Web Search Engine.* EMNLP.

[21] Chomsky, N. (1961). *Some methodological remarks on Generative grammar.*

[22] Sgall J. (1967). *Generativní popis jazyka a česká deklinace.*

[23] Lopatková, M.; Žabokrtský, Z.; Kettnerová, V.; Skwarska, K.; Bejček, E.; Hrstková, K.; Nová, M.; Tichý, M. (2007). *VALLEX 2.5 - Valency Lexicon of Czech Verbs, version 2.5.*

[24] Bojar, O.; Žabokrtský, Z.; Dušek, O.; Galuščáková, P.; Majliš, M.; Mareček, D.; Maršík, J.; Novák, M; Popel, M.; Tamchyna, A. (2011). *CzEng 1.0.*

Appendix A

The User Guide

A.1 Installation

A.1.1 Software and Tool Requirements

A user installation requires the environment with java JRE 1.5+. Before installation you need to install Tomcat web server (here we demonstrate the installation on Tomcat 7.0.27). For development installation, see **section B.1** (Development installation for Unix Systems on page 73).

A.1.2 Installation of the CD Content

The installation CD contains the tool, data for running the tool, source code of the tool and this thesis. For more information, consult **section B.2** (The Structure of the CD on page 74).

A.1.3 Before the installation

First download Tomcat 7.0.27 (or higher) for operating system you use from the URL *http://tomcat.apache.org*. Then checkout the *thesis.zip* file from

svn.ms.mff.cuni.cz/svn/undergrads/students/vandk6am/thesis/.

or copy the file from the installation cd, see **section B.2** (The Structure of the CD on page 74). Then unzip the *thesis.zip* file to a prefered directory.

Please open the *thesis.properties* file from the installation package in your favourite editor and change **home** variable to the correct value.

Example. If you unpacked the package on Windows in such a way that the *thesis.properties* file from the installation package is at the *c:\vandas\thesis\thesis.properties* directory, change the **home** variable to

home = c:\\vandas\\thesis.

Please note that using a double backslash is crucial here. Please also pay attention not to use the trailing backslash.

Example. If you unpacked the package on Unix in a way that the *thesis.properties* file from the installation package is at the */home/vandas/thesis/thesis.properties* directory, change **home** variable to

home = /home/vandas/thesis.

Please also pay attention not to use the trailing slash.
Also the **separator** variable needs to be specified.
For Windows environment set

separator=${separator_windows}

#separator=${separator_unix}

For Unix environment set

#separator=${separator_windows}

separator=${separator_unix}

A.1.4 The Installation Procedure

In the */tools/apache-tomcat-7.0.27-conf* directory there is the *server.xml* file. This file needs to be used as a replace of the Tomcat configuration file placed in the *[tomcat]/conf/* directory, see **section B.1** (Development installation for Unix Systems on page 73) for changes in the Tomcat settings.

Then the *thesis.war* file from the installation package needs to be placed into the *[tomcat]/webapps/* directory.

Now **start the server** - the package will be unpacked in

the *[tomcat]/webapps/thesis/* directory.

Stop the server and place the *thesis.properties* file from the installation package after modifications (the **home** and the **separator** variables) into

the *[tomcat]/webapps/thesis/WEB-INF/conf/* directory.

Well done, you can **start the server**, go to the URL *http://127.0.0.1:8080/thesis/* and play around.

A.1.5 Troubleshooting

Port in use

Question: My server does not start.
Answer: Make sure there is no application using the port 8080.

Different Tomcat Server Configuration

Question: My server still does not start.
Answer: Make sure your server is placed on the URL *http://127.0.0.1:8080.* Another location might be the URL *http://localhost:8080.*

Configuration Incorrect

Question: My server still does not start!
Answer: Make sure the **home** and **separator** variables are set correctly in the *thesis.properties* file from the installation package and **without** a trailing slash and **with all backslashes doubled for Windows**. Make sure the file is placed in the correct configuration folder of the Tomcat web server. Also check that path does not contain any whitespace.

Java Environment Settings Problems

Question: My server still does not start!!
Answer: Make sure Tomcat can use your Java JRE (correctly set the **$JAVA_HOME** environment variable). If you are not sure you understand the problem properly, please install Java JRE 1.5+.

Missing Files in Installation Package

Question: My server still does not start!!!
Answer: Make sure you have the full installation package containing all the data files.

When No Solution Has Been Found...

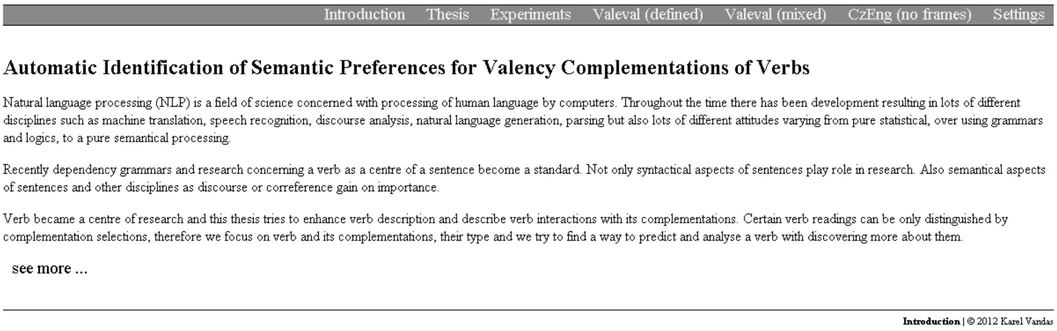
Question: My server still does not start!!!!
Answer: Send an e-mail to vandas (at) ufal.mff.cuni.cz with the description of a problem, zipped directory of the tool and zipped the Tomcat installation you are trying to run (it contains log files we can inspect).

A.2 Tool Parts

The tool works as a web application. Initial screen consists of several links - Introduction, Thesis, Experiments, Valeval (defined), Valeval (mixed), Czeng (no frames) and Settings, see Figure A.1 on page 66.

A.2.1 Introduction

The introduction section gives to the user the first three paragraphs of thesis introduction. When there are any messages from the tool workflow, they are placed to this page (once they are shown they disappear).



Automatic Identification of Semantic Preferences for Valency Complementations of Verbs

Natural language processing (NLP) is a field of science concerned with processing of human language by computers. Throughout the time there has been development resulting in lots of different disciplines such as machine translation, speech recognition, discourse analysis, natural language generation, parsing but also lots of different attitudes varying from pure statistical, over using grammars and logics, to a pure semantical processing.

Recently dependency grammars and research concerning a verb as a centre of a sentence become a standard. Not only syntactical aspects of sentences play role in research. Also semantical aspects of sentences and other disciplines as discourse or correference gain on importance.

Verb became a centre of research and this thesis tries to enhance verb description and describe verb interactions with its complementations. Certain verb readings can be only distinguished by complementation selections, therefore we focus on verb and its complementations, their type and we try to find a way to predict and analyse a verb with discovering more about them.

Figure A.1: The introduction screen of the tool.

A.2.2 Thesis

The thesis section contains a link to a page with the full version of the thesis text. An the time of writing, the thesis is available at

theURLhttp://www.ms.mff.cuni.cz/˜vandk6am/thesis/thesis.pdf

A.2.3 Experiments

The experimens section contains all the settings presented in this thesis. All those experimental settings have also a short description as it is presented in the thesis.

A.2.4 Tool Data Processing

Valeval (defined)

This section gives an overview of the tool settings, see **section A.2.5** (Settings on page 66), and the output of the tool for set of example sentences from Valeval with assigned valency frames (separated valency frames from Valeval form clusters). Each time the link is executed the data set is recounted.

Valeval (mixed)

Similarly to **section A.2.4** (Valeval (defined) on page 66), the tool gives an output for input sentences from Valeval without a verb valency assignment.

CzEng (no frames)

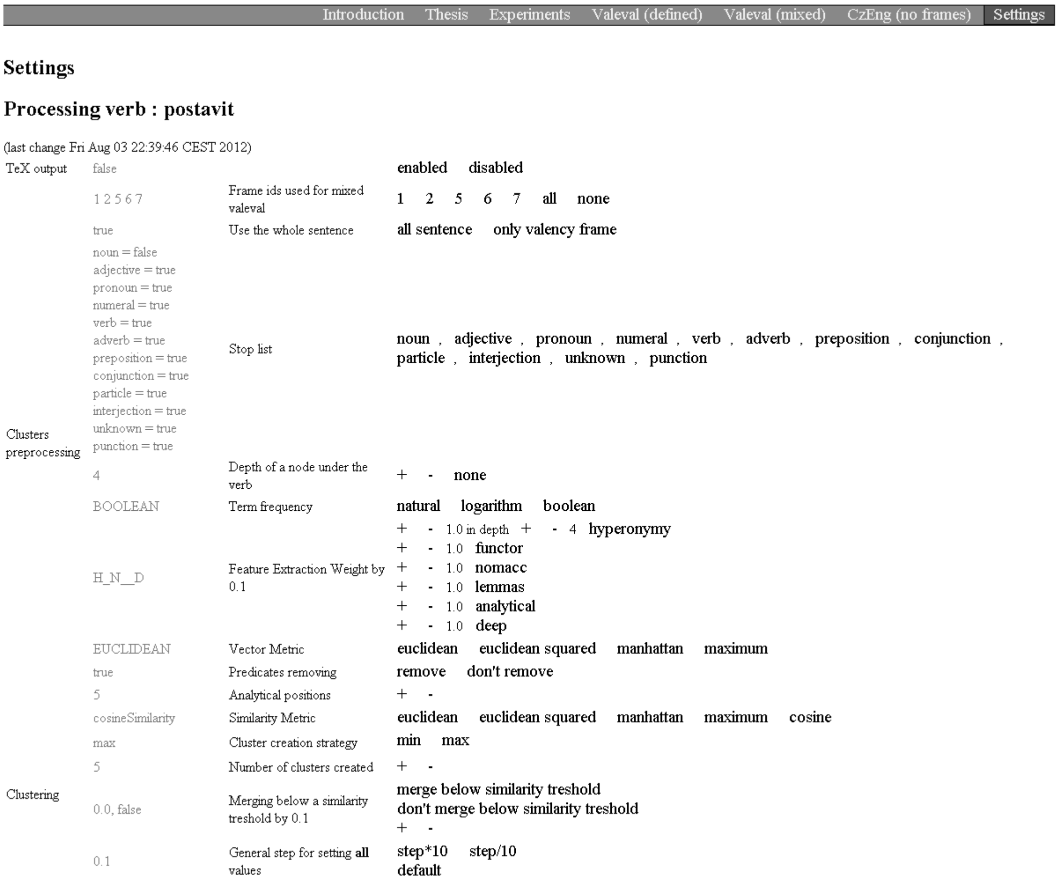
Similarly to **section A.2.4** (Valeval (defined) on page 66), the tool gives an output for input sentences from CzEng 1.0 data release without a verb valency assignment.

A.2.5 Settings

This section serves to change parameters of the tool to be applied on the data.

A.3 Tool Control

Now let's discuss the functionality of the tool.



Settings

Processing verb : postavit

(last change Fri Aug 03 22:39:46 CEST 2012)

TeX output	false		enabled disabled
Clusters preprocessing	1 2 5 6 7	Frame ids used for mixed valeval	1 2 5 6 7 all none
	true	Use the whole sentence	all sentence only valency frame

	noun = false
	adjective = true
	pronoun = true
	numeral = true
	verb = true
	adverb = true
	preposition = true
	conjunction = true
	particle = true
	interjection = true
	unknown = true
	punction = true
		Stop list	noun , adjective , pronoun , numeral , verb , adverb , preposition , conjunction , particle , interjection , unknown , punction

	4	Depth of a node under the verb	+ - none
	BOOLEAN	Term frequency	natural logarithm boolean
	H_N__D	Feature Extraction Weight by 0.1	+ - 1.0 in depth + - 4 hyperonymy + - 1.0 functor + - 1.0 nomacc + - 1.0 lemmas + - 1.0 analytical + - 1.0 deep
	EUCLIDEAN	Vector Metric	euclidean euclidean squared manhattan maximum
	true	Predicates removing	remove don't remove
	5	Analytical positions	+ -
Clustering	cosineSimilarity	Similarity Metric	euclidean euclidean squared manhattan maximum cosine
	max	Cluster creation strategy	min max
	5	Number of clusters created	+ -
	0.0, false	Merging below a similarity treshold by 0.1	merge below similarity treshold don't merge below similarity treshold + -
	0.1	General step for setting **all** values	step*10 step/10 default

Figure A.2: Settings screen of the tool.

A.3.1 What is the Tool Good For

The tool can investigate a semantic preferences for valency complementations of verbs. Apart from that it contains agglomerative hierarchical clustering implementation. The tool is suitable for users interested in clustering, verb valency and verb complementations abstraction.

A.3.2 How Can I Set the Tool

The tool has a Setting part, see **section A.2.5** (Settings on page 66), where various settings are available. For settings descriptions please consult **section 5.2** (Parameters to be Set and Their Effects on Data Analysis on page 27).

Settings can be changed by clicking on the links changing a parameter's value. Also processed verbs can be changed by clicking on them, see Figure A.2 on page 67. The read-only current settings are also shown on each screen where sets of data are processed.

A.3.3 Usual Workflow of the Tool

I want to see abstraction of Valeval separated valency frames

Please, navigate to Valeval (defined) link. Then select desired frame by navigating the link of a cluster representing the frame.

I want to see abstraction of Valeval mixed valency frames

Please, first make sure the settings for clustering is set as you wish, see Figure A.2 on page 67. Then navigate to Valeval (mixed) link. In a moment you see valency frame clusters, by clicking at the cluster number you will get the complete information about the abstraction of a cluster.

I want to see abstraction of CzEng data

Please, first make sure the settings for clustering is set as you wish, see Figure A.2 on page 67. Then navigate to CzEng (no frames) link. In a moment you see valency frame clusters, by clicking at the cluster number you will get the complete information about the abstraction of a cluster.

I want to reset the settings to default

Please restart the server. Note that by a restart you loose all the information previously gathered.

A.3.4 Screenshots and Screen descriptions

We have already presented the introduction screen, see Figure A.1 on page 66, and the settings screen, see Figure A.2 on page 67. The thesis screen only contains a frame with generated pdf with this thesis.

Clusters Overview Screen

Clusters overview screen contains four parts. The settings part visualise the settings used for generation of clusters, resulting clusters part describes found clusters with functors identified, cluster analysis describes the process of clustering with similarity for merging and "End condition" part points out the condition that has stopped the process of clustering, see Figure A.3 on page 69.

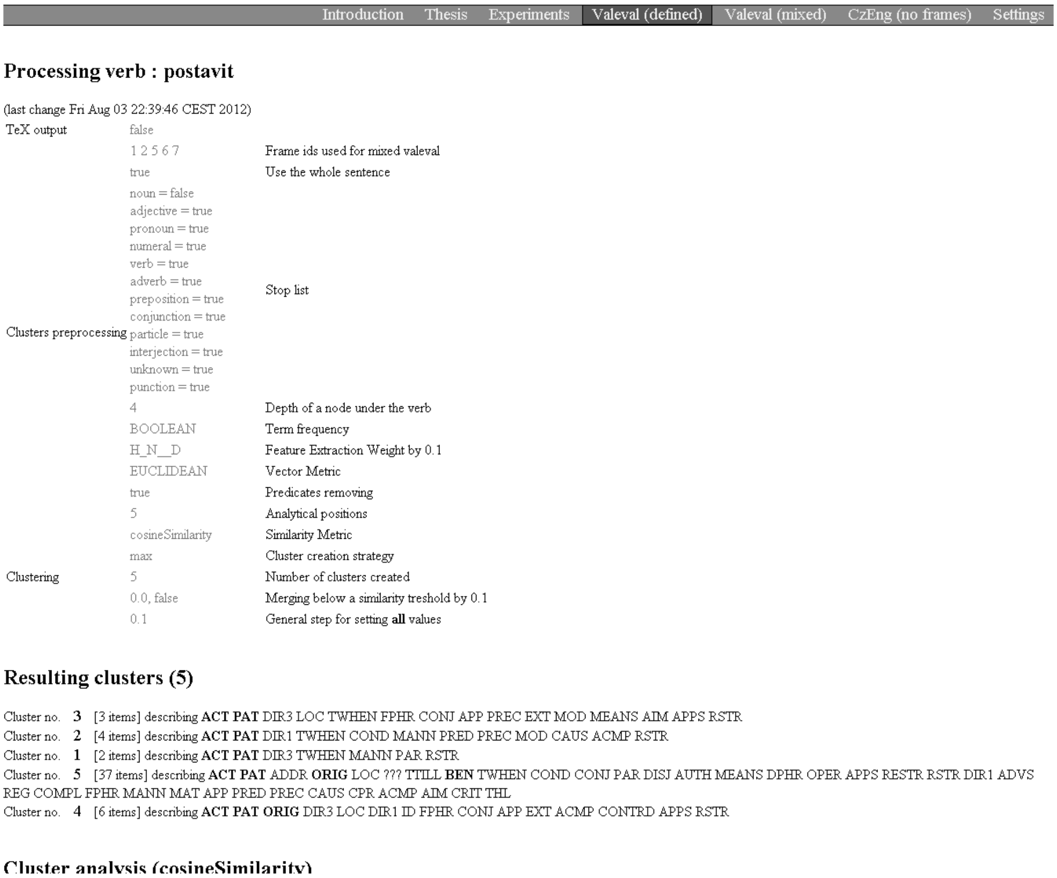
Each cluster has its own identificator, this id serves as a link to the cluster overview screen, see **section A.3.4** (Cluster Overview Screen on page 68).

Cluster Overview Screen

Each cluster has its description, where parents of the cluster are mentioned and can be inspected, followed by an abstraction analysis, (optionally) TeX definitions of the output data and Sentence overview.

Abstraction analysis contains description of found functors and their abstractions, giving coloured distinguishing of abstractions - gray is equal to baseline, red is an abstraction, see Figure A.4 on page 70. Bolded functors mark obligatory functors for valency frame (available only for Valeval data source).

Each sentence contains a link leading to the sentence in the Valeval corpus.



Processing verb : postavit

(last change Fri Aug 03 22:39:46 CEST 2012)

TeX output	false	
Clusters preprocessing	1 2 5 6 7	Frame ids used for mixed valeval
	true	Use the whole sentence

	noun = false
	adjective = true
	pronoun = true
	numeral = true
	verb = true
	adverb = true
	preposition = true
	conjunction = true
	particle = true
	interjection = true
	unknown = true
	punction = true
		Stop list

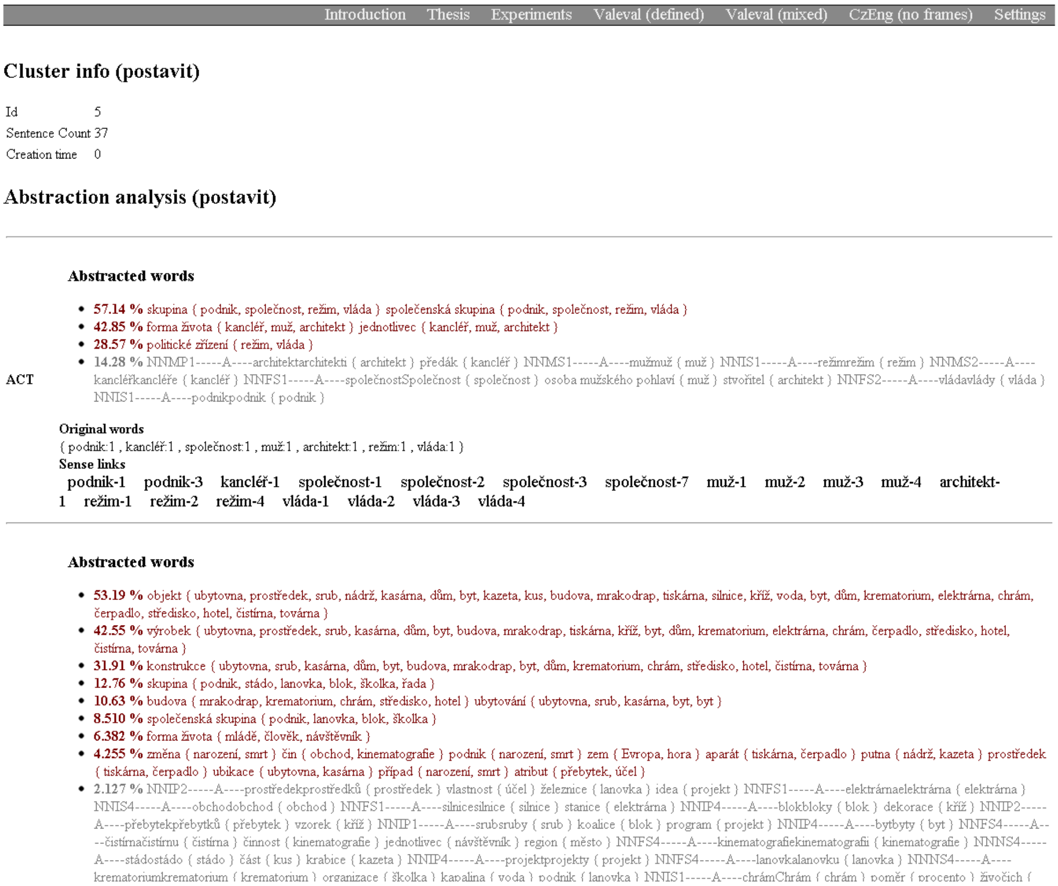
	4	Depth of a node under the verb
	BOOLEAN	Term frequency
	H_N__D	Feature Extraction Weight by 0.1
	EUCLIDEAN	Vector Metric
	true	Predicates removing
	5	Analytical positions
Clustering	cosineSimilarity	Similarity Metric
	max	Cluster creation strategy
	5	Number of clusters created
	0.0, false	Merging below a similarity treshold by 0.1
	0.1	General step for setting **all** values

Resulting clusters (5)

Cluster no. **3** [3 items] describing **ACT PAT** DIR3 LOC TWHEN FPHR CONJ APP PREC EXT MOD MEANS AIM APPS RSTR
Cluster no. **2** [4 items] describing **ACT PAT** DIR1 TWHEN COND MANN PRED PREC MOD CAUS ACMP RSTR
Cluster no. **1** [2 items] describing **ACT PAT** DIR3 TWHEN MANN PAR RSTR
Cluster no. **5** [37 items] describing **ACT PAT** ADDR **ORIG** LOC ??? TTILL **BEN** TWHEN COND CONJ PAR DISJ AUTH MEANS DPHR OPER APPS RESTR RSTR DIR1 ADVS REG COMPL FPHR MANN MAT APP PRED PREC CAUS CPR ACMP AIM CRIT THL
Cluster no. **4** [6 items] describing **ACT PAT ORIG** DIR3 LOC DIR1 ID FPHR CONJ APP EXT ACMP CONTRD APPS RSTR

Cluster analysis (cosineSimilarity)

Figure A.3: Clusters screen of the tool.



Cluster info (postavit)

Id	5
Sentence Count	37
Creation time	0

Abstraction analysis (postavit)

ACT

Abstracted words

- **57.14 %** skupina { podnik, společnost, režim, vláda } společenská skupina { podnik, společnost, režim, vláda }
- **42.85 %** forma života { kancléř, muž, architekt } jednotlivec { kancléř, muž, architekt }
- **28.57 %** politické zřízení { režim, vláda }
- **14.28 %** NNMP1-----A----architektarchitekti { architekt } předák { kancléř } NNMS1-----A----mužmuž { muž } NNIS1-----A----režimrežim { režim } NNMS2-----A----kancléřkancléře { kancléř } NNFS1-----A----společnostSpolečnost { společnost } osoba mužského pohlaví { muž } stvořitel { architekt } NNFS2-----A----vládavlády { vláda } NNIS1-----A----podnikpodnik { podnik }

Original words
{ podnik:1 , kancléř:1 , společnost:1 , muž:1 , architekt:1 , režim:1 , vláda:1 }
Sense links
podnik-1 podnik-3 kancléř-1 společnost-1 společnost-2 společnost-3 společnost-7 muž-1 muž-2 muž-3 muž-4 architekt-1 režim-1 režim-2 režim-4 vláda-1 vláda-2 vláda-3 vláda-4

Abstracted words

- **53.19 %** objekt { ubytovna, prostředek, srub, nádrž, kasárna, dům, byt, kazeta, kus, budova, mrakodrap, tiskárna, silnice, kříž, voda, byt, dům, krematorium, elektrárna, chrám, čerpadlo, středisko, hotel, čistírna, továrna }
- **42.55 %** výrobek { ubytovna, prostředek, srub, kasárna, dům, byt, budova, mrakodrap, tiskárna, kříž, byt, dům, krematorium, elektrárna, chrám, čerpadlo, středisko, hotel, čistírna, továrna }
- **31.91 %** konstrukce { ubytovna, srub, kasárna, dům, byt, budova, mrakodrap, byt, dům, krematorium, chrám, středisko, hotel, čistírna, továrna }
- **12.76 %** skupina { podnik, stádo, lanovka, blok, školka, řada }
- **10.63 %** budova { mrakodrap, krematorium, chrám, středisko, hotel } ubytování { ubytovna, srub, kasárna, byt, byt }
- **8.510 %** společenská skupina { podnik, lanovka, blok, školka }
- **6.382 %** forma života { mládě, člověk, návštěvník }
- **4.255 %** změna { narození, smrt } čin { obchod, kinematografie } podnik { narození, smrt } zem { Evropa, hora } aparát { tiskárna, čerpadlo } putna { nádrž, kazeta } prostředek { tiskárna, čerpadlo } ubikace { ubytovna, kasárna } případ { narození, smrt } atribut { přebytek, účel }
- **2.127 %** NNIP2-----A----prostředekprostředků { prostředek } vlastnost { účel } železnice { lanovka } idea { projekt } NNFS1-----A----elektrárnaelektrárna { elektrárna } NNIS4-----A----obchodobchod { obchod } NNFS1-----A----silnicesilnice { silnice } stanice { elektrárna } NNIP4-----A----blokbloky { blok } dekorace { kříž } NNIP2-----A----přebytekpřebytků { přebytek } vzorek { kříž } NNIP1-----A----srubsruby { srub } koalice { blok } program { projekt } NNIP4-----A----bytbyty { byt } NNFS4-----A----čistírnačistírnu { čistírna } činnost { kinematografie } jednotlivec { návštěvník } region { město } NNFS4-----A----kinematografiekinematografii { kinematografie } NNNS4-----A----stádostádo { stádo } část { kus } krabice { kazeta } NNIP4-----A----projektprojekty { projekt } NNFS4-----A----lanovkalanovku { lanovka } NNNS4-----A----krematoriumkrematorium { krematorium } organizace { školka } kapalina { voda } podnik { lanovka } NNIS1-----A----chrámChrám { chrám } poměr { procento } živočich {

Figure A.4: Cluster screen of the tool. Notice bolded obligatory functors. Colours mark the functor semantic preferences abstractions, gray colour is for a baseline abstraction level, whereas red gives a solution outperforming the baseline.

A.4 Troubleshooting

A.4.1 Out of Memory

The tool save all the objects that are not temporal. That means that after certain amount of time the out of memory issue rises. As a solution, simply restart the server.

A.4.2 Inspection of the server

To inspect server run you can look at the *[tomcat]/logs* directory and inspect the *localhost.[date].log* file, where the tool start is logged and the *catalina.out* file, where the tool output is logged.

A.4.3 Reset of the server

Even if this is a bit Tomcat question, we advise you to **stop the server**, remove the *[tomcat]/webapps/thesis/* directory, the *[tomcat]/work/Catalina/localhost/thesis/* directory. Server should be now ready for a fresh start.

Appendix B

Really Technical Details

B.1 Development installation for Unix Systems

User installation requires an environment with java JDK 1.5+. We used Tomcat 7.0.27, Maven 2.2.1 and Spring 2.5.

At first do a svn checkout from the subversion address

https://svn.ms.mff.cuni.cz/svn/undergrads/students/vandk6am/thesis.

The package includes the *Makefile* file that specifies directories. Before the installation please change **THESIS_FOLDER** variable to the installation directory (**without** a trailing slash).

Tomcat configuration the *server.xml* file is to be found at the */tools* directory . The only crucial change against the default Tomcat settings file is the URL encoding setting. At the *server.xml* file you should specify *URIEncoding="UTF-8"* at the *Connector* element with your favourite port (usually 8080).

For the installation instances of Maven and Tomcat are needed. This instances are referenced from the *Makefile* file, where the variable **MVN** and **SERVER_PATH** and needs to be specified. You also might need to specify the environment **$JAVA_HOME** variable.

Now you are ready to run "make". This command stops the server if it is running, build the source and place the *compiled war* file in the server the *[tomcat]/webapps* directory and then start the server.

Important. If you intend run czeng generation you need to install the tool to the ufallab machine or change the path variable in the *Makefile* file.

Known issues. Encoding of file names in the */resources/czeng* directory and the */resources/valeval* directory need to be repaired otherwise data of the thesis **are not loaded** properly.

B.1.1 Test version at ufallab.ms.mff.cuni.cz

Sample installation is at the ufallab.ms.mff.cuni.cz machine at the */home/vandas/vandas_thesis* directory.

Log in via ssh with tunelled port, i.e., *ssh vandas@ufallab.ms.mff.cuni.cz -L 8090:localhost:8080.* Change the directory to the */home/vandas/vandas_thesis/thesis* directory and execute "make start".

Now go to the URL *http://localhost:8090/thesis* in your browser. To stop the server simply type "make stop".

B.2 The Structure of the CD

The light installation placed on this CD can be generated at any time by running a command “make cd” from the *Makefile* file. The light installation is placed in the *thesis.zip* file.

The full version of the installation CD contains following:

In the */documentation* directory the javadoc documentation can be found.

In the */paper* directory the thesis and files to generate it are placed.

In the */preprocessing* directory scripts and transformation for data input accomodation are inserted.

In the */resources* directory all the resources needed for starting the thesis such as data files, abstraction definitions, etc. are placed.

In the */source* directory the source to run the thesis tool can be found.

In the */tools* directory only the *configuration* file for Apache Tomcat web server in version 7.0.27 resides.

B.3 Data Format Examples

This section introduces examples of used data formats. For an introduction into the format descriptions, please refer to **section 4.4** (Data Formats Used for This Thesis on page 23).

B.3.1 The CzEng 1.0 Data Format

CzEng 1.0 data release has a complex format. It captures morphological, as well as analytical and tectogramatical layers of a sentence. See the example at Figure B.1 on page 74.

```
<LM id=”a_tree−cs−fiction−b1−00train−f00001−s1−n3770”>
     <children id=”a_tree−cs−fiction−b1−00train−f00001−s1−n3771”>
          <form>duchu</form>
          <lemma>duch</lemma>
          <tag>NNMS6———A−−−1</tag>
          <no_space_after>0</no_space_after>
          <ord>4</ord>
          <afun>Adv</afun>
          <edge_to_collapse>1</edge_to_collapse>
          <is_auxiliary>0</is_auxiliary>
          <alignment>
...
          </alignment>
     </children>
...
</LM>
```

Figure B.1: The example of analytical description of the word *duch* (*ghost*).

B.3.2 The Vallex Data Format

The Vallex data format describes verbs and their senses. See the example at Figure B.2 on page 76.

B.3.3 The Valeval Data Format

The Valeval data format captures examples of usage of chosen verbs. See the example at Figure B.3 on page 77.

B.3.4 The WordNet Data Format

The original WordNet data format is presented in Figure B.4 on page 78.

Czech WordNet Issues and Solutions

The issues, such as non-atomic values or no root element, of Czech WordNet have been solved by a XSL transformation. The loops in hyperonymy data definitions are solved programatically by identifying elements the tool data representation algorithm has already seen when searching for a particular abstraction.

B.3.5 The Tool Data Format

The Tool Data Format. The CzEng Data Format Origin

The tool data format based on the CzEng data format is used for CzEng data representation and Valeval representation. The format is presented at Figure B.5 on page 79.

The Tool Data Format. The WordNet Data Format Origin

We decided to create a XSL transformation of a document to a well-defined XML. This XML should only contain the definition of a word, its id, literal versions of a word and a hyperonymy that it is related to. The outcome of this transformation is used as an input for further processing [1].

The modified WordNet data format is presented at Figure B.6 on page 80.

[1] The transformation source can be found at the */preprocessing* directory The outcome of the transformation can be found at the */resources/wordnet* directory.

```
<lexeme_cluster>
    <lexeme pos='v' id='lxm-v-postavit'>
        <lexical_forms>
            <mlemma aspect='pf' coindex='pf'>postavit</mlemma>
        </lexical_forms>
        <lexical_units>
            <lu_cluster id="luc-v-postavit-1">
                <blu id='blu-v-postavit-1'>
                    <frame>
                        <slot functor='ACT' type='obl'>
                            <form type="direct_case" case="1"
                                />
                        </slot>
                        <slot functor='PAT' type='obl'>
                            <form type="direct_case" case="4"
                                />
                        </slot>
                        <slot functor='ORIG' type='opt'>
                            <form type="prepos_case"
                                prepos_lemma="z" case="2" />
                        </slot>
                        <slot functor='BEN' type='typ'>
                            <form type="direct_case" case="3"
                                />
                        </slot>
                    </frame>
                    <gloss>
                        vybudovat; vytvořit
                    </gloss>
                    <example>
                        postavit dětem altánek; postavit sochu;
                            postavit z balzy model letadla
                    </example>
                    <rfl type="pass">
                        dům se postavil za několik měsíců
                    </rfl>
                    <class>
                        change
                    </class>
                </blu>
            </lu_cluster>
            <lu_cluster id="luc-v-postavit-2">
                <blu id='blu-v-postavit-2'>
                    <frame>
...
                    </frame>
                </blu>
            </lu_cluster>
        </lexical_units>
    </lexeme>
</lexeme_cluster>
```

Figure B.2: The Vallex data format. From the Vallex data format frame definitions are used to highlight valency frames in the tool. Here we demonstrate the verb *postavit*.

```
<valeval xmlns='http://ufal.mff.cuni.cz/vallex-valeval'>
    <body>
        <occurence number='65' frame='1'>
            <sentence>Na Nechranické přehradě na Ohři, která patří k
                 největším vodním nádržím v republice, bude provedena
                 oprava 3280 metrů dlouhé sypané hráze narušené
                vodními vlnami.</sentence>
            <sentence>Uvedl to mluvčí a. s. Povodí Ohře Petr Vít.</
                sentence>
            <sentence>Oprava bude podle Víta zahájena v létě při
                nejnižším stavu vody a potrvá pravděpodobně do příšt
                ího  roku.</sentence>
            <sentence is_here='1'>Nechranická vodní nádrž byla <
                word>postavena</word> v letech 1961 až 1968.</
                sentence>
        </occurence>
...
    </body>
</valeval>
```

Figure B.3: The Valeval data format. We used this data format for extracting sentences for each valency frame separatelly. Note the context sentences of each verb occurence. Here we demonstrate the verb *postavit*.

```
<SYNSET>
        <ID>00004885-v</ID>
        <POS>v</POS>
        <SYNONYM>
                <LITERAL>frkat
                        <SENSE>1</SENSE>
                </LITERAL>
                <LITERAL>zafunet
                        <SENSE>1</SENSE>
                </LITERAL>
                <LITERAL>zafrkat
                        <SENSE>1</SENSE>
                </LITERAL>
        </SYNONYM>
</SYNSET>
<SYNSET>
        <ID>00005811-v</ID>
        <POS>v</POS>
        <SYNONYM>
                <LITERAL>mrkat
                        <SENSE>2</SENSE>
                </LITERAL>
                <LITERAL>zamrkat
                        <SENSE>1</SENSE>
                </LITERAL>
        </SYNONYM>
        <ILR>00559482-v
                <TYPE>hypernym</TYPE>
        </ILR>
</SYNSET>
 ...
```

Figure B.4: The WordNet data format. The word *mrkat* indicates an existing hyperonymy relation.

```
<document>
    <utterance>
        <sentence>Nechranická vodní nádrž byla postavena v letech
           1961 až 1968.</sentence>
        <lemmas>
            <lemma tag="AAFS1----1A----" form="Nechranická" id="
               a_tree-cs-s32-n754">nechranický</lemma>
            <lemma tag="AAFS1----1A----" form="vodní" id="a_tree-
               cs-s32-n755">vodní</lemma>
            <lemma tag="NNFS1-----A----" form="nádrž" id="a_tree-
               cs-s32-n756">nádrž</lemma>
            <lemma tag="VpQW---XR-AA---" form="byla" id="a_tree-
               cs-s32-n757">být</lemma>
            <lemma tag="C=-------------" form="1961" id="a_tree-
               cs-s32-n761">1961</lemma>
            <lemma tag="C=-------------" form="1968" id="a_tree-
               cs-s32-n763">1968</lemma>
            <lemma tag="J^-------------" form="až" id="a_tree-cs-
               s32-n762">až-1_^(2_až_3)</lemma>
            <lemma tag="NNNP6-----A----" form="letech" id="a_tree
               -cs-s32-n760">rok</lemma>
            <lemma tag="RR--6----------" form="v" id="a_tree-cs-
               s32-n759">v-1</lemma>
            <lemma tag="VsQW---XX-AP---" form="postavena" id="
               a_tree-cs-s32-n758">postavit_:W</lemma>
            <lemma tag="Z:-------------" form="." id="a_tree-cs-
               s32-n764">.</lemma>
        </lemmas>
    <node t_lemma="postavit" functor="PRED" id="a_tree-cs-s32-n758"
       >
        <children>
            <node t_lemma="nádrž" functor="PAT" id="a_tree-cs-s32-
               n756">
              <node t_lemma="nechranický" functor="RSTR" id="
                 a_tree-cs-s32-n754"></node>
              <node t_lemma="vodní" functor="RSTR" id="a_tree-
                 cs-s32-n755"></node>
            </node>
            <node t_lemma="rok" functor="TWHEN" id="a_tree-cs-s32
               -n760">
              <node t_lemma="až" functor="OPER" id="a_tree-cs-
                 s32-n762">
                <children>
                    <node t_lemma="1961" functor="RSTR" id
                       ="a_tree-cs-s32-n761"></node>
                    <node t_lemma="1968" functor="RSTR" id
                       ="a_tree-cs-s32-n763"></node>
                </children>
              </node>
            </node>
        </children>
    </node>
    </utterance>
...
</document>
```

Figure B.5: The modified CzEng data format. The favourite sentence example.

```
<synsets>
    <synset id="00004885-v" pos="v">
        <literals>
            <literal lemma="frkat" sense="1"></literal>
            <literal lemma="zafunět" sense="1"></literal>
            <literal lemma="zafrkat" sense="1"></literal>
        </literals>
        <relations></relations>
    </synset>
    <synset id="00005811-v" pos="v">
        <literals>
            <literal lemma="mrkat" sense="2"></literal>
            <literal lemma="zamrkat" sense="1"></literal>
        </literals>
        <relations>
            <relation id="00559482-v" type="hypernym">
            </relation>
        </relations>
    </synset>
...
<synsets>
```

Figure B.6: The modified WordNet data format. The word *mrkat* indicates an existing hyperonymy relation.