From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description

Silvie Cinková

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1
Czech Republic
Phone: +420-221 914 257
Fax: +420-221 914 309
cinkova@ufal.mff.cuni.cz

Abstract

EngValLex is the name of an FGD-compliant valency lexicon of English verbs, built from the PropBank-Lexicon and following the structure of Vallex, the FGD-based lexicon of Czech verbs. EngValLex is interlinked with the PropBank-Lexicon, thus preserving the original links between the PropBank-Lexicon and the PropBank-Corpus. Therefore it is also supposed to be part of corpus annotation. This paper describes the automatic conversion of the PropBank-Lexicon into Pre-EngValLex, as well as the progress of its subsequent manual refinement (EngValLex). At the start, the Propbank-arguments were automatically re-labeled with functors (semantic labels of FGD) and the PropBank-rolesets were split into the respective example sentences, which became FGD-valency frames of Pre-EngValLex. Human annotators check and correct the labels and make the preliminary valency frames FGD-compliant. The most essential theoretical difference between the original and EngValLex is the syntactic alternations used by the PropBank-Lexicon, not yet employed within the Czech framework. The alternation-based approach substantially affects the conception of the frame, making in very different from the one applied within the FGD-framework. Preserving the valuable alternation information required special linguistic rules for keeping, altering and re-merging the automatically generated preliminary valency frames.

1. Introduction

The ongoing FGD[1]-based re-annotation of the Wall Street Journal subcorpus of the Penn Treebank (PTB-WSJ) (Mitchel et al., 1993) requires valency lexicons of verbs, nouns and adjectives to be mapped onto the data. We have started with resolving the valency of verbs, and it is the valency lexicon of English verbs that stands in focus of this paper.

Instead of building our own lexicon from scratch, we made use of the already existing Proposition Bank (Palmer et al., 2005). PropBank consists of a valency lexicon of verbs interlinked with a corpus annotation built above PTB-WSJ. A successful transformation of the original PropBank-lexicon into a FGD-compliant format includes preserving the links between the original lexicon and the original *manual* data annotation as well as their transformation into the new FGD-compliant lexicon and an *automatic* FGD-compliant annotation of verbal frames in PTB-WSJ. We have named the final lexicon EngValLex, after Vallex, the already existing FGD-compliant lexicon of Czech verbs.

2. Motivation

After the completion of the first FGD-implementation - the Prague Dependency Treebank 2.0[2] (PDT 2.0, 2005), we are aiming at building a dependency-based parallel Czech-English treebank with deep-syntactic ("tectogrammatical", "TR") annotation[3], the Prague Czech-English Dependency Treebank 2.0 (cf. (Čmejrek et al., 2005)). When building the English counterpart of the treebank (the Prague English Dependency Treebank - PEDT), the original surface-syntax annotation of PTB-WSJ was converted into dependency trees, above which the TR-annotation layer is being built.

While the Czech counterpart (a professional-grade translation of the PTB-WSJ) will be, for the time being, annotated only automatically using tools developed on the basis of the (Czech) PDT 2.0., PEDTwill be annotated manually. Nevertheless, we seek to save the costly and time-consuming work of human annotators by automatic pre-annotation, which also is the case of the PropBank-Lexicon and the PropBank-corpus being "recycled" into the EngValLex-lexicon and the automatic pre-annotation of PEDT.

3. From PropBank to EngValLex

Fig. 1 illustrates the main stages of the process. The transformation of both the PropBank-Lexicon together with the PropBank-Corpus annotation into EngValLex and the PEDT pre-annotation consists of the following steps:

1. Linguistic comparison of the PropBank-Lexicon and the Czech Vallex lexicon
2. Automatical generation of Pre-EngValLex

[1] Functional Generative Description: (Sgall et al., 1986).

[2] A syntactically parsed corpus of Czech texts with complex semantic annotation over 0,8 million words.

[3] The TR annotation layer is the unique contribution of the FGD-framework. Though remaining language-specific, the TR representation of a sentence in the source language A and that of its translation equivalent in the target language B are supposed to be far more similar than their surface syntax representations. This is especially desirable in Machine Translation, as the transfer between the source-language analysis and the target-language synthesis is made shorter and thus likely to be less error-prone than it is the case with surface syntax representations (Hajič, 2002). The parallel corpus is meant to be a source for experiments, which will prove (or disprove) this assumption.

3. Creating EngValLex (manual refinement)
4. Automatic mapping of EngValLex onto PEDT
5. Manual and automatic corrections of verb frames during the general manual annotation of PEDT.

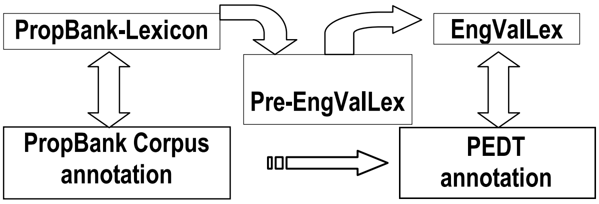

Fig. 1: The PropBank-Lexicon is automatically converted into Pre-EngValLex. Its links towards the original corpus annotation remain, although the shape of the respective frames as well as the labeling of arguments are going to be altered.

3.1. Linguistic Comparison of the PropBank-Lexicon and the Czech Vallex Lexicon

3.1.1. Characteristics of the PropBank-Lexicon

The verb entries of the PropBank-Lexicon are divided into rolesets. Rolesets roughly correspond to senses. They rely on syntactic rather than on semantic criteria, which are considered to be "subjective and potentially unlimited" (PropBank Annotation Guidelines, (2002) p.3), while syntactic distinctions be "rigorous and objective". Therefore rolesets are much more coarse-grained than e.g. WordNet senses.

Each roleset includes a set of labeled arguments ("roles") and one or more example sentences, in which combinations of the roles are rendered by the surface syntax. Rolesets are numbered within the respective entries ("roleset-ID's"). In addition, each roleset has a definition-like description attached ("roleset names"). E.g. the verb *to yell* has only one roleset, *yell.01*, which is labeled as *"to cry out loudly"*. The verb *to abandon*, on the other hand, has three rolesets (*abandon.01.-03*). They are labeled as *"leave behind", "exchange", "surrender, give_over"*, respectively. The PropBank-Lexicon comprises about 2000 roleset names, in which about 4600 rolesets are grouped. (There are 3323 verb entries in the PropBank-Lexicon. Some of them also include phrasal verbs. Phrasal verbs do not have entries of their own but are displayed as rolesets).

PropBank's conception of the argument structure (henceforth: valency) derives from Levin's assumption (Levin, 1993) that the syntactic alternations verbs participate in are not arbitrary but reflect underlying semantic components of the events denoted by each given verb. Semantically related verbs can be grouped into classes according to which alternations they take part in. The roleset names group semantically and syntactically related verb senses into classes like the Levin classes. The PropBank classes go somewhat across the Levin verb classes in accordance with the valency behavior of the living data in PTB-WSJ.

Each roleset introduces an enumeration of arguments (roles). The arguments are divided into "numbered arguments" and "adjuncts" (PropBank Annotation Guidelines, 2002). The numbered arguments are arguments that take part in the syntactic alternations analyzed by Levin, (1993) and can become syntactic subjects. The adjuncts are optional, often rendered by prepositional groups and adverbs. Each argument has two parts: the argument number and a semantic descriptor specific to the given roleset. E.g. *to yell* would acquire the following arguments:

Arg0:*Yeller*
Arg1:*Utterance*
Arg2:*Hearer*

The first roleset of *to abandon (abandon.01 "leave behind")* will have the following arguments:

Arg0:*abandoner*
Arg1:*thing abandoned, left behind*
Arg2:*attribute of arg1*

The arguments do not have to be all present on the surface of a sentence at the same time. Thus the first example sentence *And they believe the Big Board.***Arg0***, under Mr. Phelan, has abandoned their interest.***Arg1** contains only Arg0 and Arg1, while the second example sentence contains all three: *John.***Arg0** *abandoned his pursuit.***Arg1** *of an Olympic gold medal as a waste.***Arg2** *of time.*

Considering the syntactic alternations as pairs of alternation realizations, one can often identify alternation realizations in the rolesets. Each example sentence is provided with a supplementary comment, which even sometimes suggests which alternation realization the given sentence represents. Yet these comments are not formalized, nor is it explicitly stated by which alternations the respective rolesets are defined. About one half of PropBank entries are mapped onto VerbNet (Kipper et al., 2002) in which relevant alternations are listed for each verb entry. However, the linking between the PropBank-Lexicon and VerbNet does not reach as deep as to the respective example sentences. Besides that, the PropBank verb classes (i.e. the roleset names) do not correspond to the VerbNet verb classes (i.e. the original Levinian verb classes), and thus the example sentences in the PropBank-Lexicon do not necessarily show the same alternation patterns as the corresponding entry in VerbNet.

3.1.2. The Structure of Vallex/PDT-Vallex

The Czech valency lexicon describes the valency behavior of a given lexeme (verb, noun, adjective or adverb) in form of valency frames, which roughly correspond to senses. Like rolesets in the PropBank-Lexicon, the valency frames primarily rely on syntactic criteria though the syntactic criteria are "softened" with regard to the semantics of the given verb (see below). A valency frame in the strict sense consists of inner participants and obligatory free modifications (see e.g. Panevová, 2002). Free modifications are prototypically optional and do not belong to the valency frame in the strict sense though some frames require a free modification (e.g. direction in verbs of movement). Both the obligatory and the optional inner participants belong to the valency frame in the strict sense. Like the free modifications, the inner participants have semantic labels according to the cognitive roles they typically enter: ACT (Actor), PAT (Patient), ADDR (Addressee), ORIG (Origin) and EFF (Effect). However, if a verb only has one inner participant, it is automatically labeled with ACT. A two-participant verb always has an ACT and a PAT.

The valency lexicon of Czech verbs has two branches – the PDT-Vallex (i) (see Hajič et al., 2003) and Vallex (ii) (Straňáková-Lopatková, Žabokrtský, 2002). Their structure and function have been described by Lopatková (2003):

"(i) The first branch is represented by the lists of valency frames being created and used by annotators during their work. It contains valency frames of words (verbs and nouns) in their particular meanings (as they appear in PDT) and serves for consistency of annotation.

(ii) The second branch is represented by the valency lexicon, in which the words (only verbs in this stage) are analyzed in the whole complexity, in all their meanings. Rich syntactic annotation is assigned to particular valency frames, including e.g. control and reciprocity."

PDT-Vallex and Vallex are very similar in structure: each lexeme corresponds to one entry. The entry is divided into valency frames. A valency frame is modeled as a sequence of frame slots. Each frame slot corresponds to one complementation of the verb in question. Each slot is assigned a functor according to its semantic relation towards the governing verb. Each slot includes an enumeration of its surface forms. Each frame includes at least one example sentence.

PDT-Vallex notes only the valency frames in the strict sense, i.e. obligatory or optional inner participants and obligatory free modifications), while Vallex also lists optional free modifications typical of the given frame.

When delimiting the respective valency frames, syntactic as well as semantic criteria are adopted. Therefore a verb can have two valency frames with identical distribution of functors. Lopatková (2003) notes that "the change in morphemic realization signalizes the possibility of different meanings; on the other hand, particular complementation in a valency frame can have morphemic variants (if the meaning is 'sufficiently close')". Compared to the PropBank-Lexicon, the distinction of the respective valency frames appears to be somewhat more fine-grained in the annotations of Vallex and PDT-Vallex, which, in any case, are well aware of the absence of reliable semantic criteria, and also prefer syntactic distinction criteria to the semantic. However, no verb classes like the Levinian have been established for Czech. The enormous word formation potential of Czech makes it difficult to build a list of surface syntax alternations as their realizations are rendered by different verbs, though some regularity is evident, mainly in derivations. E.g. to render the Basic Transitive-Causative vs. the Intransitive-Inchoative alternation pair, Czech will often (but not always) employ reflexivization:

John *otevřel* dveře. (John *opened* the door.)

Dveře *se otevřely*. (The door *opened*.)

The Locative *with* Alternation will (in some verbs) be rendered by different prefixes:

John ***na****stříkal* barvu na zeď. (John *sprayed* paint on the wall.)

John ***po****střikal* zeď barvou (John *sprayed* the wall with paint).

3.1.3. Similarities and Differences

Both the PropBank and the ValLex-style approaches assume that sense distinctions are reflected in varying valency frames and both look into their deep syntax, employing semantic judgments. Both provide the opened valency slots with labels. The philosophy of labeling is yet different. While FGD employs rather general semantic labels, the PropBank-Lexicon combines the sheer numbering of arguments with verb-specific semantic descriptions. These verb-specific descriptions in their turn follow some regularities within the respective verb classes (as they are suggested by the names of rolesets). This implies that there is no straightforward matching between a given Arg and a given functor but relations within an entire group of verbs have to be taken into account to assign correct functors to the given Args when building EngValLex.

The most substantial difference between the PropBank-Lexicon and Vallex lies in the very conception of the frame. The PropBank-Lexicon annotation observes alternation patterns in verbs to merge them into rolesets while Vallex does not display any relations between frames within one lemma. The PropBank-Lexicon example sentences within a given roleset often represent realizations of a given alternation pair, e.g. "Intransitive - Inchoative" and "Transitive - Causative". Both sentences would refer to a roleset with the listed arguments Arg0 and Arg1, although Arg0 never emerges in the inchoative sentence. Vallex would separate such instances into two frames, the inchoative having only the "Actor" label, the causative having an Actor and a Patient. In FGD agentivity is basically not an issue. All first inner participants (typically represented by syntactic subjects) are Actor-labeled.

3.2. Pre- EngValLex

Pre-EngValLex is the product of the automatic conversion of the PropBank-Lexicon into an FGD-compliant form. It has the following features:

1. The PropBank-rolesets were automatically split into their respective example sentences. Each example sentence got a header with functors, which it inherited from the list of roles located at the beginning of the particular PropBank-roleset it used to belong to. The example sentences are now considered as preliminary FGD-valency frames.
2. The argument labels (Args) were turned into functors by means of a handful of simple rules.
3. Each preliminary valency frame has its own ID saying which roleset the original sentence used to belong to.
4. Unlike in the PropBank-Lexicon, the respective Pre-EngValLex files are not grouped as lemma files but according to the respective roleset names.

3.2.1. Assigning Functors to Args

Due to the difference in theoretical approaches no straightforward mapping could be performed. We had to make use of all hints the xml-data was offering, mainly the non-formalized attribute "role descr". Even the mapping of the PropBank-Lexicon to other lexical sources was exploited. About 50% of the PropBank-Lexicon is mapped onto VerbNet, which uses its own semantic labeling (the attribute "vntheta" in the xml data). The semantic labels from VerbNet mainly helped to classify adjuncts (ArgM's) and Args with higher numbers.

Rules for ArgM's and higher Args typically looked like this:

If <role descr="low point" n="3", functor: DIR1
If <role descr="instrument" n="5", functor MEANS
If <role descr="medium" n="5", functor DIFF.

CAU → CAUS
PRP → AIM
MNR → MANN (but whenever the frame contains an ArgA it should be EFF).

All slots corresponding to arguments with the original "agentive-subject" (Arg0) got the Actor functor, unless the ArgA was present. The ArgA always got the Actor-functor. The Arg1 always got the Patient functor, unless the ArgA was present etc.

The annotation work has proved that Arg-Ms, which roughly correspond to free modifications, were assigned quite correctly. Yet the re-labeling rules originally assumed that the frames in Pre-EngValLex would be defined by the Args used in the example sentences. The more recent technically-motivated decision to use the entire list of Args from the roleset beginning and to ignore the example-sentence annotation made the set of rules less powerful. E.g. the manual correction revealed that Actor had been systematically interchanged with Patient in all intransitive sentences that had a transitive counterpart within the same roleset.

However, it was clear already at the beginning that the functors would have to be manually corrected anyway. The rules were only meant to save the annotators' typing time, and they proved powerful enough to serve this particular purpose. Therefore no evaluation was performed.

3.2.2. Links between the PropBank-Lexicon and Pre-EngValLex

Pre-EngValLex seeks to preserve as much original information of the PropBank-Lexicon as possible. Therefore each preliminary valency frame has its own ID saying which roleset it used to belong to as its example sentence. This ID remains even if the given preliminary valency frame is later merged with another preliminary valency frame during the manual adjustment. This ID ensures that each preliminary valency frame bears the same links as the original PropBank-roleset, including the links to other lexical sources as well as the links to the original corpus annotation.

Each functor within the given frame is linked to the original Arg or "role" of its original PropBank roleset. NB that the functors are linked to the *list of roles at the beginning of each roleset*, not to the corresponding Args in the annotation of the corresponding example sentences, which is not preserved in Pre-EngValLex. This linking is of special importance for transferring the alternation information from the PropBank-Lexicon into EngValLex. To illustrate the linking policy, lets take up the Causative-Inchoative alternation case again: the (made-up) pair of sentences *John opened the door* and *The door opened* give two preliminary valency frames of the verb *to open*; the former transitive and the latter intransitive. The original roleset had the following Args: Arg0 (the agent opening the patient) and Arg1 (the thing opened). The transitive original example sentence had both the Arg0 (*John*) and Arg1 (*the door*), while the intransitive example sentence only had Arg1 (*the door*). According to our linguistic conventions set before, the frames will be kept separate in EngValLex, as they are now in Pre-EngValLex. As noted above, the valency theory of FGD requires the first argument of a verb to be ACT. Therefore, in the Causative-Inchoative Alternation pair, both the functor PAT in the transitive frame and the functor ACT in the intransitive frame will be linked to the original Arg1. As the annotation seeks to treat all major alternations as consequently as possible, this shifting of PAT to ACT will be characteristic of this type of alternation, as shifts of other functors will be characteristic of other alternation types.

3.2.3. Rearrangement of the XML-Files for Manual Editing

The editing tool FrameEditor opens a file with a roleset name, in which all rolesets (verb senses) called one particular name are gathered, no matter which lemma they actually belong to. This rearrangement was made in order to ensure that the annotators keep the consistency of the original PropBank-verb-class annotation.

3.3. EngValLex (Manual Refinement)

3.3.1. Transforming Rolesets into Valency Frames

Pre-EngValLex split the PropBank-rolesets into preliminary FGD-valency frames consisting of one example sentence each. The first task of the annotators is to merge certain types of frames. The frames are merged when:

1. the sentences in question have the same surface syntax structure.
2. a pair of sentences in question belong to an alternation whose realizations are to be merged in compliance with our linguistic conventions. E.g. a sentence with Unspecified Object is to be merged with a Transitive sentence, and so is a sentence with a Reciprocal Object, etc.

In some regular cases, complementations regarded as Args by the PropBank-Lexicon would neither have been classified as inner participants nor as obligatory free modifications, i.e. would not have been regarded as members of the frame in the strict sense, and therefore should not be listed in the valency lexicon. In order not to lose this part of the ready-made PropBank annotation, we included them as "typical free modifications" (indicated by a question mark in front of the name of the given functor), which has been an approved practice of Vallex.

3.3.2. Preserving the Alternation Information

Linguistic rules were set for the commonest pairs of alternation realizations as to whether EngValLex would list each realization as a separate frame, or whether both should be merged into one frame. When annotators recognize a sentence pair as an alternation pair or a single sentence as an unpaired realization of a particular alternation, they are supposed to follow the appropriate rule.

The following situations can occur in EngValLex during the processing of the example sentences (i.e. the frames automatically generated from a PropBank roleset):

1. The frames are merged as the sentences only give variants of surface representations, irrelevant to the verb frame. E.g.: Unspecified Object, Instructional Imperative and Reciprocity (see 3.3.3 – 3.3.5).
2. The frames are merged as either alternation realization makes an Arg to an optional free modification (FGD). This makes it fit into the frame of the other alternation realization. E.g. Instrumental Subject (see 3.3.6).
3. One of the alternation realizations is regarded as a derivation of the other. Its tectogrammatical tree structure looks different from the lexicon frame but it refers back to it. Rules are stored to generate such trees from the lexicon frames. E.g. Induced Action, Middle Alternation, Location Subject Alternation (see 3.3.7).
4. The frames remain split and each acquires functors of its own. E.g.: Causative-Inchoative alternation, Substance/Source emission alternation (see 3.3.8).

Selected examples of alternation resolutions are given below.

3.3.3. Unspecified Object

The frames are merged. The Unspecified Object (the type *John was eating*) is not captured by the lexicon but only by the data. The Patient of the normally transitive verb acquires the tectogrammatical lemma *&Gen;* (Generalized inner participant).

3.3.4. Instructional Imperative

The frames are merged. The Instructional Imperative (the type *Bake in the oven for 30 minutes*) is treated as ellipsis in the data; the frame is completed with a PAT-node whose tectogrammatical lemma is either copied from elsewhere in the text (if the object has been explicitly mentioned before or after), or newly generated and provided with a substitutional tectogrammatical lemma.

3.3.5. Reciprocity

The frames are merged. Reciprocity is also captured only in the data. E.g. *to meet* has the frame ACT PAT, which can be filled-in with the following surface syntax representations:

A met B.
A. met with B[4]*.*
A and B met.

In *A met B.* and *A. met with B* A is ACT and B is PAT. When *A and B met.* occurs in the data, A and B are joined together by coordination. The coordination node acquires the functor ACT. The functor PAT is to be added to make the frame complete. It gets the tectogrammatical lemma *&Rcp;* (Reciprocal).

3.3.6. Instrumental Subject

The frames are merged. The type *John broke the window with a hammer* fits into the same frame as *The hammer broke the window.* As noted above, agentivity is not an issue in FGD. The PropBank- Lexicon regards *John* as Arg0, *window* as Arg1 and *hammer* as Arg2 – "instrument". EngValLex, on the other hand, will treat instrument as an optional free modification, which is not part of the valency frame in the strict sense. The information from the PropBank-Lexicon will though be preserved by introducing "typical free modifications".

3.3.7. Induced Action, Middle Alternation, Location Subject Alternation

Constructions like *Sylvia jumped the horse across the fence, The father burped the baby, Crystal breaks easily* and *This room sleeps five people/We sleep five people in this room* refer to the basic frames of the respective verbs, though they are annotated in a different way in the data. Hence, *Sylvia jumped the horse across the fence* refers to the frame of *The horse jumped across the fence.*, *The father burped the baby* refers to the frame of *The baby burped.*, and *This room sleeps five people/We sleep five people in this room* refer to the frame of *Five people sleep in this room.*, which are regarded as basic. Typically, the Patients in the Induced Action and Location Subject alternations (*horse/baby, people*) derive from Actors in the basic frames.

3.3.8. Causative–Inchoative/Intransitive Alternation, Substance - Source Emission Alternation

Realizations of these alternations get each its own frame. The Inchoative *The door opened* has only ACT, while the Causative *John opened the door* has ACT and PAT. In the Substance/Source pair, the type *The sun radiates heat* will have ACT and PAT, while *Heat radiates from the sun* will have ACT PAT and DIR1 (direction "where from", obligatory free modification).

3.4. Surface Syntax Representations of Valency Slots

The possible surface representations are noted for each particular slot by means of macros based on the PTB tagset. The macros say e.g. that a particular slot can be represented by a noun or a personal pronoun or an infinitival clause or a subordinate clause (with a particular subjunction), etc. Slots that comprise parts of phrasemes also indicate a restriction in their lemmas.

4. Work in Progress and Future Prospects

The manual adjustment of the preliminary valency frames is being finished at the moment. The processing of the approx. 2000 roleset-name xml-files took about 3 months with two annotators, out of which one left the task to prepare the surface-representation description when having annotated 30% of the data. The xml-files will be regrouped back to lemma-files, and the lexicon will be examined and corrected where needed.

The next step will comprise the semi-automatic annotation of surface.representations of the respective valency slots. The commonest alternatives (e.g. "noun", "personal pronoun", etc.) will be filled in automatically to save the annotators' typing time and to decrease the risk of typing errors. The annotators will prototypically just have

[4] A met B. and A. met with B also represent the "With" Preposition Drop alternation pair. Yet in this case they both are the counterpart of the Reciprocal sentence A and B met. in the Understood Reciprocal Object Alternation pair.

to fill in prepositions into clause-macros. During the experimental course of this step, the morphosyntactic alternatives of most slots have been checked and completed according to an up-to-date dictionary of English (Rundell et al., 2002). The practice only will show whether this ambition is not too time-consuming.

The PEDT corpus is being prepared for annotation in the meantime. EngValLex will be automatically mapped onto PEDT before the manual corpus annotation has been launched. Manual as well as (semi-)automatic corrections of valency frames will go on during the general manual annotation of PEDT.

5. Conclusion

Building EngValLex is nothing but one step in the project of the Prague Czech-English Dependency Treebank 2.0. Nevertheless, valency is a strongly-focused phenomenon within the FGD-framework, and therefore we have been paying it special attention and care.

We are happy to have access to an already existing relevant lexical source. Though our work is still in progress, it proves undoubtedly worth a reasonable effort to make the PropBank-Lexicon work together with dependency-based frames and to preserve the links between the lexicon and the data in their new, dependency-based shape. Apart from saving human and financial resources involved in the corpus annotation work, the search for meeting points between two different conceptions of valency together with the hands-on experience of living data yield interesting linguistic insights, which possibly reach beyond the scope of one particular language.

6. Acknowledgements

This work has been supported by GA-UK 489/2004, IS-1ET101120413, IS-1ET201120505 and GACR 405-06-0589.

References

Čmejrek, M. et al. (2005). Prague Czech-English Dependency Treebank: Resource for Structure-based MT, In Proceedings of the 10th EAMT Conference, pp. 73-78 (eds. Hutchins, John and Kis, Balázs and Prószéky, Gábor), Budapest, Hungary, May 30-31.

Hajič, J. (2005). Complex Corpus Annotation: The Prague Dependency Treebank, In M. Šimková (Ed.) *Insight into Slovak and Czech Corpus Linguistics*. Bratislava: Veda. pp. 54-73.

Hajič, J. et al. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation, In Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 57--68 (eds. Nivre, Joakim//Hinrichs, Erhard), Vaxjo, Sweden.

Hajič, J. (2002). Tectogrammatical Representation: Towards a Minimal Transfer in Machine Translation. In: R. Frank (Ed.): Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), Venezia. pp. 216-226..

Kipper, K. et al. (2002). Extending PropBank with VerbNet Semantic Predicates. In Workshop on Applied Interlinguas, held in conjunction with AMTA. Tiburon, CA, October 2002. URL <http://www.cis.upenn.edu/group/verbnet/publications.html> [quoted 2004-10-19]

Levin, B. (1993) *English Verb Classes and Alternations*. University of Chicago Press.

Lopatková, Markéta (2003). Valency in the Prague Dependency Treebank: Building the Valency Lexicon, PBML 79-80, pp. 37--60, Prague.

Lopatková, M. et al. (2003). VALLEX 1.0 Valency Lexicon of Czech Verbs, ÚFAL-CKL Technical Report TR-2002-15. Prague.

Mitchell P. M. et al. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*

Palmer, M. et al. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71-106

Palmer, M. et al. (2004). Proposition Bank I. LDC2004T14, ISBN: 1-58563-304-6, Sep 01 2004.

Panevová, J. (2002). Sloveso: centrum věty, valence: centrální pojem syntaxe, *In Proceedings of Aktuálne otázky súčasnej syntaxe*, pp. 73-77, Budmerice, Slovakia, Nov. 7-8.

Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Prague:Academia.

Panevová, J. (1974-75). On Verbal Frames in Functional Generative Description. Part I, PBML 22,pp. 3-40, Part II, PBML 23, pp. 17-52.

PDT 2.0 URL <http://ufal.mff.cuni.cz/pdt2.0/> [quoted 2005-11-29]

PropBank Annotation Guidelines. Version 3. February 22, (2002). URL<http://www.cis.upenn.edu/,ace/propbank-guidelines-feb02.pdf > [quoted 2004-09-03]

Rundell, M. et al. (2002). *Macmillan English Dictionary for Advanced Learners. International Student Edition.* Macmillan Education.

Sgall, P. et al. (2004). Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In: A. Meyers (ed.): *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, Association for Computational Linguistics, Boston, Massachusetts, USA, pp. 32-38.

Sgall, P. et al. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht:Reidel Publishing Company and Prague:Academia

Straňáková-Lopatková, M, Žabokrtský, Z. (2002). Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In: LREC 2002, Proceedings, vol.III., pp. 949-956.