Constructing an English Valency Lexicon[*]

Jiří Semecký, Silvie Cinková
Institute of Formal and Applied Linguistics Affiliation
Malostranské náměstí 25
CZ11800 Prague 1
Czech Republic
(semecky,cinkova)@ufal.mff.cuni.cz

Abstract

This paper presents the English valency lexicon EngValLex, built within the Functional Generative Description framework. The form of the lexicon, as well as the process of its semi-automatic creation is described. The lexicon describes valency for verbs and also includes links to other lexical sources, namely PropBank. Basic statistics about the lexicon are given.

The lexicon will be later used for annotation of the Wall Street Journal section of the Penn Treebank in Praguian formalisms.

1 Introduction

The creation of a valency lexicon of English verbs is part of the ongoing project of the Prague English Dependency Treebank (PEDT). PEDT is being built from the Penn Treebank - Wall Street Journal section by converting it into dependency trees and providing it with an additional deep-syntactic annotation layer, working within the linguistic framework of the Functional Generative Description (FGD)(Sgall et al., 1986).

The deep-syntactic annotation in terms of FGD pays special attention to valency. Under valency we understand the ability of lexemes (verbs, nouns, adjectives and some types of adverbs) to combine with other lexemes. Capturing of valency is profitable in Machine Translation, Information Extraction and Question Answering since it enables the machines to correctly recognize types of events and their participants even if they can be expressed by many different lexical items. A valency lexicon of verbs is inevitable for the project of the Prague English Dependency Treebank as a supporting tool for the deep-syntactic corpus annotation.

We are not aware of any lexical source from which such a lexicon could be automatically derived in the desired quality. Manual creation of gold-standard data for computational applications is yet very time-consuming and expensive. Having this in mind, we decided to adapt the already existing lexical source PropBank (M. Palmer and D. Gildea and P. Kingsbury, 2005) to FGD, making it comply with the structure of the original Czech valency lexicons VALLEX (Žabokrtský and Lopatková, 2004) and PDT-VALLEX (J. Hajič et al., 2003), which have been designed for the deep-syntactic annotation of the Czech FGD-based treebanks (The Prague Dependency Treebank 1.0 and 2.0) (J. Hajič et al., 2001; Hajič, 2005). Manual editing follows the automatic procedure. We are reporting on a work that is still ongoing (which is though nearing completion). Therefore this paper focuses on the general conception of the lexicon as well as on its technical solutions, while it cannot give a serious evaluation of the completed work yet.

The paper is structured as follows. In Section 2, we present current or previous related projects in more detail. In Section 3, we introduce the formal structure of the EngValLex lexicon. In Section 4, we describe how we semi-automatically created the lexicon and describe the annotation tool. Finally in Section 5, we state our outlooks for the future development and uses of the lexicon.

[*] The research reported in this paper has been partially supported by the grant of Grant Agency of the Czech Republic GA405/06/0589, the project of the Information Society No. 1ET101470416, and the grant of the Grant Agency of the Charles University No. 372/2005/A-INF/MFF.

2 Valency Lexicon Construction

2.1 FGD

The Functional Generative Description (FGD) (Sgall et al., 1986) is a dependency-based formal stratificational language description framework that goes back to the functional-structural Prague School. For more detail see (Panevová, 1980) and (Sgall et al., 1986). The theory of FGD has been implemented in the Prague Dependency Treebank project (Sgall et al., 1986; Hajič, 2005).

FGD captures valency in the underlying syntax (the so-called tectogrammatical language layer). It enables listing of complementations (syntactically dependent autosemantic lexemes) in a valency lexicon, regardless of their surface (morphosyntactic) forms, providing them with semantic labels (functors) instead. Implicitly, a complementation present in the tectogrammatical layer can either be directly rendered by the surface shape of the sentence, or it is omitted but can be inferred from the context or by common knowledge. A valency lexicon describes the valency behavior of a given lexeme (verb, noun, adjective or adverb) in the form of valency frames.

2.2 Valency within FGD

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). Free modifications have semantic labels (there are some more than 40 in PDT) and they are distributed according to semantic judgments of the annotators. FGD introduces five inner participants. Unlike free modifications, inner participants cannot be repeated within one frame. They can be obligatory as well as optional (which is to be stated by the judgment on grammaticality of the given sentence and by the so-called dialogue test, (Panevová, 1974 75)). 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). Syntactic criteria are used to identify the first two participants ACT and PAT (”shifting”, see (Panevová, 1974 75)). The other inner participants are identified semantically; i.e. a verb with one inner participant will have ACT, a verb with two inner participants will have ACT and PAT regardless the semantics and a verb with three and more participants will get the label assigned by the semantic judgment.

2.3 The Prague Czech-English Dependency Treebank

In order to develop a state-of-the-art machine translation system we are aiming at a high-quality annotation of the Penn Treebank data in a formalism similar to the one developed for PDT. When building PEDT we can draw on the successfully accomplished Prague Czech-English Dependency Treebank 1.0 (J. Cuřín and M. Čmejrek and J. Havelka and J. Hajič and V. Kuboň and Z. Žabokrtský, 2004) (PCEDT).

PCEDT is a Czech-English parallel corpus, consisting of 21,600 sentences from the Wall Street Journal section of the Penn Treebank 3 corpus and their human translations to Czech. The Czech data was automatically morphologically analyzed and parsed by a statistical parser on the analytical (i.e. surface-syntax) layer. The Czech tectogrammatical layer was automatically generated from the analytical layer. The English analytical and tectogrammatical trees were derived automatically from the Penn Treebank phrasal trees.

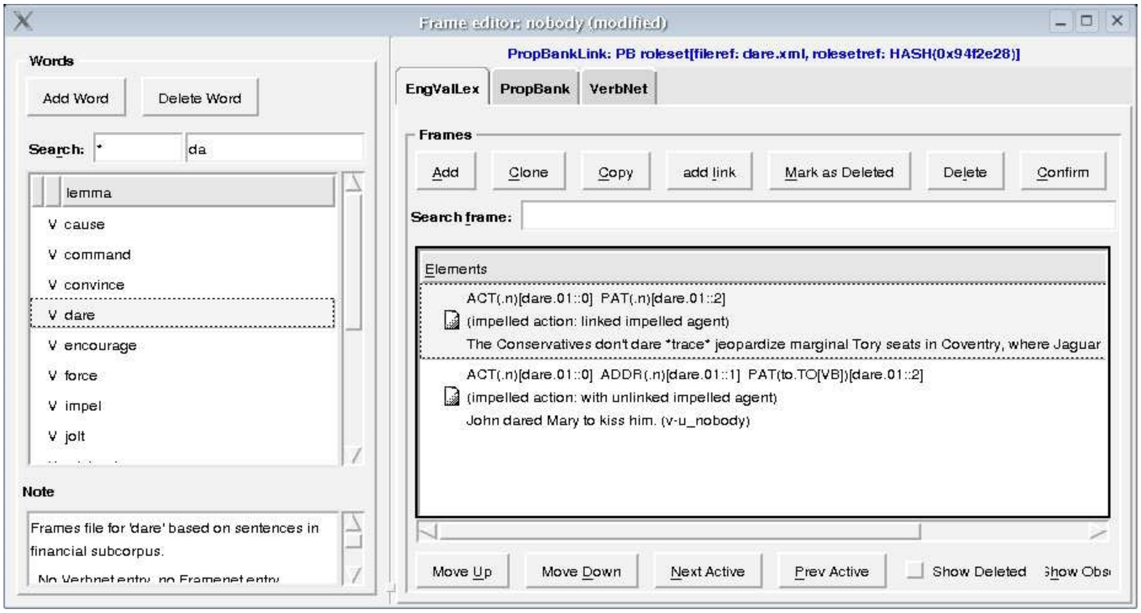
2.4 The Prague English Dependency Treebank

The Prague English Dependency Treebank (PEDT) stands for the data from Wall Street Journal section of the Penn Treebank annotated in the PDT 2.0 shape. EngValLex is a supporting tool for the manual annotation of the tectogrammatical layer of PEDT.

3 Lexicon Structure

On the topmost level, EngValLex consists of **word** entries, which are characterized by lemmas. Verbs with a particle (e.g. *give up*) are treated as separate word entries.

Each word entry consists of a sequence of **frame** entries, which roughly correspond to individual senses of the word entry and contain the valency information.



PropBankLink: PB roleset[fileref: dare.xml, rolesetref: HASH(0x94f2e28)]

Search: * da

lemma
V cause
V command
V convince
V dare
V encourage
V force
V impel
V jolt

Search frame:

ACT(.n)[dare.01::0] PAT(.n)[dare.01::2]
(impelled action: linked impelled agent)
The Conservatives don't dare *trace* jeopardize marginal Tory seats in Coventry, where Jaguar

ACT(.n)[dare.01::0] ADDR(.n)[dare.01::1] PAT(to.TO[VB])[dare.01::2]
(impelled action: with unlinked impelled agent)
John dared Mary to kiss him. (v-u_nobody)

Frames file for 'dare' based on sentences in financial subcorpus.
No Verbnet entry, no Framenet entry

Figure 1: EngValLex editor: the list of words and frames

Each frame entry consists of a sequence of **valency slots**, a sequence of **example sentences** and a textual note. Each valency slot corresponds to a complementation of the verb and is described by a **tectogrammatical functor** defining the relation between the verb and the complementation, and a **form** defining the possible surface representations of the functor. Valency slots can be marked as optional, if not, they are considered to be obligatory.

The form is listed in round brackets following the functor name. Surface representations of functors are basically defined by combination of morphological tags and lemmas. Yet to save annotators' effort, we have introduced several abbreviations that substitute some regularly co-occurring sequences. E.g. the abbreviation **n** means '*noun in the subjective case*' and is defined as follows:

```
NN:NNS:NP:NPS
```

meaning one of the Penn Treebank part-of-speech tags: *NN*, *NNS*, *NP* and *NPS* (colon delimits variants). Abbreviation might be defined recursively.

Apart from describing only the daughter node of the given verb, the surface representation can describe an entire analytical subtree whose topmost node is the daughter of the given verb node. Square brackets are used to indicate descendant nodes. Square brackets allow nesting to indicate the dependency relations among the nodes of a given subtree. For example, the following statement describes a particle *to* whose daughter node is a verb.

```
to.TO[VB]
```

The following statement is an example of a definition of three valency slots and their corresponding forms:

```
ACT(.n)  PAT(to.TO[VB])
 LOC(at.IN)
```

The ACT (Actor) can be any noun in the subjective case (the abbreviation *n*), the PAT (Patient) can be a particle *to* with a daughter verb, and the LOC (Locative) can be the preposition *at*.

Moreover, EngValLex contains links to external data sources (e.g. lexicons) from words, frames, valency slots and example sentences.

The lexicon is stored in an XML format which is similar to the format of the PDT-VALLEX lexicon used in the Prague Dependency Treebank 2.0.

4 Creating the Lexicon

The lexicon was automatically generated from PropBank using XSLT templates. Each PropBank example was expanded in a single frame in the destination lexicon. When generating the lexicon, we have kept as many back links to PropBank as possible. Namely, we stored links from frames to Propbank rolesets, links from valency slots to PropBank arguments and links from examples to PropBank examples. Rolesets were identified by the roleset *id* attribute. Arguments were identified by the roleset *id*, the name and the function of the role. Examples were identified by the roleset *id* and their name.

After the automatic conversion, we had 8,215 frames for 3,806 words.

Tectogrammatical functors were assigned semi-automatically according to hand-written rules, which were conditioned by PropBank arguments. It was yet clear from the beginning that manual corrections would be necessary as the relations of Args to functors varied depending on linguistic decisions[1].

The annotators were provided with an annotation editor created on the base of the PDT-VALLEX editor. Apart from interface for editing EngValLex, the tool contains integrated viewers of PropBank and VerbNet, which allows offline browsing of the lexicons. Those viewers can be run as a stand-alone application as well and are published freely on the web[2]. The editor allows the annotator to create, delete, and modify word entries, and frame entries. Links to PropBank can be set up, if necessary.

Figure 1 displays the main window of the editor. The left part of the window shows list of words. The central part shows the list of the frames concerning the selected verb.

For the purpose of annotation, we divided the lexicon into 1,992 files according to the name of PropBank rolesets (attribute *name* of the XML element *roleset*), and the files are annotated separately. When the annotation is finished, the files will be merged again. Currently, we have about 80% of the lexicon annotated, which already contains the most difficult cases.

5 Outlook

We have annotated the major part of EngValLex. In the final version, a small part of the lexicon will be annotated by a second annotator in order to determine the inter-annotator agreement.

The annotation of the Prague English Dependency Treebank on the tectogrammatical level will be started soon and we will use EngValLex for assigning valency frames to verbs. The annotation will be based on the same theoretical background as the Prague Dependency Treebank.

Due to the PropBank links in EngValLex, we will be able to automatically derive frame annotation of PEDT from PropBank annotation of the Penn Treebank.

As the Wall Street Journal sentences are manually translated into Czech, we will be able to obtain their Czech tectogrammatical representations automatically using state-of-art parsers.

A solid platform for testing Czech-English and English-Czech machine translation will be given. In the future we will also try to improve the translation by mapping the Czech PDT-ValLex to the English EngValLex.

References

J. Hajič, 2005. *Complex Corpus Annotation: The Prague Dependency Treebank*, pages 54–73. Veda Bratislava, Slovakia.

J. Cuřín and M. Čmejrek and J. Havelka and J. Hajič and V. Kuboň and Z. Žabokrtský. 2004. Prague Czech-English Dependency Treebank Version 1.0. (LDC2004T25).

J. Hajič et al. 2001. Prague Dependency Treebank 1.0. LDC2001T10, ISBN: 1-58563-212-0.

J. Hajič 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*, volume 9.

M. Palmer and D. Gildea and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71.

J. Panevová. 1974–75. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics (PBML)*, 22, pages 3–40, Part II, PBML 23, pages. 17–52.

J. Panevová. 1980. *Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech sentence].* Academia, Prague, Czech Rep.

J. Panevová. 2002. Sloveso: centrum věty; valence: centrální pojem syntaxe. In *Aktuálne otázky slovenskej syntaxe*, pages x1—x5.

P. Sgall, E. Hajičová, and J. Panevová. 1986. The Meaning of the Sentence in its Semantic and Pragmatic Aspects. *Academia, Prague, Czech Republic/Reidel Publishing Company, Netherlands.*

Z. Žabokrtský and M. Lopatková. 2004. Valency Frames of Czech Verbs in VALLEX 1.0. In *Frontiers in Corpus Annotation. Proceedings of the Workshop of the HLT/NAACL Conference*, pages 70–77, May 6, 2004.

[1] E.g. Arg0 typically corresponds to ACT, and Arg1 to PAT when they co-occur. Yet, a roleset including an inchoative sentence (*The door.*ARG1 *opened.*) and a causative sentence (*John.*Arg0 *opened the door.*Arg1) will be split into two FGD frames. The causative frame will keep Arg0→ACT and Arg1→PAT whereas the inchoative will get Arg1→ACT.

[2] `http://ufal.mff.cuni.cz/∼semecky/software/{propbank|verbnet}viewer/`