Advanced Searching in the Valency Lexicons Using PML-TQ Search Engine*

Eduard Bejček, Václava Kettnerová, and Markéta Lopatková

Charles University in Prague, Institute of Formal and Applied Linguistics
{bejcek,kettnerova,lopatkova}@ufal.mff.cuni.cz

Abstract. This paper presents a sophisticated way to search valency lexicons. We provide a visualization of lexicons with such built-in searching that allows users to draw sophisticated queries in a graphical mode. We exploit the PML-TQ, a query language based on the tree editor TrEd. For demonstration purposes, we focus on VALLEX and PDT-VALLEX, two Czech valency lexicons of verbs. We propose a common lexicon data format supported by PML-TQ. This format offers easy viewing both lexicons, parallel searching and interlinking them. The proposed method is universal and can be used for other hierarchically structured lexicons.

Keywords: Valency lexicon, searching, browsing, linking lexicons.

1 Motivation

Valency lexicons play a crucial role in modern theoretical and computational linguistics. The richer information they provide, the more sophisticated tools are needed for using them, namely for searching them. Search and visualization tools allow users to get useful information from the lexicons, not only for the purposes of theoretical study, but also for the lexicographical aims, e.g., providing frequency analysis, modifying annotation schemes of the lexicon, or consistency checking. However, on-line versions of current valency lexicons do not commonly allow a researcher to ask more complicated, complex queries.

In this paper, we introduce a sophisticated way to search valency lexicons. For demonstration purposes, we focus on VALLEX and PDT-VALLEX, two Czech valency lexicons of verbs. These lexicons represent a collection of manually linguistically annotated data resulting from an attempt at a formal description of valency frames of Czech verbs. Both lexicons are closely related to the Prague Dependency Treebank (PDT henceforth [1,2]) but they capture slightly different types of information and their data structures are different.

The work presented here has two goals:

1. to transform valency lexicons into a common format; this allows a user to search the lexicons in a parallel way and thus facilitate their interlinking at the level of lexical units in the future and

* This work has been supported by projects MSM 0021620838 and LC536 of the Ministry of Education and grants GAUK 4200/2009 and partially GAUK 52408/2008 of the Grant Agency of Charles University in Prague.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 51–58, 2010.

2. to provide visualization of VALLEX and PDT-VALLEX with such built-in searching that allows users to formulate complex queries in a user-friendly way by drawing their graphical representation.

We exploit the PML-TQ, a query language and search engine designed for querying annotated linguistic data [3], which is based on the TrEd toolkit [4]. There are three important reasons for adopting TrEd and PML-TQ: (i) PML-TQ incorporates a powerful query language useful for complex data and offers graphical query representation, (ii) tree editor TrEd provides us with customizable visualization of richly structured data and makes it possible to visualize query results as well, and (iii) TrEd data format proved to be a suitable common representation for both lexicons and for links between them.

As the PDT-VALLEX lexicon (a part of PDT 2.0 data) can be already searched using PML-TQ we transform the VALLEX lexicon into the format supported by this search engine.

Related Work. Let us mention some of the lexicons providing valency information and their searching interfaces.

More than 960 semantic frames can be browsed in *FrameNet* given a name of the frame or a lemma to search for; in addition, FrameGrapher visualizes relations between semantic frames and their frame elements.[1] The *VerbNet* project maps PropBank verb types to their corresponding Levin classes; on-line search tool facilitates searching only for verb lemmas; VerbNet viewer 'Inspector' can parse a VerbNet data file and print specified attributes for classes.[2] Project *SemLink* combines four lexical resources, PropBank, VerbNet, FrameNet, and WordNet; it supports lemma and semantic class on-line search through Unified Verb Index (UVI).[3] The *Corpus Pattern Analysis* shows patterns with which a verb is associated; it can be browsed only for a given lemma.[4] Verbs in another Czech valency lexicon, *VerbaLex*, can be sorted (similarly to VALLEX) by alphabet, verb roles, morphemic forms, verb classes etc.[5]

Some of these lexicons are already interlinked, like UVI for English (interlinked on the level of individual lexical units). Our long-term goal is to link the VALLEX and PDT-VALLEX lexicons on the level of individual lexical units.

Although our current effort focuses only on searching the VALLEX and PDT-VALLEX lexicons, the underlying search engine can be easily adopted for any other lexicons with structured lexical entries.

2 Two Valency Lexicons

In this section, we provide a basic description of the valency lexicons of Czech verbs, PDT-VALLEX and VALLEX. Both these lexicons take the Functional Generative Description (FGD [9]) as their theoretical background. In Section 2.1, we focus on the differences between their data formats.

[1] `http://framenet.icsi.berkeley.edu/FrameGrapher/` [5]

[2] `http://verbs.colorado.edu/verb-index/vn/reference.php` [6]

[3] `http://verbs.colorado.edu/semlink/`

[4] `http://deb.fi.muni.cz/pdev/` [7]

[5] `http://nlp.fi.muni.cz/verbalex/htmlDEMO/` [8]

PDT-VALLEX. PDT-VALLEX (see esp. [10,11]) stores the information on the valency frames of Czech verbs (and also of some nouns, adjectives, and adverbs), which occur at least once in PDT 2.0. Valency frames in PDT-VALLEX are linked with the occurrences of verbs in PDT 2.0. One of the main purposes of building PDT-VALLEX was to ensure the data consistency of PDT.

VALLEX 2.5. The VALLEX lexicon (see esp. [12]) aims at describing valency behavior of verbs in each of their senses, i.e., at providing analysis of whole verb lexemes. In addition to valency frames, further syntactic information is rendered there, esp. the information related to the surface manifestation of verbal valency (e.g., reciprocity, reflexivity, grammatical control), and syntactico-semantic class for a substantial subset of verbs.

In VALLEX, the concept of a *lexeme* plays a crucial role – aspectual counterparts[6] are treated within a single lexeme, which may be therefore represented by more than one lemma. Moreover, a particular lemma may have different orthographic variants. A lexeme associates individual *lexical units* (LUs) representing different verb meanings. The concept of lexeme can be exemplified by the verbs *započítávat*^{impf} and *započítat*^{pf} 'to count' as aspectual counterparts and the verb *započíst*^{pf} as an orthographic variant of the verb *započítat*^{pf}. In VALLEX, all these verbs are treated within one lexeme (Figure 1, left column). Let us mention at least the most important reasons for such convention:

- theoretical adequacy: aspectual counterparts have (in principle) the same meaning; in the FGD theory, they are considered as different forms of one verb lexeme;
- compact representation: aspectual counterparts prototypically share the set of lexical units describing their valency characteristics, see LU1 in Figure 1; thus this representation effectively reduce the redundant information in the lexicon;
- convenience for human users when searching the lexicon.

2.1 Data Formats

Although both VALLEX and PDT-VALLEX are stored in XML format, their data formats differ as the lexicons are developed separately and they contain slightly different types of information. In this section, we discuss and exemplify these differences in more detail. The full format description can be found in [13] for the VALLEX format and [14, Section 6.2] for the PDT-VALLEX format.

VALLEX format. VALLEX, version 2.5, is stored in a complex format that reflects the concept of lexeme associating aspectual counterparts of verbs.

The complicated XML format makes searching in VALLEX format rather difficult from the technical point of view: for each lemma, it is necessary to identify correctly the relevant XML elements. For instance, some lexical units are ascribed to all lemmas, see LU1, whereas others are assigned exclusively to some of them, see LU2 ascribed only to the lemma *započítat*^{pf} (Figure 1).

[6] Roughly speaking, perfective and imperfective aspectual counterparts are verbs with the same meaning, which differs in presenting the event either as completed, or as ongoing, like e.g. *pokrýt*^{pf} and *pokrývat*^{impf} as in 'he covered the floor with the carpet' and 'he was covering the floor with the carpet'. Aspectual counterparts usually form pairs, but also triples or even quadruples may appear.

PDT-VALLEX format. In PDT-VALLEX, neither the aspectual counterparts nor the orthographic variants are clustered together. Therefore the XML format of the lexicon is much simpler in comparison with the format of VALLEX. E.g., unlike VALLEX, the verbs *započítávat*, *započítat*, and *započíst* 'to count' are described by three separate word entries in PDT-VALLEX. As a result, these word entries are characterized by the identical set of valency frames.

To facilitate comparing, parallel viewing, and interlinking the VALLEX and PDT-VALLEX data, it is necessary to have a common data representation. This representation must be powerful enough to store different type of information from both lexicons. For this purposes, we exploit TrEd toolkit and its native PML format, which is a part of the pmltq extension to the tree editor TrEd.

3 Exploiting TrEd Toolkit for Valency Lexicons

The Tree Editor (TrEd) is a graphical editor that was primarily designed as an annotation tool for the syntactic annotation of the PDT. However, the editor can also be used for data viewing and for advanced data searching. TrEd supports any tree-like structure (which every XML exactly is), thus it is possible to use it for our valency lexicons as well.

TrEd supports an XML-based format called PML (Prague Markup Language [15]). PML data are described in a form of a PML-schema (similarly as DTD describes XML data). In principle, a PML-schema can be obtained automatically from a DTD for an XML document. In addition, it is necessary to specify 'PML roles'. These roles identify XML elements that ought to be visualized, those XML elements that serve as tree nodes etc. Lastly, there is a stylesheet in TrEd that defines the style, color and layout of nodes, edges and labels of a tree.

The PML data format makes it possible to exploit the PML Tree Query language (PML-TQ[3]), a search tool designed for linguistically annotated data in PML format.

3.1 Common Lexicon Format `vallex_pml`

The first task is to prepare common representation of both lexicons. We use the PML format derived from the PDT-VALLEX format; it is referred to as `vallex_pml` here.

VALLEX format to `vallex_pml`. To transform VALLEX 2.5 data into the `vallex_pml` format, each lexeme from VALLEX have to be split – separate *verb entries* must be created for each lemma. To each verb entry, an appropriate set of *frame entries* describing valency frames (and some other syntactic information) has to be assigned, see Figure 1.

However, the information on aspectual counterparts and their corresponding valency frames belong to the core information stored in the lexicon. To retain this information, the resulting verb entries and corresponding frame entries are interlinked in the target format: each verb entry contains a reference to relevant verb entries for aspectual counterpart(s) (if applicable); similarly, each frame entry contains a reference to corresponding frame entries, see Figure 1. Orthographic variants of lemmas are treated in the same way.

By this splitting and linking, we overcome the aforementioned difficulty with searching in the complex VALLEX format.

PDT-VALLEX format to `vallex_pml`. Though PML is not a native format for PDT-VALLEX, it can be straightforwardly transformed into this format. As a result, PDT-VALLEX is compatible with PML-TQ and TrEd toolkit.

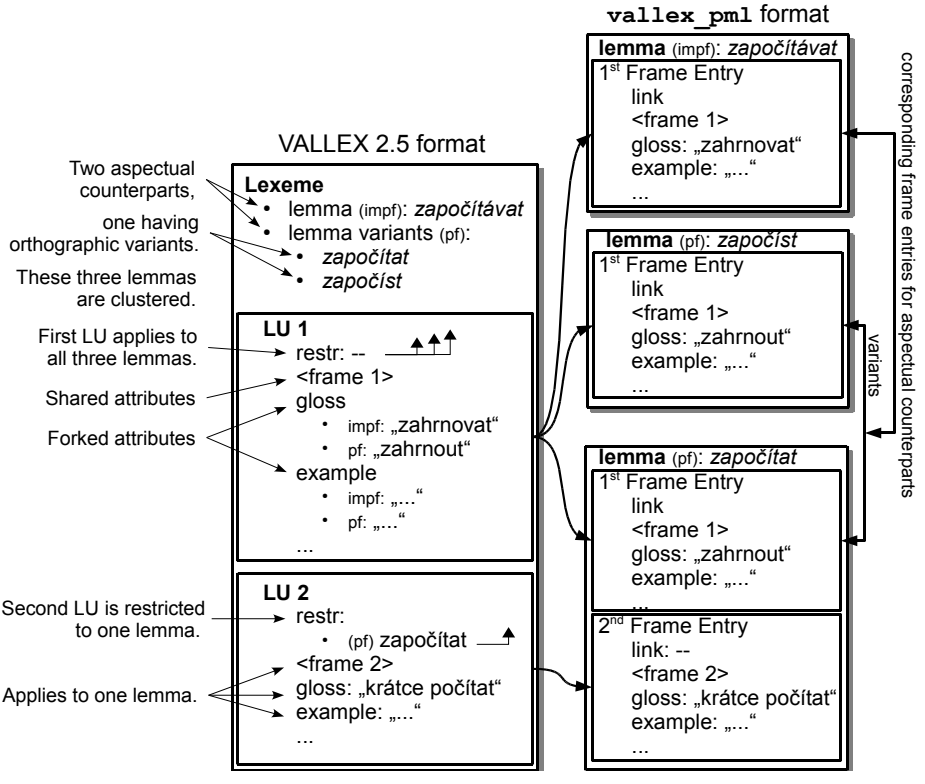

Fig. 1. The transformation of VALLEX into the `vallex_pml` format. The lexeme represented in VALLEX 2.5 by the lemmas *započítávat*^{impf}, *započítat/započístpf* ‘to count’ associated with two lexical units LU1 and LU2 is schematically displayed in the left column. The right column shows three verb entries for these lemmas and the relevant frame entries for each of these lemmas in the `vallex_pml` format.

After converting both valency lexicons into `vallex_pml` format, they can be loaded into TrEd and three tasks are easier to process: viewing and editing both lexicons (3.2), parallel searching the lexicons (3.3), and linking them together (4).

3.2 Viewing and Browsing in TrEd

Both VALLEX and PDT-VALLEX can be displayed in TrEd as it is shown in Figure 2 (the style depends on the provided PML-schema and the stylesheet).

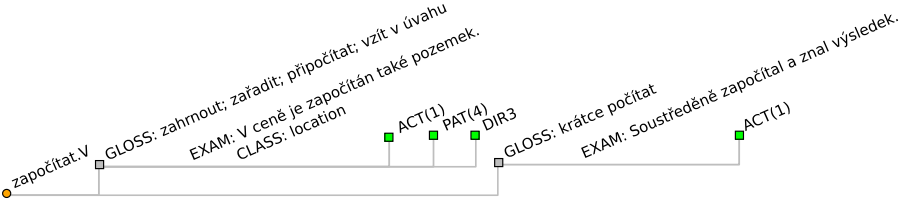

Fig. 2. The verb *započítat* 'to count' displayed in TrEd. The left node represents the lemma of the verb, its two children are two frame entries with different meanings. Both are provided with a gloss and an example. The upper level of nodes represents individual valency complementations and their possible possible morphological forms.

Furthermore, in VALLEX, each member of aspectual group displays the reference(s) to its counterpart(s). By clicking on the reference, the corresponding frame entry (or the whole corresponding verb entry) is displayed in a new window. VALLEX and PDT-VALLEX are interlinked by similar reference on the level of verb entries.

Let us anticipate that the links between frame entries (which correspond to individual meaning of verbs) across the lexicons can be viewed in the same way, see Section 4.

3.3 Searching the Lexicons Using PML-TQ

The lexicons can be not only viewed but also searched using PML Tree Query language. TrEd with PML-TQ extension allows users to formulate complex queries in a user-friendly way. All queries can be created in a graphical mode, a query having a form of a subtree with possible constraints on nodes and edges. Graphical interface enables users to insert nodes into a query subtree, to interconnect the nodes and to formulate constraints on their attributes. (Alternatively, a textual form of the query can be used.)

Let us exemplify some types of possible queries in PML-TQ. There are simple queries, as e.g., 'search for verbs with obligatory ADDRESSEE'. Queries with quantification can be asked too, e.g., 'find a verb with more than twenty valency frames'. Moreover, PML-TQ makes it possible to formulate complex queries concerning diverse properties of verbs, as e.g. the query in Figure 3. We can also search the previous queries in both lexicons in a parallel way.

The output from PML-TQ can be either just viewed in TrEd lemma by lemma, or it can be further processed – one can, for instance, ask for statistics (as, e.g., 'display

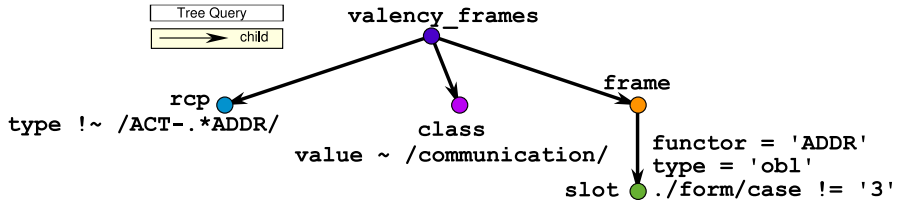

Fig. 3. Example PML-TQ query: searching for verbs from the class 'communication' with obligatory ADDRESSEE realized in other than dative case that cannot be in a reciprocity relation with ACTOR

frequency of found lemmas in individual verb classes'). This is achieved by 'filters' that can be appended to every query and that generate simple text tables.

3.4 Other XML-Based Lexicons

Czech valency lexicons serve here only as sample lexicons allowing us to demonstrate pros and cons of proposed common representation of lexicons. In fact, any XML-based[7] lexicon can be – after some necessary modifications – viewed in TrEd; however, hierarchical, highly structured lexicons benefit from this format most. First necessary step consists in creating a PML-schema; this includes automated transformation from DTD and manual assignment of a few PML roles (they create the required tree structure). Secondly, stylesheet for viewing the lexicon in the required way must be specified. It includes layout of displayed lexicon elements, their descriptions, colors etc. Thirdly, optionally, lexicon data can be transferred into the database for faster querying.

It is profitable not only for our example lexicons VALLEX and PDT-VALLEX but also for e.g. VerbaLex, the other Czech valency lexicon. Its format is different but captures similarly structured information and would be easily transformed to PML.

4 Conclusion and Future Work

In this paper we have presented a format for linking valency lexicons and an effective way how to visualize them in the tree editor TrEd. We have exploited a powerful PML-TQ search engine offering a graphical query representation for comfortable work of linguists. These tools can be used for any lexicon after transformation to PML format, which is mostly automated. It is especially profitable for lexicons with a hierarchical structure such as our example lexicon VALLEX (with several levels of lemma clustering) or as PDT-VALLEX (with structured frame slot information).

The `vallex_pml` format introduced here proved to be suitable common representation for these lexicons. This new data format, which overcomes different logical structures of VALLEX and PDT-VALLEX, poses an important prerequisite for interlinking both valency lexicons – more precisely, for (semi)automatic interlinking corresponding lexical units from VALLEX and PDT-VALLEX – and thus making available information stored in both lexicons (including references to external language resources). This represents an effective way of enriching particular lexical resources.

The `vallex_pml` format being supported by the tree editor TrEd offers parallel visualization of VALLEX and PDT-VALLEX and thus facilitates manual checking and necessary follow-up corrections of the automatic phase of interlinking the affected lexicons as well as viewing and searching in the interlinked system in the future.

References

1. Hajič, J.: Complex Corpus Annotation: The Prague Dependency Treebank. Veda, Bratislava, Bratislava, Slovakia, pp. 54–73 (2006)
2. Hajič, J., et al.: Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia (2006)

[7] Naturally, other format can be used, too, yet the transformation into PML is not automated.

3. Pajas, P., Štěpánek, J.: System for Querying Syntactically Annotated Corpora. In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, pp. 33–36. Association for Computational Linguistics (2009)
4. Pajas, P., Štěpánek, J.: Recent Advances in a Feature-Rich Framework for Treebank Annotation. In: Scott, D., Uszkoreit, H. (eds.) Proceedings of The 22nd International Conference on Computational Linguistics, The Coling 2008 Organizing Committee, Manchester, UK, vol. 2, pp. 673–680 (2008)
5. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice (2006)
`http://framenet.icsi.berkeley.edu/book/book.html`
6. Kipper, K., Korhonena, A., Ryant, N., Palmer, M.: Extending VerbNet with Novel Verb Classes. In: Fifth International Conference on Language Resources and Evaluation, LREC 2006 (2006)
7. Hanks, P.: Mapping meaning onto use: a Pattern Dictionary of English Verbs. In: AACL 2008 (2008)
8. Hlaváčková, D., Horák, A.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Computer Treatment of Slavic and East European Languages, Bratislava, Slovakia, Slovenský národní korpus, pp. 107–115 (2006)
9. Panevová, J.: Valency Frames and the Meaning of the Sentence. In: Luelsdorff, P.A. (ed.) The Prague School of Structural and Functional Linguistics, pp. 223–243. John Benjamins Publishing Company, Amsterdam (1994)
10. Hajič, J., et al.: PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In: Nivre, J., Hinrichs, E. (eds.) Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 57–68. Vaxjo University Press, Vaxjo (2003)
11. Urešová, Z.: The Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View. In: Šimková, M. (ed.) Insight into Slovak and Czech Corpus Linguistics, Veda, Bratislava, pp. 93–112 (2006)
12. Žabokrtský, Z., Lopatková, M.: Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. Prague Bulletin of Mathematical Linguistics, 41–60 (2007)
13. Žabokrtský, Z.: Valency Lexicon of Czech Verbs. Ph.D. thesis, ÚFAL MFF UK, Prague, Czech Republic (2005)
14. Mikulová, M., et al.: Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep. (2006)
15. Pajas, P., Štěpánek, J.: A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to Prague Dependency Treebank 2.0. Technical Report TR-2005-29, ÚFAL MFF UK, Prague, Czech Rep. (2005)