Joint search in a bilingual valency lexicon and an annotated corpus

Eva Fučíková **Jan Hajič** **Zdeňka Urešová**

Faculty of Mathematics and Physics
Charles University in Prague, Czech Republic
Institute of Formal and Applied Linguistics
{fucikova,hajic,uresova}@ufal.mff.cuni.cz

Abstract

... so I say to you ... search, and you will find ...

In this paper and the associated system demo, we present an advanced search system that allows to perform a joint search over a (bilingual) valency lexicon and a correspondingly annotated linked parallel corpus. This search tool has been developed on the basis of the Prague Czech-English Dependency Treebank, but its ideas are applicable in principle to any bilingual parallel corpus that is annotated for dependencies and valency (i.e., predicate-argument structure), and where verbs are linked to appropriate entries in an associated valency lexicon. Our online search tool consolidates more search interfaces into one, providing expanded structured search capability and a more efficient advanced way to search, allowing users to search for verb pairs, verbal argument pairs, their surface realization as recorded in the lexicon, or for their surface form actually appearing in the linked parallel corpus. The search system is currently under development, and is replacing our current search tool available at http://lindat.mff.cuni.cz/services/CzEngVallex, which could search the lexicon but the queries cannot take advantage of the underlying corpus nor use the additional surface form information from the lexicon(s). The system is available as open source.

1 Introduction

For linguistic research and for manual inspection of corpora, treebanks and lexicons, many different search tools exist (PML Tree Query[1] (Štěpánek and Pajas, 2010; Bejček et al., 2010), KonText (Klyueva and Straňák, 2016),[2] NoSketch Engine[3] (Rychlý, 2007), Tgrep,[4] BNC-search,[5] LAPPS Grid,[6], and many others. Every electronic lexicon (monolingual or bilingual) also comes with a basic search, typically allowing to search for headwords, or within any text using some form of fulltext search. Specifically both valency lexicons developed at the Institute of Formal and Applied Linguistics, PDT-Vallex[7] (Urešová, 2011b; Urešová, 2011a) and VALLEX[8] (Lopatková et al., in print; Žabokrtský and Lopatková, 2007), come with a search-allowing interface.[9]

However, we are not aware of any system that would allow structured search (a) in both a lexicon with rich information and an annotated corpus at the same time, *and* (b) bilingually. This could be caused also by the lack of parallel (bilingual or multilingual) corpora that are annotated by such rich lexicon entries

[1] https://ufal.mff.cuni.cz/pmltq
[2] https://ucnk.ff.cuni.cz/intercorp/?req=page:manual_kontext_en,
http://ufal.mff.cuni.cz/lindat-kontext
[3] https://nlp.fi.muni.cz/trac/noske
[4] https://tedlab.mit.edu/~dr/Tgrep2
[5] http://www.natcorp.ox.ac.uk
[6] http://galaxy.lappsgrid.org
[7] https://ufal.mff.cuni.cz/pdt-vallex-valency-lexicon-linked-czech-corpora
[8] http://ufal.mff.cuni.cz/vallex/3.0
[9] http://lindat.cz

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 40–44, Osaka, Japan, December 11-17 2016.

(usually, the corpora contain - in some cases - lemmas, which can be then searched for in *independent* lexicons). Our task was to make the Prague Czech-English Dependency Treebank (PCEDT 2.0) (Hajič et al., 2012) efficiently searchable, aiming to be a help for various applications of computational and traditional linguistics as well as for NLP studies. The PCEDT is a bilingual corpus that contains both rich dependency and predicate-argument annotation itself, as well as links to valency lexicons used (not only) for predicate-argument annotation consistency. In this respect, it is similar to the PropBank (Palmer et al., 2005), which annotates predicate-argument structure on top of the Penn Treebank (Marcus et al., 1993), indexing also by pointing to the frame files, from which additional information about the predicates (verbs) can be extracted. However, PropBank is a monolingual resource. Also, because English is not a very morphologically rich language, PropBank's frame files do not contain much more than a list of arguments and sense distinctions; in contrast, Czech language is quite rich in this respect, and consequently, the Czech valency lexicon entries (Urešová, 2011b) contain additional information on the required form of verb arguments in terms of case, prepositions to be used, etc. Fig. 1 shows a simple example of a valency entry for the Czech verb *kalkulovat*, which has two senses: a *compute sth* sense, and a more abstract sense on *counting on something, on somebody* (*to happen*). In the first sense, it adds an optional third argument *to compute ... from something*, and both senses also differ in the possible surface form - the second, more abstract sense requires a particular preposition and case (specified by `s+7`, lit. `with`, and `7` for instrumental case), while the deep object of the former sense is a simple direct (prepositionless) accusative (specified by `4`, as cases are typically numbered in Czech).

Figure 1: Czech Valency lexicon (PDT-Vallex) entry (two senses of "kalkukovat": lit. "compute" and "count on *sth/sb*")

In addition, not only the Czech and English treebanks are aligned in the PCEDT, but so are the associated valency lexicons for Czech - PDT-Vallex[10] (Urešová, 2011b) and English - EngVallex[11] (Cinková, 2006), forming a bilingual parallel CzEngVallex lexicon (Urešová et al., 2016) which explicitly aligns verb senses as well as verb arguments between the two languages.

What was missing after having explicit alignments annotated was a tool that would allow inspection of the resulting corpus and lexicons, allowing cross-lingual queries with reasonable flexibility to support linguistic studies, NLP tasks, manual check of results of automatic tools, etc.

In our previous work (Fučíková et al., 2015), we have developed a tool that can search the lexicon(s) in a cross-lingual manner, allowing to formulate queries such as *show me all pairs of verbs and their translations where the English verb is a phrasal verb, while the Czech one is not.* Fig. 2 shows the old interface (taken from (Fučíková et al., 2015)), and the result of such a query.

The tool did not allow the user to formulate the query over the associated parallel corpus, but it was at least showing the associated examples with the bilingual lexicon entries found. While useful as such, there was a demand both from linguists and from computer researchers to allow for more detailed queries: specifically, to be able to use the surface form constraints in the lexicon, and also constraints on the actual use in the associated parallel corpus.[12] For example, there was no way to specify exactly which particular phrasal verb the user wants to find (cf. Fig. 2). The new tool presented here answers to such demands

[10] http://lindat.mff.cuni.cz/services/PDT-Vallex

[11] http://lindat.mff.cuni.cz/services/EngVallex

[12] This has been in part due to the fact that the English valency lexicon unlike the Czech lexicon does not often contain any information about the required form, such as required or typical prepositions, and thus the corpus is the only place where such information is available.

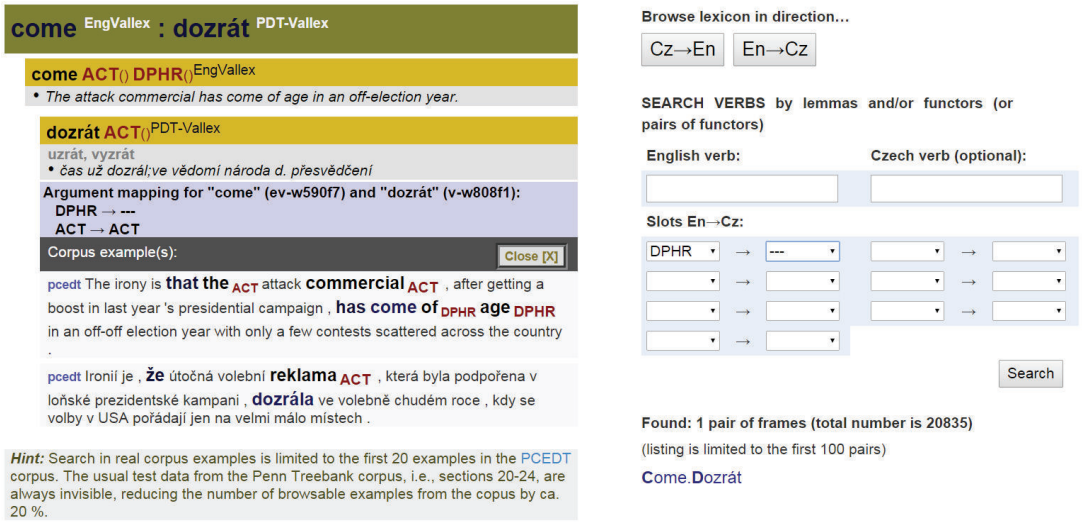

Figure 2: The old search interface

by implementing the possibility to search for the surface form constraints in the lexicon as well as in the corpus, in a bilingual setting.

2 System overview

Fig. 3 depicts the new search interface. We demonstrate the new system capabilities on the displayed example.

The query interface is on the right-hand side, the results appear left to it. In this case, the user searched for all Czech verbs—paired with the English verb *to cater (to)*—which express the corresponding argument (a deep object) by a prepositional phrase using the Czech preposition `o` (lit. `about`) with accusative, or by a subordinate clause introduced by the conjunction `aby` (lit. `so that`). There was only one result (*starat se*, lit. *to take care*) in Czech. One example from the PCEDT is also shown (which fully corresponds to the requirement that the English verb occurrence in the corpus has to be complemented by a prepositional phrase with `to`).

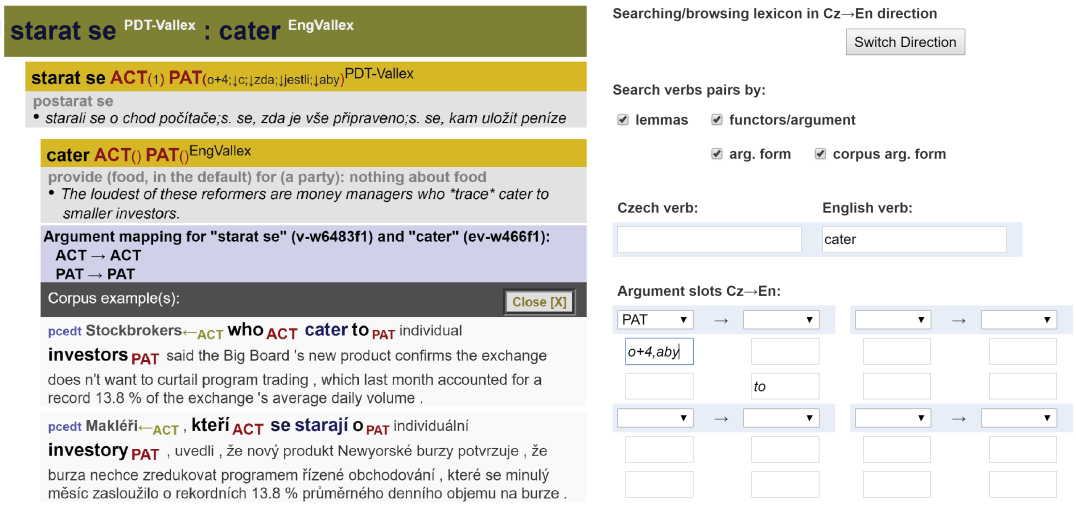

Figure 3: The new search interface and a result of search query

Let's have a closer look at the query: the checkboxes (lemmas, functors, arg. form, corpus arg. form) are all checked, bringing up the corresponding search fields to be filled. There is always a pair of these fields - Czech on the left, English on the right.[13] Apart from lemma, there are three search fields for each language: selector for the argument label ("functor"), lexicon argument form specification, and field for specifying the required corpus argument form. In Fig. 3, the user selected `PAT` as the argument label on the Czech side, limiting the search results to "deep objects" (label: `PAT`) of the Czech verb while leaving the field on the English side empty (any label - actor, deep object, addressee, ... could be paired with the Czech `PAT` on the English side). In addition, the (surface) argument form of the Czech `PAT` has been restricted to an accusative prepositional phrase headed by `o` (lit. `about`) which is expressed by `o+4`. As an alternative for this arg. form, the user allowed for this argument to be a subordinate clause headed by the conjunction `aby`, (lit. `to`), expressed as `aby`.[14] On the English side, the user has put the preposition `to` on the second line of the two form fields, since it is to be found in the *corpus*, not in the English valency lexicon.[15] These expressions are a shorthand for fully expressing the exact dependencies of the verb and its arguments; the relatively complex expansion of these fields and execution of a proper match of the treebank data is performed by the search engine.

In general, it is possible to use also much more complicated queries, using the usual logical operators ("and", "or") grouping and precedence, all combined with the possibility of using regular expressions on the literals (strings, whether lemmas, forms, or tags).

At query time, the search engine does not use the treebank (PCEDT) and the bilingual lexicon (CzEng-Vallex) directly, but they are pre-processed and indexed to make the search efficient. It also re-formats the treebank annotation to a linear annotation within the text (as can be seen in the examples), to make it more readable and avoid the need for additional visualization for the trees.[16]

Information about every pair of verbs is collected into a separate `.php` file (which also includes CSS-based formatting). In addition, information about the form for each argument of each valency frame for both PDT-Valex and EngVallex is extracted to another file for efficient search; similar index is created for the parallel corpora. There is also one more set of `.php` files for the display of dictionary examples, one file for each valency frame pair.

3 Conclusions and Future Development

We have described a search system over a bilingual lexicon and a parallel corpus. The tool builds on our previous simple search system, but substantially extends it for the use of surface form both as recorded in the lexicon as well as allowing to restrict the search to particular forms of argument expression in the associated corpus.

In the future, we intend to add more search possibilities, such as the option to search for particular form combinations in verb argument description, statistics (occurrence counts etc.) and their visualization.

The system is open, but at the moment its adaptation to other similar treebanks will require certain amount of work, namely to converts and index such treebanks and lexicons to the form which the search system uses at search time. This factorization allows, on the other hand to accommodate diverse corpora to be used, without regard to original formats or exact annotation schemas.

The system is available from the LINDAT/CLARIN language resource repository[17] as open source, as is the current system and the associated lexicons and corpora. The search interface can be used openly through any browser.

Acknowledgments

This work has been directly supported by the grant No. DG16P02019 of the Ministry of Culture of the Czech Republic.

[13] The direction can be switched for convenience.

[14] Czech prepositions do not overlap with conjunctions, so their lexical forms can be used without ambiguity in the queries.

[15] Which typically do not have the required preposition marked.

[16] For those interested in the annotation itself, there are other tools, such as the PML-TQ search system, see above.

[17] `http://lindat.cz`

In addition, it has also been using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (projects LM2010013 and LM2015071), which also hosts the resulting software.

References

Eduard Bejček, Václava Kettnerová, and Markéta Lopatková. 2010. Advanced searching in the valency lexicons using PML-TQ search engine. In *Text, Speech and Dialogue. 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings*, pages 51–58.

Silvie Cinková. 2006. From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2170–2175, Genova, Italy. ELRA.

Eva Fučíková, Jan Hajič, Jana Šindlerová, and Zdeňka Urešová. 2015. Czech-English Bilingual Valency Lexicon Online. In *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, pages 61–71, Warszawa, Poland. IPIPAN.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th LREC 2012)*, pages 3153–3160, Istanbul, Turkey. ELRA.

Natalia Klyueva and Pavel Straňák. 2016. Improving corpus search via parsing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of 10th LREC 2016*, pages 2862–2866, Paris, France. ELRA.

Markéta Lopatková, Václava Kettnerová, Eduard Bejček, Anna Vernerová, and Zdeněk Žabokrtský. in print. *Valenční slovník českých sloves VALLEX*. Nakladatelství Karolinum, Praha.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Pavel Rychlý. 2007. Manatee/bonito - a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masarykova univerzita.

Jan Štěpánek and Petr Pajas. 2010. Querying diverse treebanks in a uniform way. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1828–1835, Valletta, Malta. European Language Resources Association.

Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2016. CzEngVallex: a bilingual Czech-English valency lexicon. *The Prague Bulletin of Mathematical Linguistics*, 105:17–50.

Zdeňka Urešová. 2011a. *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.

Zdeňka Urešová. 2011b. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia.

Zdeněk Žabokrtský and Markéta Lopatková. 2007. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87):41–60.