Automatic Valency Derivation for Related Languages

Natalia Klyueva, Vladislav Kuboň
Charles University in Prague
Faculty of Mathematics and Physics
Czech Republic

Abstract

This paper describes an experiment combining several existing data resources (parallel corpora, valency lexicon, morphological taggers, bilingual dictionary etc.) and exploiting them in a task of building a valency lexicon for a related language (Russian) derived from a high quality manually created valency lexicon for Czech (Vallex) containing several thousands of verbs with very rich syntactic and semantic information. The experiment is restricted only to nominal constituents in both simple and prepositional cases. The results discussed in the second half of the paper seem to justify the method used and to encourage further experiments in this direction. The paper also discusses most frequent sources of errors.

Introduction

The notion of valency plays a crucial role in all attempts to syntactically analyze natural language texts using traditional rule-based methods. The lack of reliable large scale valency lexicon makes it practically impossible to create a grammar adequately describing syntactic rules of a particular natural language. Although the current mainstream research direction leads to an extensive use of data-driven methods, valency lexicons still constitute a valuable and important resource in some areas, as, e.g., in the area of machine translation between related languages or in the field of second language acquisition.

Building a large scale high quality valency lexicon is a costly and time-consuming effort which requires years of thorough linguistic work. The automatization of this process is challenging, especially for some types of natural languages, as, e.g., the languages with high degree of word-order freedom. In the sentences of free word order languages it is impossible to rely on the order of individual constituents and thus their identification constitutes a complicated problem.

Another problem in automated valency frame extraction is the difficulty of classifying obligatory and optional constituents. The frequency in a corpus is not a sufficient factor because many obligatory constituents are omitted at the surface level because they can be identified in the previous context. This makes the classification practically impossible. Unfortunatelly, not even the treebanks which capture all syntactic relationships at the surface and deep syntactic level can help. Adding the deep syntactic representation of the constituents which are omitted at the surface level is a challenging task by itself, which actually requires a thorough linguistic investigation of a particular word, or, if such a resource exists, a valency dictionary containing this information. So, it is actually a kind of a logical loop - we need valency dictionary for an annotation at the deep syntactic level, which then might provide the information necessary for creating the valency dictionary.

In order to overcome these problems we have decided to test a crosslingual method of determining valency. The idea is very simple - if we have a manually created and thoroughly tested valency dictionary for one language, why not to try to exploit its information across the language border? It is of course obvious that both languages have to belong to one language group, or, at least have some common history. Due to the large number of language resources available for two Slavic languages, Czech and Russian, we have decided to test the method on this pair of languages.

The Background of the Experiment

Although there is a large number or linguistic resources available for both Czech and Russian, the critical issue of these resources seems to be their incompatibility - both Czech and Russian have a rich and long linguistic tradition and the existing resources mostly had been created according to the theories developed by linguists of each particular language. The Czech linguistic resources mostly follow the theory of the Functional Generative Description of Sgall and Hajičová (Sgall et al., 1986)(Sgall, Hajičová, and Panevová 1986), while the Russian linguists stick to the Meaning Text Theory of Mel'chuk (Mel'chuk 1981). This dichotomy actually means that the linguistic resources with rich information and complex structure are practically impossible to combine and to use them in applications or experiments which require compatibility of resources, as, e.g., in machine translation.

Valency constitutes a major source of translation errors for rule-based MT systems even for closely related languages (Kuboň and Homola 2012), having the valency of both languages available (and using the same format and style) should boost the translation quality of MT between the two related languages.

In our experiment we are trying to create a valency lex-

icon for Russian on the basis of the existing valency lexicon for Czech Vallex (Lopatková, Žabokrtský, and Benešová 2006). There are other valency lexicons for Czech, some of them not available in the electronic form (as, e.g., the first valency lexicon of Czech verbs described in (Svozilová, Prouzová, and Jirsová 1997), which contains valency for only about 800 Czech verbs), some of them created more or less automatically (Czech Syntactic Lexicon, described in (Skoumalová 2001)) and thus containing less reliable information, some of them being restricted only to a description of surface level valency (as, e.g., BRIEF, described in (Pala and Ševeček 1997)).

Basic Notions - Valency

Before we describe our experiment, let us first specify what we understand under the notion of valency. This notion has been understood differently by various researchers. Generally, for a particular word - mostly verb - it presents the number of dependent words in a sentence - complements and their morphosyntactic characteristics. Although the valency lexicons provide primarily syntactic and lexically semantic information, it is morphological information which is crucial for processing texts especially in Slavic languages, as verbs determine the case of the depending noun. Let us take a sample phrase:

en: *He thanked his mother*
cz: *Poděkoval své mamince.DAT*
ru: *Poblagodaril svoju mamu.ACC*

The Czech verb *poděkovat* governs the noun in a Dative case, whereas in Russian the corresponding verb *poblagodarit'* governs a noun in Accusative. The situation where the cases of dependent nouns of some verb differ in Czech and Russian are not so frequent. According to (Klyueva and Kuboň 2010) the percentage of such cases is around 10%. Verbal valency slots may also contain infinite forms of (other) verbs, subordinated clauses, etc., but we are restricting our experiment only to nominal dependents, because other types of valency slots exhibit similar behaviour in both languages.

Within the linguistic tradition of the Prague school (cf. (Tesnière 1959)), the dependent words - complementations - are presented as a valency frame consisting of functors - inner participants and free modifications, that express the relation between a verb and a complement. Each functor is assigned by a morphemic realization of a complement. In our experiment we do not go as deep as to consider the functors and will stick only to the surface representation relying on the relatedness of the languages. We also ignore the dichotomy obligatory/optional complementation for the reasons we have already mentioned above - the obligatorness is usually not clearly manifested at the surface level, i.e. the level we are working with, because in the first phase of our experiments we intentionally avoid using any kind of parser, we are aiming at an experiment determining how much can be achieved by the combination of morphology, parallel data and a monolingual valency lexicon.

Vallex

The best source of data from our point of view is Vallex. It constitutes a very complex source of linguistic information, it contains verbal lexical entries with as many valency frames as they were found in the available corpora. Each valency frame contains a range of syntactic elements (verbal complementations) either obligatory (required at the deep syntactic level) or optional (grammatically permitted by this verb). In accordance with the theory of FGD, each verbal valency frame of a particular verb consists of valency slots for inner participants, both obligatory and optional, and for obligatory free modifications (adverbial modifications, adjuncts, etc.). An example of a Vallex record is presented in Fig.1. It shows two valency frames of the Czech verb *vyžadovat* (to demand).

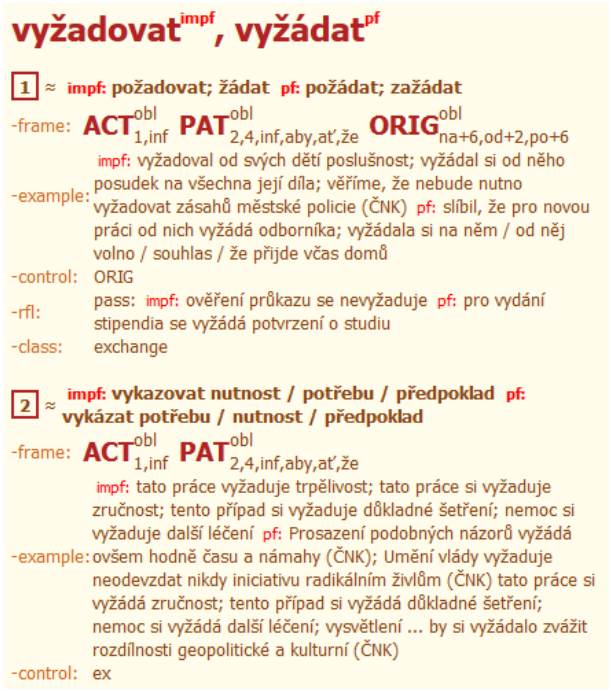

Figure 1: Vallex frames for the Czech verb "vyžadovat"

As we can see, the first valency frame contains three obligatory complements, namely Actor, Patient and Origin. The Actor may take a surface form either of a word in the nominative case, or of an infinite verb. The Patient may either be in genitive or accusative case, or it may be an infinite verb, or a subordinated clause starting with either of the three conjunctions *aby*, *ať* and *že*. The Origin manifests itself at the surface level by prepositional phrases with either a preposition *na* (on) and a noun in the locative case, or a preposition *od* (from) with a noun in genitive case, or a preposition *po* (after) with a noun in locative case. The second frame represents a different meaning of the verb (it can be translated as *to demonstrate the necessity*, which does not require the Origin to be present in a sentence.

For our experiment we are not exploiting valency frames in their full details. We ignore the classification of individ-

ual complements, we are only concerned with the verbal valency and the information about morphological characteristics (typically cases) of their nominal dependents. The core of our experiments is the identification of Czech nominal constituents in the Czech part of the parallel corpus and search for their Russian counterparts in the Russian part of the corpus. The restriction on nominal constituents constitutes the first phase of our research, which is supposed to give an answer whether this method is able to provide expected results or not. We are working with both types of nominal constituents, prepositional and non-prepositional ones.

The problem of extracting bilingual lexicons for Slavic languages is a well-studied problem. In (Bojar and Šindlerová 2010) authors collect valency translation equivalents for Czech and English verbs exploring the parallel treebank. (Rosa, Mareček, and Tamchyna 2013) built a simple probabilistic valency model and exploited this information to correct valency errors in the machine translation output. Last but not least, we have also done some preliminary study on the valency of Czech and Russian based on the small manually created dictionary (Klyueva and Kuboň 2010).

The Setup of the Experiment

We are aiming at using the simplest possible means in our experiment. In the first phase we avoid using syntactic parsers to identify the dependencies in sentences and types of identified nominal groups. Our experiment consist of the following stages:

Simplifying Vallex Frames

The original valency frame from Vallex contains complex linguistic information:

- lemma - the basic form of a verb;
- a deep semantic role called functor (Actor, Patient, Addressee etc.,);
- a surface realization of the functor;
- a semantic class of the verb;
- examples of using the verb in a real context;
- an information on reflexivity, aspect, idioms and some others.

In out experiment we are exploiting only the surface realization of verb complements (functors), typically having the form of a case or a combination of a case and preposition. For the moment we are leaving out the subject complements, assuming that the subject is mostly nominative in both Czech and Russian[1] and thus it can be included into the Russian valency frame automatically.

Let us present an example of the simplification of the valency frame of the Czech verb vyžadovat (to demand or to require) from Fig.1:
Original:
vyžadovat-V: ACT-1 PAT-2,4 ORIG-na+6,po+2,od+6
Simplified:
vyžadovat+2,4+(na+6,po+2,od+6) (codes 2,4 and 6 mark genitive, accusative and locative case in Czech, respectively).

Dictionary Lookup

For each lemma from Vallex we search for the Russian translation equivalent in the Czech-Russian commercial [2] dictionary, the translations can be multiple. The equivalents are then searched for in the parallel corpus in the next stage.

Parallel Corpus Lookup

The search is performed in the Czech-Russian part of two multilingual corpora, both containing 242 242 sentences for each language [3]. The texts are morphologically tagged, the tags contain a lemma, part-of-speech tag and other morphological characteristics. They are assigned to each word in each sentence in the format form|lemma|tag [4].

In the first step of our algorithm, the corpus is searched sentence by sentence, until we identify a verb whose valency frame is contained in Vallex. Vallex then provides its valency pattern - Czech lemma and the surface realization of the nominal dependents - this can be either an adjective, a noun or a pronoun within the same clause. The bilingual dictionary then provides corresponding lemma(s) which are looked up in the corresponding Russian sentence. In case of success (the verb corresponds to one of the lexical equivalents found in the translation dictionary), the respective case of a valency candidate (noun/adjective/pronoun) is extracted and stored in the hypothesis set. Following is a part of a sample tagged sentence from the corpus (An Arabic-Israeli peace requires a complex approach, because ...) and an illustration of how we process it. The Czech tagger outputs the following information:

Arabsko|arabský `A2--------A----`
-|- `Z:------------`
izraelský|izraelský `AAIS1----1A----`
mír|mír_ (opak_valky) `NNIS1-----A----`
vyžaduje|vyžadovat_:T `VB-S---3P-AA---`
komplexní|komplexní `AAIS4----1A----`
přístup|přístup `NNIS4-----A----`
,|, `Z:------------`
neboť|neboť `J^------------`

The bilingual dictionary then provides the translation of the Czech verb *vyžadovat* (demand, require) into the corresponding Russian lemma *trebovat'*.

This lemma is then identified in the tagged Russian sentence (the actual output has been transcribed into latin

[1] We have excluded 422 verbs the subject of which is in not in the nominative case. The set is to be processed separately.

[2] http://www.langsoft.cz/

[3] http://ufal.mff.cuni.cz/umc/cer/, http://www.korpus.cz/intercorp/

[4] The TreeTagger (http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/) has been used for Russian and the Feature-Based tagger for Czech (http://ufal.mff.cuni.cz/czech-tagging/)

alphabet):

Arabsko-izrael6skij|Arabsko-izrael6skij|arabsko-izrael6skij|Afpmsnf
mir|mir|mir|Ncmsnn
trebuyet|trebovat'|trebuyet|Vmip3s-a-e
vsestoronnego|vsestoronnij|vsestoronnego|Afpns-g-f
podchoda|podchod|podchoda|Ncmsgn
,|,|,|
potomu|potomu|potomu|P-----r
c'to|c'to|c'to|C

According to Vallex, the verb *vyžadovat* has two additional constituents apart from the Actor in nominative. One of these constituents is either in genitive or in accusative case, the other one is prepositional. Because there are no prepositions in our sample clause, we may narrow our search to nominal groups in genitive or accusative cases. Genitive case is not found either, so the only possibility how to fill the valency slot of this verb constitutes the noun *přístup* (approach).

With the Czech constituent identified, we get its Russian equivalent from the dictionary. Unfortunatelly, the translation of this Czech noun is highly ambiguous, it has the following Russian equivalents:
podchod, *podstup*, *pravo vchoda*, *pod'ezd*, *dopusk*, *pristup*, *obras'c'enie* and *dostup*.
The only candidate present in the tagged Russian sentence is the noun *podchod*. Its morphological tag Ncmsgn tells us that the corresponding case in Russian is genitive (the *g* tag on the 5th position).

The algorithm applied on this clause therefore provides a frame hypothesis:

(cz)*vyžadovat+Acc* => (ru)*trebovat'+Gen*
(An accusative case in the Czech valency frame probably corresponds to a genitive case in Russian).

Verbs requiring prepositional valency pattern are processed in a similar manner, it is only necessary to identify both the preposition and the case in the Czech text and to take into account that a prepositional case in Czech may correspond to a non-prepositional one in Russian and vice versa.

Russian Valency Frame Identification

The final phase consists in collecting all hypotheses established in the preceding phases for a particular Russian verb and choosing the most frequently occurring Russian valency frame from this set. This is a rather simplistic solution because the verb may have several valency frames (if it has more than one meaning), but a more subtle solution remains a task for the future research. In this initial experiment we are aiming mainly at testing the viability of the proposed method, more subtle solutions will be developed in subsequent phases of our research.

Results Achieved

The first (and simplest) type of evaluation we have performed after the experiment concerned the total number of patterns identified in our corpora. The exact results are presented in Table 1. The fact that we have been able to identify almost one third of verbs and their constituents on the basis of only slightly more than 240 000 sentences seems to be promising. Many patterns were simply not present in the data, Vallex had been created on a much larger set of data and thus it contains also many variants of valency frame which are less frequent and thus rarely found in smaller corpora.

verbs in the lexicon	5293
patterns "verb+constituent(s)"	14046
extracted patterns for Cz and Ru	4286

Table 1: Statistics of the experiment

Table 1 shows only the identified patterns, it does not reflect whether these patterns are correct or not. The errors we have discovered are discussed further in this chapter.

The second interesting result of our experiment is the confirmation of the hypothesis that Czech and Russian valency frames differ only in about 10% cases. This hypothesis has been formulated in (Klyueva and Kuboň 2010). In order to give a more precise answer we have splitted the set of frames into two parts - those with simple case and those with prepositional ones.

The results for the simple case correspondences are presented in the Table 2. According to this table, out of the total of 1727 simple case constituents only 343 are different. This represents almost 20%, (more exactly 19.86%) of the total. This number is 10% higher than the result presented in (Klyueva and Kuboň 2010).

		Czech			
		Gen	Dat	Acc	Ins
Russian	Gen	21	20	196	15
	Dat	1	159	12	2
	Acc	8	23	1026	22
	Ins	4	6	34	178

Table 2: Co-occurence of the same simple case in Czech and Russian

The results for prepositional valency are presented in the table 3. Due to a large number of very rare correspondences it was necessary to include only those detected at least 10 times into the table.

Out of the total of 841 pairs included into 3 there were 208 different pairs. That represents 24.7% of the total, slightly more than in the case of non-prepositional cases. Even this number contradicts the hypothesis set in (Klyueva and Kuboň 2010). This difference may be attributed to a very different set of data used in the initial experiment. Unlike in this case, the valency frames tested in (Klyueva and Kuboň 2010) had been manually created by linguists for the MT

Czech	Russian	freq
na+Acc	na+Acc	159
k+Dat	k+Dat	82
s+Ins	s+Ins	78
z+Gen	iz+Gen	58
v+Loc	v+Loc	56
za+Acc	za+Acc	52
do+Gen	v+Acc	50
od+Gen	ot+Gen	42
o+Loc	o+Loc	35
na+Loc	na+Loc	33
na+Acc	v+Acc	32
na+Acc	na+Loc	22
z+Gen	s+Gen	20
v+Acc	v+Acc	18
na+Acc	k+Dat	18
k+Dat	na+Acc	16
na+Acc	v+Loc	14
před+Ins	ot+Gen	12
proti+Dat	protiv+Gen	12
za+Acc	na+Acc	11
na+Acc	o+Loc	11
k+Dat	v+Acc	10

Table 3: Prepositional case correspondence.

system RUSLAN, cf. (Oliva 1989), no automatic method had been involved. The lexical items in the RUSLAN dictionary had also been very strictly domain dependent, aiming at the translation of manuals to operating systems of mainframe computers, i.e. a very technical domain with a specific language.

Error Analysis

One of the possible reasons for a difference in the estimation of the number of differing valency slots mentioned above may also be the frequency of errors in our automatic experiment. In order to detect the errors, to discover the source of the errors and to improve the algorithm on the basis on this information, we have perfomed a manual evaluation of a small sample of valency frames. Out of the set of 4286 extracted frames we have manually evaluated 200 frames. Among those, 24 frames, i.e. 12% of the sample, were recognized as incorrect. The analysis of errors presented in 4 shows that there were several reasons for those errors. Some errors were caused by tagging inaccuracy, some others resulted from an erroneous match of Czech and Russian nouns), and the rest can be attributed to other factors, as, e.g., bilingual dictionary issues.

tagging issues	7
experiment setup	7
others	10
Total	24 (12%)
Evaluated verbs	200

Table 4: Error types according to the manual evaluation

After having analysed the major categories errors, we have tried to predict which pairs of frames in Czech and Russian are most likely to cause an error:

- The most frequent error has a pattern

  Czech: Verb+Acc => Russian: Verb+Gen.

  This error pattern has its roots in the tagger inaccuracy. In Russian, an animated masculine noun has the same form in genitive and accusative cases, and the tagger often confuses them. So even if the algorithm matches all the dependencies correctly, the extracted case of the Russian noun is incorrect. Let us present an example:

  ERR: (cz) *najímat*+Acc (ru) *nanimat'*+Gen (to rent smth., should be also in the accusative case in Russian)

- The second group of 'suspicious' cases contains a prepositional valency frame in one language and a simple one in the other. Let us illustrate this on the following entry:

  (cz) *odebrat*+Acc(take smth.) => (ru) *otobrat'*+u+Gen(take from smb.)

  Clearly, due to the free word order, the semantic roles of the nouns has got transposed. Let us look at this example more closely and examine a sentence from our corpus containing this example ('They took a cake from Simeonov'): [5].

  (cz)Simeonovovi|3 odebrali dort|4
  'Simeonov.Dat took.3Pl cake.Acc'

  (ru)tort|4 otobrali u|prep Simeonova|g
  'cake.Acc take.3Pl from|prep Simeonov.Gen'

  Although our algorithm has identified both dependencies - object and indirect object, the latter has got mixed up because of the reversed word order in Russian. The same situation was observed in many sentences - when the algorithm has chosen the most frequent variant, it turned out that it was an incorrect one for that particular verb. It should be noted, that the correct valency frame for the indirect object was generated as well:

  (cz)*odebrat*+Dat(take from smb.) => (ru) *otobrat'*+u+Gen(take from smb.)

  This mistake is beyond the abilities of our simple algorithm, a possible solution of this problem is to use some deeper parsing strategy which would be able to identify the type of the noun phrases involved.

Conclusions and Future Research Directions

Although the experiment is relatively simple, the results achieved so far are quite encouraging, the percentage of cor-

[5] In order to simplify the text we leave only the relevant morphological tags

rectly identified frames suggests that the direction we are taking might lead to a relatively fast method of creating large scale valency lexicon for Russian containing verbs occurring most frequently in the parallel corpus.

The method introduced in this paper might be improved through the exploitation of a syntactic parser. This would enable to include also the complements which cannot be easily identified in a sentence, as, e.g., the long-distance dependencies. The problem will be solved as soon as we obtain a parser for Russian compatible in format with the Czech one, which is our primary plan for future. The incorporation of the parser will also help to extend the scope of our experiments and to include also other types of complements, not only the nominal ones.

Acknowledgments

The research was supported by the grants GACR P406/12/0557 and GAUK 639012.

References

Bojar, O., and Šindlerová, J. 2010. Building a bilingual vallex using treebank token alignment: First observations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, 304–309. Valletta, Malta: ELRA.

Klyueva, N., and Kuboň, V. 2010. Verbal valency in the mt between related languages. In *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features*.

Kuboň, V., and Homola, P. 2012. *Machine Translation Among Related Slavic Languages*. Cambridge, United Kingdom: Cambridge Scholars Publishing. 283–307.

Lopatková, M.; Žabokrtský, Z.; and Benešová, V. 2006. Valency lexicon of czech verbs VALLEX 2.0. Technical Report 34.

Mel'chuk, I. 1981. *Meaning-text Models: a Recent Trend in Soviet Linguistics*. Annual Review of Anthropology 10.

Oliva, K. 1989. *A Parser for Czech Implemented in Systems Q*. Explizite Beschreibung der Sprache und automatische Textbearbeitung. Matematicko-fyzikální fakulta UK.

Pala, K., and Ševeček, P. 1997. Valence českých sloves. In *Sborník prací FFUB*.

Rosa, R.; Mareček, D.; and Tamchyna, A. 2013. Deepfix: Statistical post-editing of statistical machine translation using deep syntactic analysis. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, 172–179. Sofia, Bulgaria: Bălgarska akademija na naukite.

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

Skoumalová, H. 2001. *Czech Syntactic Lexicon*. Ph.D. Dissertation, ÚTKL FF UK.

Svozilová, N.; Prouzová, H.; and Jirsová, A. 1997. *Slovesa pro praxi*. Prague, Czech Republic: Academia.

Tesnière, L. 1959. *Eléments de syntaxe structurale*. Paris: Librairie Klincksieck.