Comparing Czech and Russian Valency on the Material of Vallex

Natalia Klyueva
Institute of Formal and Applied Linguistics
Charles University in Prague
kljueva@ufal.mff.cuni.cz

Abstract

In this study we have compared Czech and Russian valency frames based on monolingual and bilingual data. We assume that Czech and Russian are close enough to have, for the majority of their verbs, similar valency structures . We have exploited Vallex as a source of valency frames and have used a Czech-Russian dictionary to automatically translate Czech verbs into Russian. Afterwards, we have manually checked whether the Czech frame fits the Russian verb and, in case it was different, we have added the verb to the set that will be described in our paper. We suggest that there is a connection between the semantic class of some verbs and the type of difference between their Czech and Russian valency frames.

1 Introduction

Verbal valency is an important topic in Natural Language Processing which has been broadly studied within various linguistic branches - theoretical and practical. Bilingual research on valency is crucial for practical fields such as Machine Translation, or second language acquisition. There are many sources of information on both valency and word classes - WordNet(Fellbaum, 1998), FrameNet(Baker et al., 1998), VerbaLex(Hlaváčková and Horák, 2006) and Vallex(Lopatková et al., 2006) to name some of them. The central resource for our research has been the Czech Valency Lexicon Vallex. The Resource for Russian Valency, Explanatory Combinatorial Dictionary of Modern Russian(Mel'čuk and Zholkovsky, 1984), which is comparably big and rich in terms of language information is not available on-line, so we can not make a straightforward comparison. Instead, we have looked at the Russian verbal valency through the prism of the Czech one.

Czech and Russian are Slavic languages which are related and therefore share many morphological and syntactic features. A valency frame of a Czech verb is, in the majority of cases, similar to that of a Russian one. The focus of our study is on the verbs which have a different valency structure between the two languages. It seemed interesting for us not just to collect the set of those verbs, but rather to find out whether those Czech and Russian verbs that present some dissimilarity between their valency have some regularity or rule, or if this discrepancies are merely coincidental. Our hypothesis is that these differences have something to do with the semantic class of a verb.

There is a resource of valency bilingual data for Czech and Russian - the dictionary Ruslan (Oliva, 1989) that contains this information. But it is not big and it can only give us 'absolute' numbers - the percentage of verbs with different valency structure (Klyueva and Kuboň, 2010), without a insight into the nature of these dissimilarities. Vallex enables us to browse various verbs classes and see the underlying connections between the semantics of these verbs and the difference of frames in the two languages.

The idea of using data from Czech language in order to create new data for Russian was exploited by (Hana and Feldman, 2004), who constructed a

Proceedings of KONVENS 2012 (LexSem 2012 workshop), Vienna, September 21, 2012

morphological tagger for Russian language upon Czech data and tools. In (Benešová and Bojar, 2006) authors compare the similarity between the automatically extracted valency frames and the manually annotated frames.

2 Vallex and Verb Classes

Vallex - a manually created Lexicon of Czech Verbs - is based on the valency theory in the Functional Generative Description (Panevová, 1994),(Sgall et al., 1986). It provides an information on valency frames of the most frequent verbs (in version Vallex 2.5 over 2.700 lexeme multiplied by different senses of the verbs). The frame consists of a slot that reflects the number of complements the verb may govern. A slot includes a functor (a deep semantic role, written after a period attached to the word) and the surface realization of it(mostly morphosyntactic case, written in brackets). The main deep semantic roles that have frequently been used in our work are:
ACT:Actor, ex. I.ACT love peaches.
PAT:Patient, ex. Cats love rats.PAT
ADDR:Addressee(a person or an object to whom/to which the action is performed - more in the paper below), ex. He gave him.ADDR a book
DIFF:Difference measure, ex. Prices have fallen twice.DIFF

Verbs are classified into verb classes according to their meaning, which we have used in our research as well. Vallex distinguishes 22 verb classes, among them are communication, exchange, motion, perception, transport, psych verb, just to mention some. Naturally, words that belong to the same semantic field or share some component of meaning will have a similar valency frame. Vallex entry also provides other valuable information on aspect, reciprocity, reflexivity etc. that we have not used in our work, so it will not appear in our examples. Here is an example of a Czech verb frame that belongs to the Mental Action verb class:
apelovat Act(Nom) Pat(na+Acc)-(on+Acc). This means that the verb *to appeal* governs two arguments: an Actor in the Nominative case and a Patient in prepositional phrase on+Accusative. The case systems in Czech and Russian are very similar and prepositions have almost identical surface form which simplifies the process of comparison.

3 Czech Vallex for Russian verb frames

We made a comparison of Czech and Russian frames based on the Czech Lexicon in the following way. We took a Czech verb and said if its frame fits the frame of a Russian equivalent verb as well. At this stage, it was impossible to evaluate a big amount of verb frames (totally 2,903 lexical units have a verb class assigned), so we took the selection of the most frequent verbs distributed among the following verb classes: motion, communication, change, exchange and mental action, as the most representative ones.

This verb set contains frequent verbs of various semantic types. Our assumption is that the difference in valency frames might be related to a verb class, in other words, verbs from certain classes might have tendency to have different valency in Czech and Russian. In our study we focus on morphemic forms of **noun complements**, leaving aside verb complements and sentence complements of verbs. Within a semantic class for each Czech verb we state whether or not a Russian equivalent has the same valency structure.

For example, (1) shows the verb with the same valency frame and the verb in example (2) has two discrepancies in it..[1]

(1cz)*obhajovat* ACT(Nom) PAT(Acc),to defend
(1ru)*zaščiščat'* ACT(Nom) PAT(Acc),to defend
The frame is the same in both languages

(2cz)*blahopřál mu.ADDR(**Dat**) **k** narozeninám*
'congratulated him.ADDR(**Dat**) **to** birthday'
(2ru)*pozdravljal ego(**Acc**) **s** dnem roždenija*
'congratulated him.ADDR(**Acc**) **with** birthday(with+Ins)'

In the example (2) it is illustrated that in Czech and Russian different prepositions and different cases are used to express the same semantic roles - Patient and Addressee. Especially diverse in this case is the surface realization of Patient as a prepositional phrase across the languages: Czech

[1]There are 6 cases in Russian and 7 cases in Czech (7th, Vocative, is not relevant for our study) and case endings are very similar in both languages. Czech and Russian prepositions are almost identical as well. All this makes it rather easy to detect differences in valency frames.

- congratulate to, Russian - congratulate with, English - congratulate on/upon.

We consider the Russian frame to be similar to the Czech one if it has the same number complements, the same semantic roles and if these semantic roles have the same surface realization. All the verbs we observed met the first two conditions because we tried hard to find the closest translation equivalent in Russian. It was always the surface form that was different in two languages. If a surface form is represented by a preposition with some case, we judge the default translation of prepositions as the similar realization.

Further on, to simplify the examples, we will leave only the slot of the frame that is different in the languages and leave out the slots that are irrelevant to our comparison. So the example verb (2cz) will be shortened to *blahopřát* PAT(k+Dat) ADDR(Dat) and (2ru)*pozdravljat'* PAT(s+Ins) ADDR(Acc) ' leaving aside the functor ACT(Nom) which is almost always the same in Czech and Russian. The examples in this paper are either taken from corpus, invented or taken straightly from Vallex examples.

4 Differences According to the Verb Classes

While analyzing Czech and Russian frames, it became evident that the differences between Valency frames can be either regular or occasional. In this paper we will present the description of the differences according to the semantic classes of the verbs. Some groups of verbs that have some regular discrepancy in a valency frame may belong to different classes, as it will be illustrated below.

4.1 Class of Change

Verbs of the class Change often have the complement DIFF, and we observed that it often has different realization in Czech and Russian, namely the slot cz:DIFF(o+Acc)-(about+Acc) generally corresponds to ru:(na+Acc)-(on+Acc) in Russian (other variations are possible), see examples (3) and (4).
(3cz)*ceny klesly o 20%* 'prices fall **about** 20%'
(3ru)*ceny upali na 20%* 'prices fall **on** 20%'
(4cz) *Administrace zkrátila dovolenou **o** 2 dny*
'administration cut off the holiday **about** 2 days'
(4ru) *Administracija sokratila otpusk **na** 2 dnja*
'administration cut off the holiday **on** 2 days'

For the functor DIFF, we should mention, that the form (about+Acc) is typical of Czech while Russian language uses the preposition 'o' (about) mainly with mental predicates like English(forget about+Loc) or communication verbs (tell about+Loc) and does not occur with the Accusative case at all.

4.2 Class of Motion

We have not found many dissimilarities in Czech and Russian valency frames within the class of Motion verbs. One most evident is that verbs of classes motion with the semantic component of 'going away from somewhere' in the case they have the surface realization of PAT as (před+Ins)-(before+Ins) in Czech are translated into Russian with the respective verb plus the prepositional phrase (ot+Gen)-(from+Gen), not the expected ru:(pered +Ins): prchat, ujíždět, unikat.

(5cz)*prchat* **před** policii-'run before police'
(5ru)*ubegat'* **ot** policii 'run from police'

In other words, Russian prefers the preposition 'from' whereas Czech uses 'before' in this context. Verbs of other semantic classes with the similar component of meaning, ex. class location - share this rule as well(cz:schovat před+Ins - 'to hide before' vs. ru:sprjatat' ot+Gen - 'to hide from').

The following example illustrates a coincidental difference in verb frame :
(6cz)*trefit* PAT(**Acc**) 'to hit smth '
(6ru)*popast'* PAT(**v+Acc**) 'to hit into+Acc'

4.3 Verbs of Exchange

One of the regular and rather evident differences between Czech and Russian frames was described in (Lopatková and Panevová, 2006). This is the case of some exchange verbs with the meaning of removing something from someone, ex. sebrat(take away), krást(steal), brát(take) etc. The addresse here is a person or an object from

whom/which something is taken.

(7cz)brát ADDR(**Dat**)-'take +Dat'
bere dítěti hračku -'takes baby.Dat toy'
(7ru)brat' ADDR(**u+Gen**)-'take (u+Gen)'
beret u rebenka igrushku 'He takes of baby.Gen toy'
(7en)'He takes a toy from a baby'

(8cz)zabírat ADDR(**Dat**)-take(time)+Dat
studium mi zabírá hodně času
'study me.Dat takes many time'
(8ru) otnimat' ADDR(**u+Gen**)
ucheba otnimaet u menja mnogo vremeni
'study takes from me many time'
(8en)'Study takes me a lot of time. '

In this cases if the sentence (8cz) was translated into Russian according to the Czech valency pattern, they would have the reverse meaning in Russian, because the Dative case of the noun in this context is understood as Benefactor (taken TO someone), not Addressee (taken FROM someone). Especially this difference causes big problems to learners of foreign languages: they project the known pattern from their native language onto the phrase in the foreign language and, given that the surface form of the preposition is the same, they make a mistake.

This scheme does not work for all words with this meaning in this class, for example a semantically related word 'odpírat'(to deny) in Russian has the same surface form of Addressee in Russian(ADDR(Dat)) as in Czech, yet another non-direct realization of Patient:

(9cz)odpírat ADDR(Dat) PAT(**Acc**)
odpíral mu pomoc
'denied him help'
(9ru)otkazyvat' ADDR (Dat) PAT(**v+Loc**)-(in+Loc)
on otkazal emu v pomošči
'denied him in help'
(9en)'he denied to help him'

On the example of this verb class we can see that the semantically related group of words has different surface realizations of a functor (ADDR in this case) in Czech and Russian. This makes us believe that difference in valency frames can depend on the semantic class. Only two words of this class with different valency framedo not belong to the group described, and we consider them to be occasional discrepancies. The number of occasional discrepancies in the verb classes is not so big in comparison with ones that have some regular difference.

4.4 Class of Communication

Czech and Russian verbs belonging in this class have many differences with respect to valency. Here we could not observe some of the leading difference present in the previous classes. Differences may concern several functors and several surface forms. They may be considered coincidental, but we can allocate several groups of verbs with some dissimilarity in valency frames.

1. The functor Addressee with the surface form ADDR(na+Acc)-(on+Acc) in Czech is presented in another way in Russian

(10cz)*mluvit* (**na+Acc**)-'speak on smb(Acc)'
(10ru)*obraščat'sja* ADDR(**k+Dat**)-'speak to smb(Dat)'

(11cz)*zavolat* (**na+Acc**)-'call on smb(Acc)'
(11ru)*pozvat'* (**Acc**) -'to call smb(Acc)'

2. Patient with the surface form (na+Loc)-(on+Loc) in Czech corresponds to another realization in Russian, generally the morphemic form is (o+Loc)-(about+Loc) for such verbs used for 'asking question' as (ze)ptát se, tázat se etc.:

(12cz)*ptát se* PAT(**na+Acc**) *zdraví* -'ask on health'
(12ru)*sprosit'* PAT(**o+Loc**) *zdorov'je* -'ask about health'

Other verbs with a frame slot PAT(na+Loc)-(on+Loc) are also very similar to the above sample:

(13cz)*domlouvat se* PAT(**na+Loc**) - 'to agree on'
(13ru)*dogovorit'sja* PAT(**o+Loc**) 'to agree about'

3. Addressee in Dative case for the following verbs corresponds to Accusative in Russian:

(14cz)*poblahopřát* ADDR(**Dat**)'congratulate +Dat'
(14ru)*pozdravit'* ADDR(**Acc**) 'congratulate +Acc'

(15cz)*děkovat* ADDR(**Dat**) 'thank +Dat'
(15ru)*blagodarit'* ADDR(**Acc**) 'to thank + Acc'

4. Similar to the verbs of Exchange class, some Czech communication verbs with surface form (o+Acc)-(about+Acc) will be translated in another manner in Russian due to the fact that, unlike in Czech, the preposition 'o'-'about' does not combine with the Accusative:

(16cz)hlásit se PAT(**o+Acc**):
hlásí se o slovo 'ask about word'
(16ru)prosit' PAT(**Gen**)
Ona prosit slova
'She ask word.gen'
(16en)'ask for a word'

Coincidental differences occurring only once or twice are not going to any scheme:

(17cz)*doznávat se* PAT(**k+Dat**) 'confess to smth'
(17ru)*priznavat'sja* PAT(**v+Loc**) 'to confess in smth'
(18cz)*konzultovat* PAT(**Acc**) 'to consult +Acc'
(18ru)*konsultirovat'* PAT(**po+Loc**)-(about+Loc)

4.5 Class of Mental Action

Verbs of this class often have differences in valency frames, but they are rather coincidental and we have found only one regular difference - when Czech PAT(na+Acc)-(on+Acc) corresponds to Russian PAT (o+Loc)-(about+Loc), (pro+Acc)-(about+Acc) or (k+Dat)-(to+Dat). The surface form (na+Acc) in Czech is also different for verbs belonging in the class Communication, but for that class it was regularly translated as (o+Loc)-(about+Loc) whereas for the class of Mental Action no common translation equivalent exists.

(19cz)*pamatovat PAT*(**na+Acc**) 'to remember on'
(19ru)*pomnit'* PAT(**pro+Acc**) 'to remember about'
(20cz)*myslet* PAT(**na+Acc**) 'to think on'
(20ru)*dumat'* PAT(**o+Loc**) 'to think about'

(21cz)*zvykat si* PAT(**na+Acc**) 'get used on'
(21ru)*privykat'* PAT(**k+Dat**) 'to get used to'

The structure of the following verb coincides a lot with that from ex. (14) and (15) though the functor is PAT, not ADDR:
(22cz)*rozumět* PAT(**Dat**) 'understand'
(22ru)*ponimat'* PAT(**Acc**) 'understand'

Other coincidental differences:
(23)*pohrdat* PAT(**Ins**)
(23)*prezirat'* PAT(**Acc**) 'to despice'
(24cz)*mrzet* ACT(**Acc**)
(24ru)*sožalet* ACT(**Nom**) 'to be sorry for'
The example (24) is the one of a very few verbs with different surface realization of ACTor.

4.6 Overall results

We have compared Czech and Russian valency frames of verbs from 5 semantic classes, totally 1473 lexical entries. 111(7.5%) of them were different in Czech and Russian. The comparison was rather straightforward because of the relatedness of the languages. If some more distant languages were compared, more complicated method of evaluation should be chosen. From the examples above we can make the following observations:

- most dissimilarities occur in prepositional phrases.
- the regular discrepancies are more frequent than the coincidental ones.
- Within a verb class we can find some typical valency patterns of Czech verbs which correspond regularly to the different Russian pattern.

The table 1 presents the distribution of verbs with different frames according to the verb classes.

From this table we can see that verbs of physical activity(change, motion, exchange) have in

Verb class	same frame	different frame	# of verbs
Change	309(95%)	14(5%)	323
Exchange	166(92%)	13(8%)	179
Motion	305(99%)	3(1%)	308
Communication	312(88%)	42(12%)	354
Mental Action	270(87%)	39(13%)	309
Total	1362(92%)	111(8%)	1473

Table 1: Differences according to the verb classes

some sense less complicated valency structures than verbs of mental activity(communication, mental action) and that in most cases, their valency structure corresponds to that of Russian verbs.

5 Conclusion

In this paper we have described the dissimilarities in Czech and Russian Valency based on the material of the Czech lexicon. Our main hypothesis was that the differences in valency structure might be explained by the semantics of verbs , so we have exploited the classification of the semantic classes provided by Vallex. In almost in each verb class we have found some regular dissimilarity that is typical of this class. Still, there are some cases when verbs from other classes are subjected to this regularity as well, so other aspects (such as surface realization) should also be taken into con- sideration. A practical result of our paper is that we have made a draft version of a small bilingual Czech-Russian lexicon with different frames in the Vallex format.

Acknowledgments

The research is supported by the grants P406/2010/0875 GAČR and GAUK 639012.

References

Ch. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press

C. Baker, C. Fillmore and J. Lowe. 1998. The Berkeley FrameNet Project. *Proceedings of the 17th international conference on Computational linguistics - Volume 1 (COLING '98), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 86-90.*

V. Benešová and O. Bojar. 2006. Czech Verbs of Communication and the Extraction of their Frames. *Proceedings of the 9th International Conference, TSD 2006, pages 29-36.*

J. Hana and A. Feldman. 2004. Portable Language Technology: Russian via Czech. *Proceedings of the Midwest Computational Linguistics Colloquium*, June 25-26, 2004, Bloomington, Indiana.

D. Hlaváčková and A. Horák. 2006. VerbaLex - New Comprehensive Lexicon of Verb Valencies for Czech. *Computer Treatment of Slavic and East European Languages*. Bratislava, Slovakia: Slovenský národný korpus, p. 107-115.

N. Klyueva and V. Kuboň. 2010. Verbal Valency in the MT Between Related Languages. *Proceedings of Verb 2010, Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features* Universita di Pisa - Dipartimento di Linguistica, Pisa, Italy, pp. 160-164.

M. Lopatková and J. Panevová. 2006. Recent developments in the theory of valency in the light of the Prague Dependency Treebank. *In Mária Šimková, editor, Insight into Slovak and Czech Corpus Linguistic, pages 83-92. Veda Bratislava, Slovakia.*

M. Lopatková, Z. Žabokrtský and V. Benešová. 2006. *Valency Lexicon of Czech Verbs VALLEX 2.0.* Technical Report 34, UFAL MFF UK.

I. Mel'čuk and A. Zholkovsky. 1984. *Explanatory Combinatorial Dictionary of Modern Russian.* Semantico-syntactic Studies of Russian Vocabulary. Vienna: Wiener Slawistischer Almanach.

K. Oliva. 1989. *A parser for Czech implemented in systems Q.* Praha:MFF UK

J. Panevová. 1994. Valency Frames and the Meaning of the Sentence. *The Prague School of Structural and Functional Linguistics* (ed. Ph. L. Luelsdorff), Amsterdam-Philadelphia, John Benjamins, pp. 223-243.

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