Resources in Conflict: A Bilingual Valency Lexicon vs. a Bilingual Treebank vs. a Linguistic Theory

Jana Šindlerová, Zdeňka Urešová

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

Abstract

In this paper, we would like to exemplify how a syntactically annotated bilingual treebank can help us in exploring and revising a developed linguistic theory. On the material of the Prague Czech-English Dependency Treebank we observe sentences in which an Addressee argument in one language is linked translationally to a Patient argument in the other one, and make generalizations about the theoretical grounds of the argument non-correspondences and its relations to the valency theory beyond the annotation practice. Exploring verbs of three semantic classes (Judgement verbs, Teaching verbs and Attempt Suasion verbs) we claim that the Functional Generative Description argument labelling is highly dependent on the morphosyntactic realization of the individual participants, which then results in valency frame differences. Nevertheless, most of the differences can be overcome without substantial changes to the linguistic theory itself.

Keywords: treebank annotation, bilingual corpora, valency lexicon

1. Introduction

In the light of recent development in the area of applied linguistics it seems convenient to explore annotated multilingual corpora in order to gain a better balance between theoretical systems of language description and needs of NLP applications. Today there exists a considerable amount of annotated corpora covering different depth and width of linguistic description, a wide range of content domains, and a large amount of world languages. Many of these corpora are accompanied by additional resources, such as valency lexicons. Parallel valency lexicons, accompanying multilingual corpora, well satisfy the call for capturing complex lexical information, i.e. the information on both verbal translational equivalents and their valency frames. Having resources of this kind at our disposal gives us an opportunity to study similarities and differences between syntactic structures of different languages and to critically review theories developed on mother tongue language material and applied to a foreign language material.

In this paper, we focus on Czech and English deep syntactic (tectogrammatical) valency structures in a contrastive perspective. On the material of the Prague Czech-English Dependency Treebank (PCEDT)[1] (Hajič et al., 2012), we study sentences with selected differences in argument labelling between languages. There are two ways in which argument labelling differences manifest in a bilingual corpus. Either there appears different argument labelling of corresponding participants of a particular sentence and its translation (see Fig. 1), or there may be a difference in argument labelling between verbs within a particular language, such that the verbs belong to the same verb class and share the same translational equivalent. These two forms can appear separately or mingled in the data.

[1] http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4

Regarding the fact that the tectogrammatical layer is so far the deepest layer of syntactic description within the Functional Generative Description theory (FGD) (Hajičová and Sgall, 2003), declared to convey linguistic meaning and represent an input for semantic interpretation of the sentence (Sgall, 2006), it has been considered a suitable layer of representation for machine translation systems in the past. It is believed that at the tectogrammatical level, the differences between semantically equivalent analytical (surface) syntactic structures of distinct languages wipe off and the structures become close in appearance to each other. Then, the differences in tectogrammatical labelling between translationally equivalent structures may reveal interesting facts about the nature of individual languages.

In the light of our observations, the argument non-correspondences found in the PCEDT data may have different grounds:

- annotator's mistake or misunderstanding;
- a result of the translation of the verb, either an invalid or inconvenient translation equivalent, or a translation using a verb with slightly different semantic preferences and presuppositions;
- different grammatical (morphological and/or syntactic) properties of the two languages;
- different annotator intuitions considering argument interpretation due to the vagueness of the used linguistic theory.

Within the research we are particularly interested in the last two items of the list.

Our goal is to unveil the way a contrastive approach with the use of a bilingual treebank and bilingual valency lexicon can shed light on the theoretical questions of verbal valency and treebank annotation.

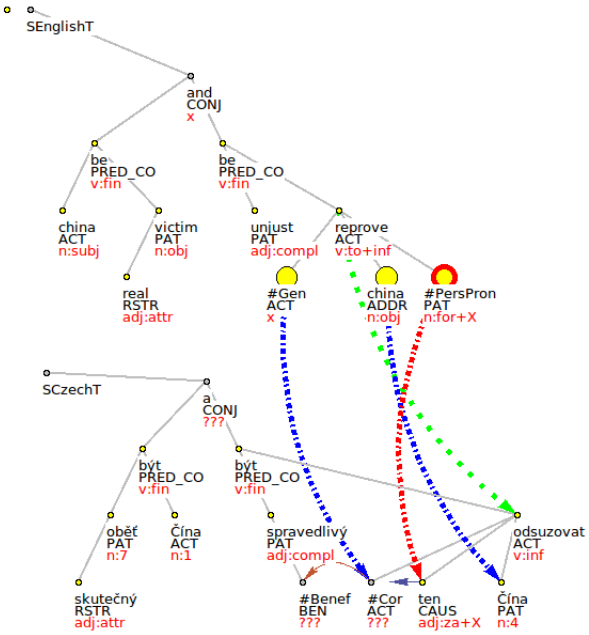

Figure 1: An example of different argument alignment between Czech and English.

2. Data

Our research is closely connected to a project aimed at interlinking two existing valency lexicons, the PDT-Vallex (Urešová, 2011a; Urešová, 2011b) and the Engvallex (Cinková, 2006) in the sense of gaining a database of frame-to-frame, and subsequently, argument-to-argument pairs for the purposes of machine translation experiments. The interlinking is gained via semi-manual alignment of the valency slots directly in the ca. 50 000 sentence pairs of the PCEDT. The links are then automatically extracted and stored in a separate file.

The PDT-Vallex[2] has been developed as a resource for annotating argument relations in the Prague Dependency Treebank[3] (Hajičová et al., 2010) and later used also for annotation of the Czech part of Czech-English parallel dependency data, the PCEDT. Valency frames in the PDT-Vallex consist of participant slots represented by tectogrammatical functors (deep syntactic role labels) and they roughly correspond to individual verb meanings. Each slot is marked as obligatory or optional and its typical morphological realization forms are listed. Frame entries are supplemented with illustrative sentence examples.

The Engvallex[4] was created as an adaptation of the already existing resource of English verb argument structure characteristics, the Propbank (Kingsbury and Palmer, 2002). The original Propbank argument structure frames have been adapted to the scheme of the PDT-Vallex, though some minor deflections from the original scheme have been allowed in order to save some important theoretical features of the original Propbank annotation.

[2]http://hdl.handle.net/11858/00-097C-0000-0023-4338-F

[3]http://hdl.handle.net/11858/00-097C-0000-0001-B098-5

[4]http://hdl.handle.net/11858/00-097C-0000-0023-4337-2

Both the PCEDT and its incorporated valency lexicons are sheltered under the FGD theoretical framework (Panevová, 1994). providing a thorough linguistic approach to deep syntactic roles labelling.

In the project, we also deal with semantic categories and verb classes. Since this topic is not covered within the FGD theory, we have consulted other available resources of native speaker's intuition regarding valency characteristic of English verbs: the Propbank, the Framenet (Ruppenhofer et al., 2006) and Levin's classification (as presented in (Levin, 1993) and used in Verbnet (Schuler, 2005)). We are more than aware of the fact that theoretical grounds for each of the mentioned projects differ in several aspects, nevertheless, frequent attempts to interrelate the resources (Palmer, 2009) show that they are in a way comparable.

3. Argument Labelling in the FGD Framework

In the FGD, five *actants*, i.e. arguments of a valency frame, are recognized: ACT (Actor), PAT (Patient), ADDR (Addressee), EFF (Effect) and ORIG (Origin). In the theoretical framework, it is declared that ACT and PAT stand for more general concepts of "the first" and "the second" argument in the valency structure, in other words, these positions are described more syntactically than semantically. In the valency frame, the actants are subject to the so-called "shifting of cognitive roles". If a verb has two arguments only, semantic Effect, semantic Addressee and/or semantic Origin are being shifted to the Patient position. Whereas such a conception appears fitting for monolingual data, unfortunately, it brings certain degree of confusion to the parallel data. A certain amount of argument mismatches found in the parallel data are caused by semantic arguments being "renamed" due to the shifting principle.

When it comes to actant role labels, the otherwise complex FGD theory appears surprisingly vague. Though much attention is paid to the criteria for the theoretical distinction of actant and free modifier roles, and for the distinction of obligatory and optional positions in FGD, only little is said about the nature of the individual actant roles per se. It is somehow taken for granted that native speaker intuition in this respect recognizes semantic aspects of the actant roles well, but the annotation practice often shows otherwise. The annotation guidelines usually offer a broad description for each of the roles (e.g. PAT = "affected object") and a non-exhaustive list of prototypical semantic modifications. Nevertheless, the treebank data offer a range of examples for which such a description comes short.

4. Analysis of Selected Issues

Though the majority of verb-to-verb alignments in PCEDT is consistent in the argument mapping, we can find mismatches for each individual deep syntactic argument label. In this paper, we focus on a selection of these argument mismatches. We concentrate on those argument non-correspondence including an ADDR position on either side of the translation. Of these we have chosen three compact semantic verb classes, each including verbs with three major roles in their semantic structure. It must be said that

ACT ADDR PAT	ACT PAT CAUS
accuse – obvinit	
obviňovat – accuse	
vinit – accuse	
charge (with) – obvinit	
charge (with)	obžalovat
charge (with)	žalovat
obvinit	blame (for)
připisovat – blame (on)	
připisovat	blame (for)
přisuzovat – blame (on)	
přičítat – blame (on)	
obvinit – convict	
usvědčit – convict	
obvinit	fault
reprove	odsuzovat
sue	žalovat
sue – soudit se	
vytknout	chastise
	kárat – chastise
	potrestat – chastise
	kritizovat – criticize

Table 1: Frame distribution for selected Czech and English Judgement Verbs in the PCEDT

though we have searched the treebank for ADDR argument mismatches, usually, ADDR was not the source of the mismatch issue. More likely, it was the non-agent-non-addressee argument in the structure that carried the controversy, whereas ADDR was linked to a PAT argument whose semantics was addressee-like and its label was just an outcome of the application of the shifting principle.

4.1. Verbs of Judgement

The first class of verbs which gained our attention were verbs of judgement and communicating judgment.[5]
In our sample, we have looked at the following verbs: *accuse, blame, charge, chastise, convict, criticize, fault, reprove, sue* and their Czech equivalents. According to the three resources of English verbs argument structures, these verbs share the following argument roles: *the judge*, *the judged entity* and *the reason for judgment*. In the PCEDT (and its valency lexicons), the annotation practice is divided as shown in Table 1.[6] The individual rows of the table represent selected translation verb pairs, columns show the distribution of the verbs among different frames. If both verbs of the translation pair belong to the same frame, they are both inscribed in the same cell.
The split of the annotation is apparently caused by different approaches to *the reason for judgement* argument. Either it is interpreted as an actant, i.e. belonging to the valency structure, or it is considered an adjunct, a free modification external to the valency structure. Unfortunately, the question of argumenthood, i.e. "what exactly is an argument (theta role, participant etc.) and how many of them there really are", has not been satisfyingly answered yet. A very nice and summarizing debate of this issue can be found e.g. in (Dowty, 1991), concluding that while using criteria from different levels of linguistic description for describing argumenthood, we on the one hand aim at more exact linguistic description, but on the other hand we end up with more confusion and theoretical clashes. Since there is no significant difference in the verb semantics, the difference in labelling of *the reason for judgement* may be the result of the influence of its morphological form, which is or is not imposed by the verb. As opposed to the direct object form which builds almost immediately the actant interpretation, prepositional phrases are ambiguous with respect to possible interpretations. Considering the third argument itself, there are equally relevant criteria for both interpretations (Patient and Cause). The semantics of the argument in question bears causal features (Framenet e.g. names this role *Reason*). On the other hand, it is often expressed (in lexicalized alternations (Kettnerová, 2012) of the verbs in question) in a direct object position, which is typical for Patient and atypical for Cause. Nevertheless, there is one important theoretical difference between a Patient and a Cause in the FGD. A Patient is an actant role, thus being considered a part of valency structure of the verb whether syntactically obligatory or not. On the other hand, Cause is an adjunct role, thus being considered part of the valency structure of the verb only when syntactically obligatory. Interestingly, due to the lack of reliable criteria for obligatoriness in the FGD, Cause rarely appears as obligatory in the annotation practice. Rather, phrases with causal semantics and syntactically obligatory character appear to be labelled consistently as Patients.
With respect to the fact that the verbs in question to a great extent share both the grammatical behaviour and semantic features, we argue that their annotation within the dependency treebank and the valency lexicons should be uniform. We have encountered no serious theoretical contradiction in the FGD approach which would speak against such a unification of the description. Taking into account the analysis above, we propose that the resultant valency frames for judgement verbs should be ACT ADDR PAT. The main advantages of the solution are the following:

- The annotation stays consistent with other theoretical approaches which consider *the reason for judgement* a part of the inner argument structure of a judgement verb, disregarding its actual morphosyntactic form.
- It enables us to treat uniformly all judgement verbs having both *the judged entity* and *the reason for judgement* in their argument structure. As a result, the tectogrammatical structures of parallel trees of different languages would appear more similar.
- Such labelling enables us to treat uniformly lexicalized alternations for individual verbs.

[5] In Framenet, these two categories are considered separate, for our purposes it seems convenient to treat them jointly, e.g. as they appear in Levin's classification.

[6] In this paper, in order to stay as clear as possible, we oversimplify the FGD valency theory. We disregard the question of obligatoriness and we treat FGD adjunct labels (such as CAUS or REG) as parts of the valency frame when occupying a relevant position in the conceptual structure.

4.2. Verbs of Teaching

A similar, though less complicated situation can be observed in the class of Teaching verbs. Here we deal with the functor label non-correspondence at the position of the argument describing the taught *subject, or skill*. The consulted resources of valency characteristics for English differ in the number of acknowledged participants of the valency structure. Whereas Propbank and Verbnet[7] distinguish in accordance with PDT-Vallex and Engvallex three arguments, Framenet splits the taught skill argument into more semantic labels (Subject, Skill, Precept, Fact), disregarding the fact that normally they occupy a single syntactic position and are therefore in a complementary distribution. From Table 2 we can see that the verbs in question behave rather homogeneously throughout the languages. Most Czech verbs share the valency frame ACT PAT REG, probably reflecting the fact that the *Subject/Skill* argument is restricted as for the morphosyntactic form – in the majority of cases the form of prepositional phrase v+loc (in+loc) is the only available, it is not imposed syntactically by the verb and it is also a typical form expressing the "regard" semantics. There are two exception being assigned the ACT ADDR PAT frame. First, there is the verb "učit" (teach), which expresses both the ADDR and the PAT argument with accusative, and allows expressing the PAT argument with infinitive, i.e. forms typical of actants and always imposed syntactically by the governing verb. And second, rather surprisingly and from unknown reasons, verb "školit", the derived form of which "vyškolit" has the ACT PAT REG valency frame.

The fact that English verbs share (with the exception of "coach") the ACT ADDR PAT frame can be attributed to the fact that Engvallex has been generated from Propbank lexicon, therefore the annotators were likely to assign actant labels to Propbank roles if possible (in order to keep as many roles in the frame as possible). On the other hand the fact that Czech verbs of Teaching have been assigned ACT PAT REG frame is probably connected to two facts: the most frequent morphosyntactic form with which the Subject/Skill argument is expressed (v+loc) is a typical "regard" form, and there is a theoretical possibility of expressing the *Subject/Skill* argument with other morphosyntactic forms typical of "regard" adjunct and at the same time highly non-typical of any actant role (secondary prepositions "co do", "ohledně" etc). Such sentences would not be ungrammatical and would be easily understandable to a native speaker. Nevertheless, if we look up the verbs in question in the Czech National Corpus, section SYN2010 (Křen, 2009), we can see, that there is no occurence of the secondary preposition phrase realizations of the argument, which makes the point rather hypothetical. Therefore, we propose that the verbs of teaching should also be annotated in a unified manner, with the ACT ADDR PAT frame.[8]

[7] We are aware of the fact that Verbnet even does not count the verbs "teach, coach, train, educate etc." in the same class. According to Verbnet, each of the verbs in question belongs to a different class.

[8] In order to stay consistent with basic principles of FGD, in case a secondary preposition of the "regarding" type appeared, the PAT argument might be considered overtly unexpressed and the REG label should be assigned.

Table 2: Frame distribution for selected Czech and English Verbs of Teaching in the PCEDT

ACT ADDR PAT	ACT PAT REG
train	trénovat
train – školit	
train	vyškolit
train	vycvičit
educate	vzdělávat
teach – (na)učit	
	coach – vyškolit

4.3. Attempt Suasion

With verbs of Attempt Suasion we encounter a different issue. As can be seen in Table 3, the valency frame attribution to each of the verbs is quite consistent within the individual languages. There are three participants in the frame, which can be labeled as *Agent*, *Addressee* and *Desired Action*. All the verbs found in the data are treated quite uniformly in various valency characteristics and semantic class resources. The frames include actants only, so there is no controversy considering the status of the participants. Moreover, there is no significant difference across the Czech verbs considering the morphosyntactic forms the individual tectogrammatic actant labels are usually assigned to. Some of the Czech verbs mentioned express *the Addressee* with dative case, others with accusative, and others with the prepositional phrase na+acc (on+acc), while all the forms being comparably frequent and all of them being imposed by the verb syntactically. Similarly, the *Desired Action* participant may be expressed with infinitive, subordinate clause or prepositional phrase without any serious inferences drawn considering the semantics.

The only semantic difference between the two frames in question, since the PAT position is the syntactic position for argument "shifting", is that in case of ACT ADDR PAT the ADDR position is accented as semantically "distinctive", whereas in the case of ACT PAT EFF, it is the EFF (result) position. Actually, there exists an annotation manual suggestion, that the ADDR frame should be used with verbs with which the ADDR position is typically involving an animate entity. In this respect we may claim that assigning the ACT PAT EFF frame should be considered a lexicon annotator mistake and that the lexicon entries shall all be corrected to the ACT ADDR PAT variant.

5. Conclusion

On the example of verbs of three different verb classes, we have sketched the way in which contrasting different resources can reveal different aspects of capturing valency across verbs similar in meaning and across languages. The FGD argument labelling is highly dependent on the morphosyntactic realization of the individual participants of the structure. If the morphosyntactic form is imposed by the governing verb, an actant label is used. Otherwise the annotation practice seems to be variable. Still, it appears to

Table 3: Frame distribution for selected Czech and English Verbs of Attempt Suasion in the PCEDT

ACT ADDR PAT	ACT PAT EFF
donutit	press
naléhat	press
tlačit	press
pressure – tlačit	
nařídit	direct
nutit	urge
naléhat	urge
dotlačit	urge
push – dotlačit	

be possible to achieve unification of the annotation practice to a considerable extent without the need for a deeper (conceptual) layer of linguistic description.

6. Acknowledgements

The project has been partially supported by the grant No. GPP406/13/03351P of the Grant Agency of the Czech Republic
This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

7. References

S. Cinková. 2006. From Propbank to Engvallex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy*.

D. Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

J. Hajič, E. Hajičová, J. Panevová, P. Sgall, O. Bojar, S. Cinková, E. Fučíková, M. Mikulová, P. Pajas, J. Popelka, J. Semecký, J. Šindlerová, J. Štěpánek, J. Toman, Z. Urešová, and Z. Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey*, pages 3153–3160.

E. Hajičová and P. Sgall. 2003. Dependency Syntax in Functional Generative Description. *Dependenz und Valenz–Dependency and Valency*, 1:570–592.

E. Hajičová, A. Abeillé, J. Hajič, J. Mírovský, and Z. Urešová. 2010. Treebank Annotation. In *Handbook of Natural Language Processing, Second Edition*, pages 167–188. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, ISBN 978-1-4200-8592-1, pp. 167-188, 678 pp.

V. Kettnerová. 2012. *Lexikálně-sémantické konverze ve valenčním slovníku*. Ph.D. thesis, Charles University, Prague, Czech Republic.

P. Kingsbury and M. Palmer. 2002. From Treebank to Propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993. Citeseer.

M. Křen. 2009. The SYN Concept: Towards One-Billion Corpus of Czech. In *Proceedings of the Corpus Linguistics Conference*.

B. Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago and London, ISBN 0-226-47533-6, 348 pp.

M. Palmer. 2009. Semlink: Linking Propbank, Verbnet and Framenet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15.

J. Panevová. 1994. Valency frames and the meaning of the sentence. *The Prague School of Structural and Functional Linguistics*, 41:223–243.

J. Ruppenhofer, M. Ellsworth, M. RL Petruck, Ch. R Johnson, and J. Scheffczyk. 2006. Framenet II: Extended theory and practice.

K. K. Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia.

P. Sgall. 2006. *Language in Its Multifarious Aspects*. Charles University in Prague, The Karolinum Press, Prague, ISBN 80-246-1158-9, 556 pp.

Z. Urešová. 2011a. *Valence sloves v Pražském závislostním korpusu*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, ISBN 978-80-904571-0-2, 229 pp.

Z. Urešová. 2011b. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, ISBN 978-80-904571-1-9, 375 pp.