Zero Alignment of Verb Arguments in a Parallel Treebank

Jana Šindlerová **Eva Fučíková** **Zdeňka Urešová**
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Czech Republic
{sindlerova,uresova,fucikova}@ufal.mff.cuni.cz

Abstract

This paper analyses several points of interlingual dependency mismatch on the material of a parallel Czech-English dependency treebank. Particularly, the points of alignment mismatch between the valency frame arguments of the corresponding verbs are observed and described. The attention is drawn to the question whether such mismatches stem from the inherent semantic properties of the individual languages, or from the character of the used linguistic theory. Comments are made on the possible shifts in meaning. The authors use the findings to make predictions about possible machine translation implementation of the data.

1 Introduction

In Machine translation tasks lately, paraphrases have been used and studied intensely. They basically serve to improve the evaluation metrics of MT systems. The ability to generate valid paraphrases also plays an important role in information retrieval tasks, textual entailment etc. The so-called paraphrase tables can be automatically extracted from parallel corpora (Denkowski and Lavie, 2010; Ganitkevitch et al., 2013).

So far, only lexical paraphrases have been explored for Czech (Barančíková et al., 2014), with syntactic (structural) paraphrases intended for future enhancement of the systems. For English, experiments with both lexical and syntactic paraphrases are employed (Dorr et al., 2004).

This paper presents a preliminary linguistic analysis of structural paraphrases based on valency representations. It appears that certain types of paraphrases affect the valency structure of verbs, and possibly the semantic structure of the sentence, in terms of foregrounding or backgrounding different arguments.[1]

We believe that the analysis of possible syntactic variation within paraphrases, especially such that involves a kind of “disproportion”, in the parallel treebank data, would be beneficial for further MT experiments.

By a disproportion in dependencies, we mean such structural configurations that involve different number of dependencies in corresponding syntactic structures, i.e., an alignment of “something” on one side of the translation to “nothing” on the other side. For the purposes of this paper, we call it a “zero alignment”.

2 Related Work

The analysis in this paper goes in a similar direction as that of (Sanguinetti et al., 2013), though our interest in what they call a “translation shift” is of a different kind. The authors claim that dependency structures are finely apt to account for the alignment of syntactically different treelets between languages, because of the subtree structures constituting similar semantic units. We take their findings as our starting point and provide a linguistic analysis of some of the well-identified categories of translation shift from their research, in order to get a better understanding of different linguistic grounds for different syntactic structures for a parallel semantic content. Also, our analysis is based on the deep syntactic layer (in contrast to the surface structure alignments used in the paper mentioned above), therefore it does not have to deal with those structural phenomena that might not have important semantic consequences, but only serve for topic-focus hierarchization purposes (such as word order variation, simple passivization etc.).

[1] Here, we use the label "argument" in a simplifying manner. Any element which is included in the valency frame is referred to as an argument.

Our research is also inspired by (Bojar et al., 2013), an attempt to generate as many possible translation paraphrases as possible, in order to enlarge the reference set of translations for MT evaluation purposes. The experiment described in the paper used mostly a flat approach, and was carried out with substantial work provided by human annotators. We believe that our research might help establish rules for automatic extraction of true syntactic paraphrases (without unnecessary noise) from parallel corpora, based on the valency patterns of words, so that most of the work could be done automatically, with minimal human control.

3 Methodology and Data

In the research, we took the advantage of the existence of Czech-English parallel data, namely the Prague Czech-English Dependency Treebank 2.5 (PCEDT 2.5) (Hajič et al., 2012).[2]

It is a collection of about 50 000 sentences, taken from the Wall Street Journal part of Penn treebank (Marcus et al., 1993),[3] translated manually to Czech, transformed into dependency trees and annotated at the level of deep syntactic relations (called tectogrammatic layer). In short, the tectogrammatic layer contains mostly content words (with several defined exceptions) connected with oriented edges and labelled with syntactico-semantic functors according to the Functional Generative Description approach (FGD), see (Sgall et al., 1986). Ellipsis and anaphora resolution is also included, as well as automatic alignment of corresponding nodes. The PCEDT 2.5 is annotated according to the the FGD valency theory (FGDVT) and two valency lexicons (one for each language) are part of the release.

PDT-Vallex[4] (Hajič et al., 2003; Urešová, 2011) has been developed as a resource for annotating argument relations in the Prague Dependency Treebank (Hajič et al., 2006). The version used here contains 11,933 valency frames for 7,121 verbs. Each valency frame in the PDT-Vallex represents a distinct verb meaning. Valency frames consist of argument slots represented by tectogrammatic functors (slots). Each slot is marked as obligatory or facultative and its typical morphological realization forms are listed. Frame entries are supplemented with illustrative sentence examples.

EngVallex[5] (Cinková, 2006) was created as an adaptation of an already existing resource of English verb argument structure characteristics, the Propbank (Palmer et al., 2005). The original Propbank argument structure frames have been adapted to the FGD scheme, so that it currently bears the structure 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. This lexicon includes 7,148 valency frames for 4,337 verbs.

PDT-Vallex and EngVallex have been interlinked together into a new resource called CzEngVallex (Urešová et al., 2015a; Urešová et al., 2015). Beside the complete data of the two lexicons, the CzEngVallex contains a database of frame-to-frame, and subsequently, argument-to-argument pairs for the purposes of machine translation experiments (Urešová et al., 2015b). PCEDT and the CzEngVallex data have already been used successfully in several MT experiments aimed at valency frame detection and selection (Dušek et al., 2014) and also for word sense disambiguation (Dušek et al., 2015).

The interlinking of CzEngVallex frames was carried out via an annotation over the PCEDT. First, an automatic alignment procedure was run over the data, which suggested translational links between nodes of the tectogrammatic layer. Corresponding verb pairs[6] and argument pairs were highlighted. Then, manual revision and correction of the alignments by two annotators was carried out. Thus, as a by-product of building the lexicon, a collection of illustrative annotated tree pairs is available for each verb pair of the CzEngVallex.

4 Zero Alignment in the Data

In the following sections, we will describe the most important, consistent and frequent points of zero alignment found in the data. For each section, we will comment on the linguistic background of the phenomena described and the possible consequences for semantic interpretation in the individual languages.

[2] https://catalog.ldc.upenn.edu/LDC2012T08

[3] https://catalog.ldc.upenn.edu/LDC99T42

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

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

[6] As a basic stage of building the CzEngVallex, only verb-verb pairs were taken into account.

4.1 Catenative Verbs - Single vs. Double Object Interpretation

One of the prominent points of alignment disproportion in the data are sentences with catenative verbs. Catenative verbs are usually defined as those combining with non-finite verbal forms. Between the finite catenative verb and the non-finite verb form, there might appear an intervening NP that might be interpreted as the subject of the dependent verbal form. In this section, we will be concerned with exactly those verbs allowing the sequence of a finite catenative verb – NP – a non-finite catenative verb.

4.1.1 ECM Constructions, Raising to Object

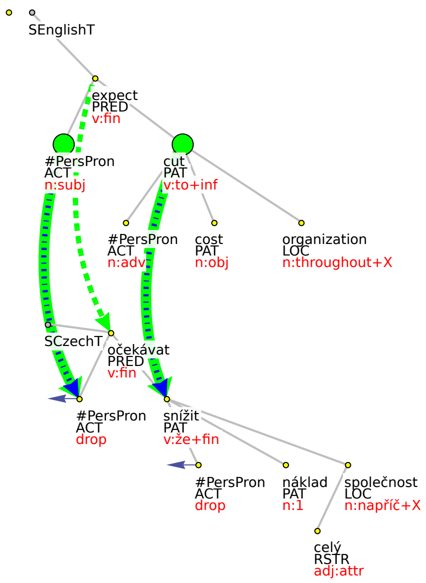
Most Czech linguistic approaches do not recognize the term Exceptional Case Marking (ECM) in the sense of "raising to object", instead they generally address similar constructions under the label "accusative with infinitive". The difference between ECM and control verbs is not being taken into account in most of Czech grammars. In short, raising and ECM are generally considered a marginal phenomenon in Czech and are not being treated conceptually (Panevová, 1996), except for several attempts to describe agreement issues, e.g., the morphological behaviour of predicative complements described in a phrase structure grammar formalism (Przepiórkowski and Rosen, 2005).

The reason for this negligent approach to ECM is probably rooted in the low frequency of ECM constructions in Czech. Czech sentences corresponding to English sentences with ECM mostly do not allow catenative constructions. They usually involve a standard dependent clause with a finite verb, see Fig.1,[7] or they include a nominalization, thus keeping the structures strictly parallel.

The only exception are verbs of perception (*see, hear*), which usually allow both ways of Czech translation – with an accusative NP followed by a non-finite verb form (1a), or with a dependent clause (1b), not speaking about the third possibility involving an accusative NP followed by a dependent clause (1c).

(1) He saw Peter coming.

a. Viděl Petra přicházet.
He saw Peter.ACC to come.



En: They expect him to cut costs...

Cz: Očekávají, že sníží náklady...

Figure 1: Alignment of the ECM construction

b. Viděl, že Petr přichází.
He saw that Peter.ACC is coming.

c. Viděl Petra, jak přichází.
He saw Peter.ACC, how is coming.

In this type of accusative-infinitive sequence, the accusative element is in FGDVT analysed consistently as the direct object of the matrix verb (the PATient argument) and the non-finite verb form then as the predicative complement of the verb (the EFFect argument).

The PCEDT annotation of verbs of perception is shown in Fig. 2, with frame arguments mapped in the following way:

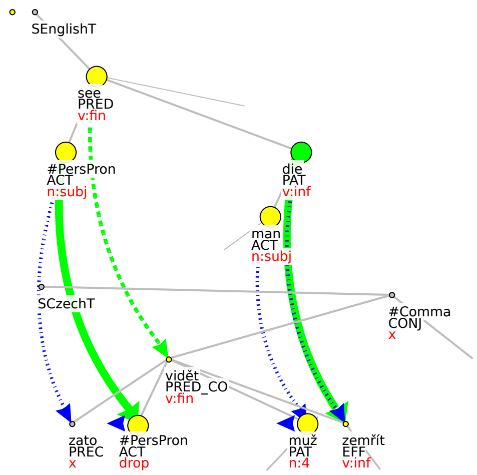
ACT→ACT; PAT→EFF; ----→PAT

The literature mentions two ways of ECM structural analysis, a flat one, representing the NP as dependent on the matrix verb, and a layered one, representing the intervening NP as the subject of the dependent verb. This mirrors the opinion that verbs allowing ECM usually have three syntactic, but only two semantic arguments. It is then a matter of decision between a syntactic and semantic approach to tree construction.

The English part of the PCEDT data was annotated in the layered manner,[8] thus most of the pairs in the treebank appear as strictly parallel. The consistency of structures is one of the most impor-

[7] In the examples displayed, the green dashed lines connect the annotated verb pair, the dotted lines connect verb dependents, the thick arrows mark collected verb arguments, the automatic node alignment is displayed in blue, the manually corrected alignment is marked in red. The images have been cropped or otherwise adjusted for the sake of clarity.

[8] The annotation followed the original phrasal annotation of the data in the Penn Treebank.



En: I have seen [one or two] men die...
Cz: Zato jsem viděla [jednoho nebo dva] muže zemřít...

Figure 2: Alignment of the perception verbs' arguments. The corresponding arguments man-muž are interpreted as belonging to verbs in different levels of the structure.

tant advantages of the layered approach; there is no need of having two distinct valency frames for the two syntactic constructions of the verb, therefore, the semantic relatedness of the verb forms is kept. Also, there are other specific constructions supporting the layered analysis for English, like the there-constructions intervening instead of the NP, see (2).

(2) We expected there to be slow growth.

On the other hand, the Czech part of the PCEDT data uses flat annotation, partly because the catenative construction with raising structure is fairly uncommon in Czech (cf. Sect. 4.1.1). The flat structure is easier to interpret, or translate in a morphologically correct way to the surface realization, but it requires multiple frames for semantically similar verb forms (the instances of the verb *to see* in *see the house fall* and *see the house* are in the FGD valency approach considered two distinct lexical units) and it also leaves alignment mismatches in the parallel data.

The treatment of ECM constructions in English and in Czech is different. It reflects both the differences internal to the languages and their consequences in theoretical thinking. Contrary to English, Czech nouns carry strong indicators of morphology - case, number and gender. The rules for the subject-verb agreement block overt realization of subjects of the infinitives. The accusative ending naturally leads to the interpretation of the presumed subject of the infinitive as the object of the matrix verb. The morphosyntactic representation is taken as a strong argument for using a flat structure in the semantic representation, and a covert co-referential element for filling the "empty" ACTor position of the infinitive. In English, in general, there is no such strong indication and therefore the layered structure is preferred in the semantic representation.

4.1.2 Object Control Verbs, Equi Verbs, Causatives

Contrary to the ECM constructions, object control verbs constructions (OCV), involving verbs such as *make, cause, or get*, are analyzed strictly as double-object in both languages, i.e., the intervening NP is dependent on the matrix verb (and licensed by it) and there is usually a co-referential empty element of some kind in the valency structure of the dependent verb form. OCV constructions are similarly frequent in Czech and English and their alignment in the PCEDT data is balanced, see Fig. 3.[9]

Interestingly, it is sometimes the case that English control verbs in the treebank are translated with non-control, non-catenative verbs on the Czech side, and the intervening NP is transformed to a dependent of the lower verb of the dependent clause (see Fig. 4), or even a more complex nominalization of the dependent structure is used.

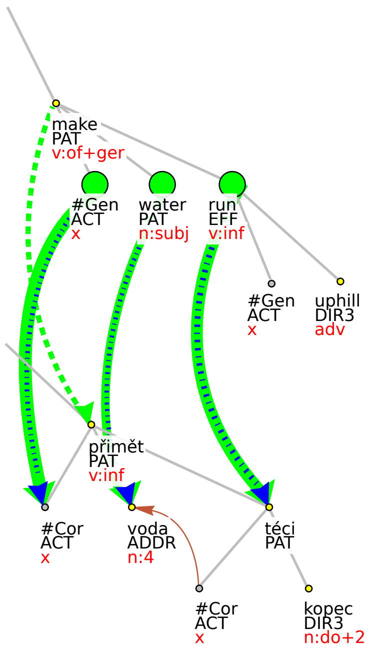
The verb involved in this kind of translation shift may be either a more remote synonym, or a conversive verb.[10]

Such a translation shift brings about (at least a slight) semantic shift in the interpretation, usually in the sense of de-causativisation of the meaning (*prompt→lead to*).[11] Nevertheless, this type of semantic shift does not prevent the use of the struc-

[9] In Fig. 3, English ACT of *run* does not show the coreference link to *water* since the annotation of coreferential relations has not yet been completed on the English side of the PCEDT, as opposed to the Czech side (cf. the coreference link from ACT of *téci* to *voda*).

[10] Semantic conversion in our understanding relates different lexical units, or different meanings of the same lexical unit, which share the same situational meaning. The valency frames of conversive verbs can differ in the number and type of valency complementations, their obligatoriness or morphemic forms. Prototypically, semantic conversion involves permutation of situational participants.

[11] Note that the de-causativisation process is possible without objections whereas the reverse shift, from non-control verb to a control verb, is rare if it at all exists.



En: ...making water run...

Cz: ...přimět vodu téct...

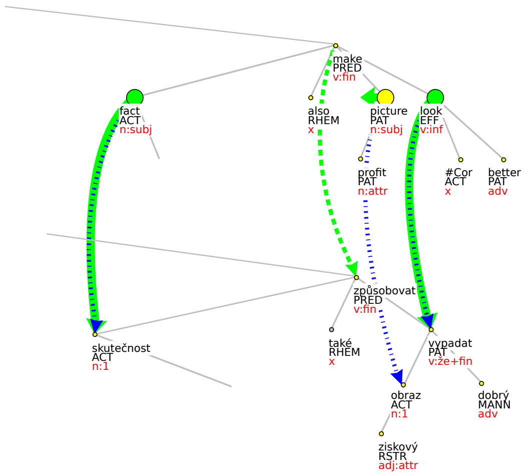
Figure 3: Alignment of the control verbs' arguments

ture as a sufficiently equivalent expression of the semantic content. We approach this as an inherent property of (any) language to suppress certain aspects of meaning without losing the general sense of synonymity.

4.2 Complex Predication

By "complex predication" we mean a combination of two lexical units, usually a (semantically empty, or "light") verb and a noun (carrying main lexical meaning and marked with CPHR functor in the data), forming a predicate with a single semantic reference, e.g., *to make an announcement*, *to undertake preparations*, *to get an order*. There are some direct consequences for the syntactically annotated parallel data.

First type of zero alignment is connected to the fact that a complex predication in one language can be easily translated with a one-word reference, and consequently aligned to a one-word predication, in the other language. This is quite a trivial case. In the data, then, one component of the complex predication remains unaligned. There are basically two ways of resolving such cases: either one can align the light verb with the full verb in the other language, or one can align the full verb with the dependent noun in the complex predication, based on the similarity of semantic content. In the CzEngVallex, the decision was to align the verbs, reflecting the fact that the verb and the noun phrase form a single unit from the semantic point of view.



En: The fact... ...will also make the profit picture look...

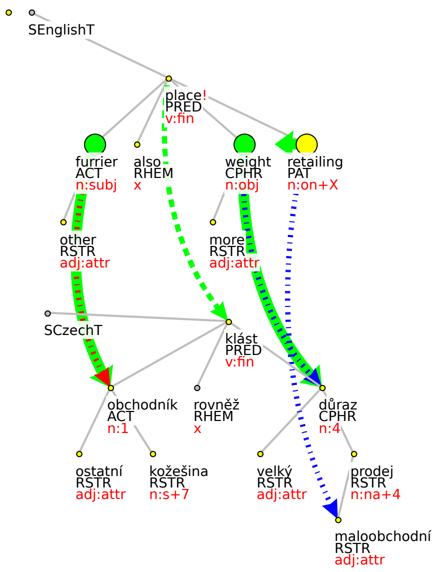
Cz: Skutečnost......způsobuje, že ziskový obraz vypadá...

Figure 4: Alignment of English OCV with Czech non-OCV construction

The second type of zero alignment is connected to the presence of a "third" valency argument within the complex predication structure, e.g., En: *placed weight on retailing* - Cz: *klást důraz na prodej*, see Fig. 5.

Complex predicates have been annotated according to quite a complicated set of rules on the Czech side of the PCEDT data (for details, see (Mikulová et al., 2006)). Those rules include also the so-called dual function of a valency modification. There are two possible dependency positions for the "third" valency argument of the complex predicate: either it is modelled as the dependent of the semantically empty verb, or as a dependent of the nominal component. The decision between the two positions rely on multiple factors, such as valency structure of the semantically full use of the verb, valency structure of the noun in other contexts, behaviour of synonymous verbs etc. On the Czech side, the "third" valency argument was strongly preferred to be a dependent of the nominal component.

On the English side of the PCEDT, the preferred decision was different. The "third" argument was annotated as a direct dependent of the light verb



En: Other furriers have also placed more weight on retailing.

Cz: Ostatní obchodníci s kožešinami rovněž kladou větší důraz na maloobchodní prodej.

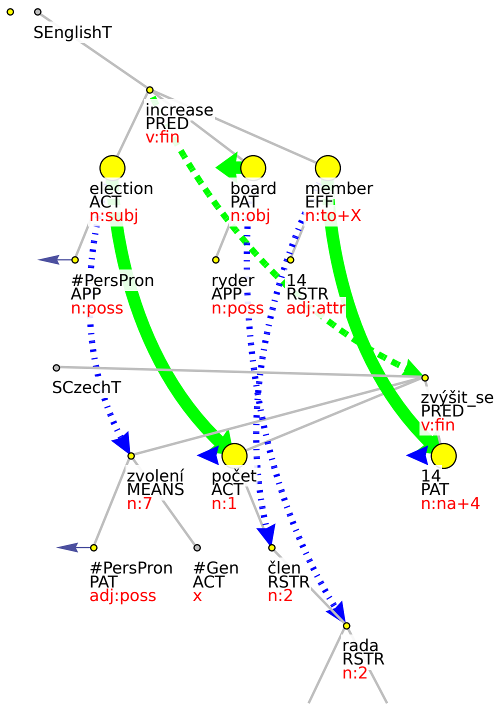
Figure 5: Mismatch due to complex predication solution

(probably due to lower confidence of non-native speaker annotators in judging verb valency issues).

There is probably no chance of dealing with the dependencies in one of the two above stated ways only. The class of complex predicates in the data is wide and heterogeneous with respect to semantic and morphosyntactic qualities. Nevertheless, the data suggest several points of interesting inconsistencies stemming from the imperfection or lack of reliability of the theoretical guidelines. For example, the dependency of the valency complementation of the complex predicate *klást důraz* 'place emphasis', as can be seen in Fig. 5, is solved as a dependency on the nominal component, whereas in the complex predicate *klást požadavek* 'place claim', the valency lexicon entry involves a direct dependency on the verb. Keeping in mind that the verb *klást* 'to place' has three arguments in its semantically full occurrences, we would expect direct dependency on the verb in both cases.

4.3 Conversive Verbs

A considerable number of unaligned arguments in the data is caused by the translator's choice of a verb in a conversive relation to the verb used in the original language. For some reason (e.g., frequency of the verbal lexical unit, topic-focus articulation etc.), the translator decides not to use the syntactically most similar lexical unit, but uses a conversive one (cf. also Sect. 4.1.2), thus causing the arguments to relocate in the deep syntactic structure, see Fig. 6.



En: His election increases Ryder's board to 14 members.

Cz: Jeho zvolením se počet členů správní rady společnosti Ryder zvýšíl na 14.

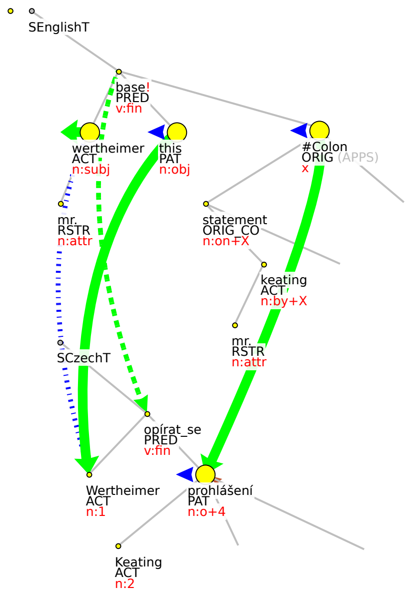
Figure 6: Mismatch due to the the use of conversive verbs

The relocation of arguments frequently goes together with backgrounding of one of the arguments, which then either disappears from the translation, or is transformed into an adjunct, or into a dependent argument embedded even lower in the structure.

The first argument (actant)[12] in the FGD approach is strongly underspecified. It is mostly delimited by its position in the tectogrammatic annotation. Its prevalent morphosyntactic realization is nominative case, but certain exceptions are recognized (verbs of feeling etc.). Also, the ACT position (first actant) is subject to the process called "shifting of cognitive roles" (Panevová, 1974), i.e., other semantic roles can take the nominative case and the corresponding place in the structure

[12] Under the term "actant", FGDVT distinguishes five core constituting valency complementations, ACT, PAT, ADDR, EFF, and ORIG.

in case there is no semantic agent in the structure. Thus we get semantically quite different elements (e.g., +anim vs. -anim) in the ACT position, even with formally identical verb instances, see the English side of Figs. 7 and 8.



En: Mr. Wertheimer based this on a statement by Mr. Keating...

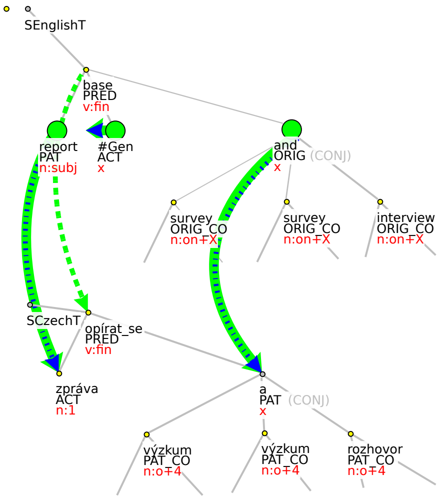
Cz: Wertheimer se opírá o prohlášení Keatinga...

Figure 7: Conflict due to the underspecification of the ACT position

This formal feature of the FGDVT gives rise to a number of conflicts in the parallel structures considering structures that undergo semantic de-agentization or (milder) de-concretization of the agent.

Here the question arises, whether such verb instances correspond to different meanings of the verb (represented by different verb frames), or whether they correspond to a single meaning (represented by a single valency frame). It is often the case, that the Czech data tend to overgeneralize the valency frames through considering the different instances as realizations of a single deep syntactic valency frame, when there is no other modification intervening in the frame. Therefore, this approach chosen for the Czech annotation sometimes shows a conflict, as in Fig. 7.

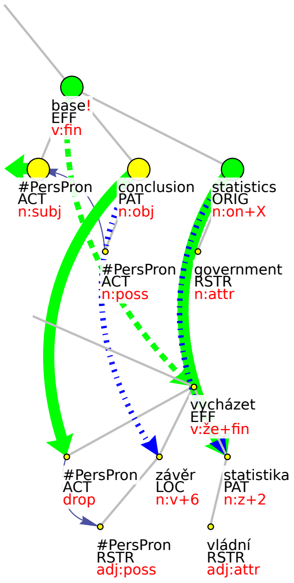
The valency structure for both instances of *base* is identical, only in the first case, the verb is used in active voice, whereas in the second case, it takes passive morphology. There are three semantic arguments in the structure. We will call them the Person that expresses an opinion, the Expressed Opinion and the Resource for the opinion. The Person bases the Expressed Opinion on the Resource. With the English verb, the Expressed Opinion always takes the PAT position and the Resource the ORIGin position in the valency structure. On the other hand, on the Czech side of the data, there is a conflict. In both cases, there are seemingly only two arguments. In the first case, the Expressed Opinion is sort of backgrounded from the semantic structure. If there were a need of overtizing it, it would probably appear with locative morphology, as an adjunct: *Wertheimer se **v tomto** opírá o prohlášení...* ‘Wertheimer **in this** relies on a statement’ (see also an authentic example from the data in Fig. 9). In the second case, on the other hand, the structure follows the passivized English structure in backgrounding the Person (note that the *se* morpheme does NOT stand for a passive morphology here). If there were a need for expressing the Person, it would probably appear as a specifying dependent to the ACT position: ***Jejich** zpráva se opírá o telefonický výzkum.* **‘Their** report is based on a phone survey’. In the second case, the Expressed Opinion does not take the PAT position, but the ACT position in the structure, which is the cause of the conflict. We are able to reformulate the first case



En: The report was based on a telephone survey...

Cz: Zpráva se opírá o telefonický výzkum...

Figure 8: Original collect for the verbs *base* and *opírat se*



En: ...they based their conclusions on government statistics.

Cz: ...vycházejí z vládních statistik.

Figure 9: Original collect for the verbs base and vycházet with LOC argument linked to PAT

in a corresponding manner to show the Expressed Opinion argument in the ACT position and the Person backgrounded from the structure, see (3):

(3) a. Wertheimer se ve svém názoru opírá o
Wertheimer REFL in his opinion leans to
prohlášení Keatinga.
the statement by Keating

b. Wertheimerův názor se opírá o
Wertheimer's opinion REFL leans to
prohlášení Keatinga.
the statement by Keating

c. Wertheimer opírá svůj názor o
Wertheimer leans his opinion to
prohlášení Keatinga.
the statement by Keating

The problem of the status of a Czech verbal-adjoining *se*-morpheme is a complex one and there is no clear scientific consensus in this respect. The *se*-morpheme in Czech has a variety of functions, e.g., a passivization morpheme for the so-called "reflexive passive" form, a "dispositional diathesis" morpheme, a reflexive morpheme for lexical derivation of impersonal verbal variants, or an accusative reflexive pronoun.

These variants differ with respect to the way they are reflected in the data and in the lexicon. Some are treated as individual verb lemmas, some as surface variants of a common non-reflexive lemma.

The conflicts in annotation have a substantial reason – the ways in which English and Czech express backgrounding of the agent are multiple and they differ across the languages. Czech uses the *se*-morphemization often, in order to preserve the topic focus articulation (information) structure, whereas English does not have such a morpheme to work with, so it often uses simple passivization, or middle construction.

Moreover, the first valency position in Czech is often overgeneralized, allowing a multitude of semantically different arguments, which is, due to "economy of description", sometimes not reflected in the linguistic theory.

4.4 Arguments Mapped to Adjuncts

In the previous section, we have described the bilingual treebank data manifestation of the fact that languages have different means of expressing a content, and we have noted that these can also variate between argument and adjunct interpretation. This variation appears both within a single language (one language expresses a largely synonymous content with either argument or adjunct means) and across languages (a direct consequence of the former case: an argument (actant) in one language can be translated into another language using an adjunct construction). The language may differ in the preference for either of the possibilities.

Observing such mismatches in a parallel treebank occasionally leads us to hesitate whether our interpretation of an argument as an argument or an adjunct is proper or justifiable. There may be two possible consequences drawn from the observation of a mismatch – either there are some (rather subtle) semantic reasons for formulating an argument as an argument/adjunct, or there is some imperfection in our theoretical thinking about the internal system of a particular language.

The theoretical distinction between arguments and adjuncts is subject to serious debates in the world of linguistics (Hwang, 2011; Tutunjian and Boland, 2008), and so far there is no approach known to us that would overcome this problem easily. Still, we can see that the real data indicate some remarkable points that stand at the roots of the argument/adjunct distinction problem. Most prominently - the nature of the relation between the form of the argument and its semantics.

Let us mention cases like alignment of an actor with a temporal adjunct (4) and an actor with a causal adjunct (5). etc.

(4) Americans haven't forgiven China's leaders for the military assault of June 3-4 ***that*** killed hundreds, and perhaps thousands, of demonstrators.

a. Američané neodpustili čínským vůdcům
Americans haven't forgotten Chinese leaders
vojenský útok z 3.-4. června,
military assault from 3-4 June,
při kterém zahynuly stovky, možná i
during which died hundreds, maybe even
tisíce demonstrantů.
thousands demonstrators

(5) ***The purchase*** will make Quebecor the second-largest commercial printer in North America.

a. ***Díky této koupi*** se společnost
Thanks to this purchase REFL the company
Quebecor stane druhou největší
Quebecor will become second largest
komerční tiskárnou v Severní Americe.
commercial printer in North America

The interpretation of the argument in the above stated examples is driven mainly by its morphological form, which is a surprising finding considering that we are dealing with deep syntax, or even semantics.

It is believed that the form of the expression more or less mirrors its function in the language. The width of the paraphrasing range though, both within and across languages, leads us to questioning whether it is appropriate to lay much stress on the difference between arguments and adjuncts in the description of a language.

5 Conclusion

We have encountered several reasons for the presence of a zero alignment in the data. Though these reasons have different grounds they tend to be interconnected in the language.

1. Language is flexible in paraphrasing linguistic content with different syntactic means. Even pairs of sentences which include semantic backgrounding or foregrounding of different arguments are easily interpreted as synonymous.

2. It is possible to use predicates that are in a conversive relation, or predicates of different complexity.

3. The backgrounding and foregrounding of arguments leads to syntactic relocation of other arguments in the structure, and consequently to the shift in their morphosyntactic properties, to the shift in their valency status, or even to their complete disappearance from the structure.

4. The FGD, having been built on a morphologically rich Czech language, relies strongly on the morphosyntactic form of the individual arguments. Therefore, disproportions of the zero alignment or argument mismatch kind must appear when it is applied to other languages with different typological properties.

Points 1, 2 and 3 belong among inherent deeply rooted properties of (perhaps any) natural language. Such differences are not to be overcome by means of possible theoretical unification of description.

Point 4, on the other hand, belongs to the properties of a certain linguistic theory. We will leave it open, whether it were appropriate to change the very roots of a linguistic theory in order to make it more flexible for use across different languages. Nevertheless, it appears that it is at least possible to change those aspects that cause individual and otherwise unjustifiable conflicts in the data.

Acknowledgements

This work has been supported by the grant GP13-03351P of the Grant Agency of the Czech Rep.

References

P. Barančíková, R. Rosa, and A. Tamchyna. 2014. Improving Evaluation of English-Czech MT through Paraphrasing. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, and J. Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 596–601, Reykjavík, Iceland. European Language Resources Association.

O. Bojar, M. Macháček, A. Tamchyna, and D. Zeman. 2013. Scratching the surface of possible translations. In *Text, Speech, and Dialogue*, pages 465–474. Springer.

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.*

M. Denkowski and A. Lavie. 2010. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In *Proceedings* *of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342. Association for Computational Linguistics.

B. J. Dorr, R. Green, L. Levin, O. Rambow, D. Farwell, N. Habash, S. Helmreich, E. Hovy, K.J. Miller, T. Mitamura, et al. 2004. Semantic annotation and lexico-syntactic paraphrase. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC) Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 47 – 52.

O. Dušek, J. Hajič, and Z. Urešová. 2014. Verbal valency frame detection and selection in Czech and English. In *The 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 6–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

O. Dušek, E. Fučíková, J. Hajič, M. Popel, J. Šindlerová, and Z. Urešová. 2015. Using Parallel Texts and Lexicons for Verbal Word Sense Disambiguation. In *Proceedings of the Third International Conference on Dependency Linguistics, Depling 2015*, page this volume. Uppsala University.

J. Ganitkevitch, B. Van Durme, and Ch. Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

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 LREC*, pages 3153–3160.

J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolářová, and P. Pajas. 2003. PDT-Vallex: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9, page 57–68.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková Razímová, and Z. Urešová. 2006. *Prague Dependency Treebank 2.0*. Number LDC2006T01. Linguistic Data Consortium, Philadelphia, PA, USA.

J. D. Hwang. 2011. Making verb argument adjunct distinctions in English. *Synthesis paper, University of Colorado, Boulder, Colorado.*

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

M. Mikulová, A. Bémová, J. Hajič, E. Hajičová, J. Havelka, V. Kolářová, L. Kučová, M. Lopatková, P. Pajas, J. Panevová, M. Razímová, P. Sgall, J. Štěpánek, Z. Urešová, K. Veselá, and Z. Žabokrtský. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30, Prague, Czech Rep.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

J. Panevová. 1974. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics*, 22:3–40.

J. Panevová. 1996. More remarks on control. *Prague Linguistic Circle Papers*, 2(1):101–120.

A. Przepiórkowski and A. Rosen. 2005. Czech and Polish raising/control with or without structure sharing. 3:33–66.

M. Sanguinetti, C. Bosco, and L. Lesmo. 2013. Dependency and constituency in translation shift analysis. *DepLing 2013*, page 282.

P. Sgall, E.Hajičová, and J. Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht, Reidel, and Prague, Academia, Prague.

D. Tutunjian and J. E. Boland. 2008. Do we need a distinction between arguments and adjuncts? Evidence from psycholinguistic studies of comprehension. *Language and Linguistics Compass*, 2(4):631–646.

Z. Urešová, E. Fučíková, and J. Šindlerová. 2015. CzEngVallex: Mapping Valency between Languages. Technical Report TR-2015-58, Charles University in Prague, Institute of Formal and Applied Lingustics, Prague. To appear at `http://ufal.mff.cuni.cz/techrep/tr58.pdf`.

Z. Urešová, O. Dušek, E. Fučíková, J. Hajič, and J. Šindlerová. 2015b. Bilingual English-Czech valency lexicon linked to a parallel corpus. In *Proceedings of the The 9th Linguistic Annotation Workshop (LAW IX 2015)*, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Z. Urešová, O. Dušek, E. Fučíková, J. Hajič, and J. Šindlerová. 2015a. Bilingual English-Czech Valency Lexicon Linked to a Parallel Corpus. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 124–128, Denver, Colorado, USA, June. Association for Computational Linguistics.