A Multilingual Crossover between Types of Resources

Jan Hajič, Eva Hajičová and Zdeňka Urešová

Charles University, Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics, Malostranské n. 25, 11800 Prague 1, Czech Republic

{hajic,hajicova,uresova}@ufal.mff.cuni.cz

**Abstract.** We present two case studies that demonstrate that co-habitation of various (sub)disciplines of Computational Linguistics and Natural Language Processing can reach novel findings and help the advancement of the field(s). One example covers a study of synonymy by searching for information in different lexical resources with regard to multilinguality, and the other demonstrates on the study of some phenomena concerning information structure and word order in English and Czech how a parallel multilingual and/or multilayered corpus if properly annotated can be used for a study of some aspects of the deep syntactic structure. Both cases support our conviction that the creation of language resources is a fundamental step in the domain of computational linguistics, that well-founded annotation is an important step forward towards both testing the original linguistic theory and developing it further.

**Keywords:** Lexical resources, parallel corpora, treebanks, discourse, information structure, synonymy, syntax, semantics

1. Introduction

Brno and Prague are often understood by local people as competing localities within the same country and Masaryk University in Brno and Charles University in Prague as two competing academic institutions. However, one can also find examples where the two communities live comfortably side by side, and not only that, overlapping of interests are easy to spot. One of such examples is the interest in computational linguistics personalized by the man to whom the current volume is dedicated, Professor Karel Pala. Karel and all the three authors of this contribution have common roots, long ago at different historical periods they all were introduced to the field by Petr Sgall, the pioneer of Czech formal and computational linguistics, frequently collaborated under different projects since then but with concentration and emphasis on different subfields. In our contribution, written on his honor, we would like to demonstrate on two case studies that in our scientific discipline co-habitation may help to reach novel findings and may shift the field a little bit forward. One of the examples concerns a possible advancement in the study of synonymy by searching for information in different lexical resources with regard to multilinguality, and the other demonstrates on the study of some phenomena concerning information structure and word order in English and Czech how a parallel multilingual and/or multilayered corpus if properly annotated can be used for a study of some aspects of the deep syntactic structure. Both cases support our conviction that the creation of language resources is a fundamental step in the domain of computational linguistics, their many-sided theoretically well-founded annotation an important step forward towards both testing the original linguistic theory and developing it further, and a combination of resources of most different types and orientations is still another step forward to make use of different kinds of linguistic (and in some cases non-linguistic) information to achieve novel results.

1. Bilingual lexical resource – verbal synonym classes based on linking syntax and semantics

Verbs can describe many events and states depending on collocates they appear with, which in turn leads to the problem of ambiguity of verbs related to their meanings (senses). In addition, the same verb with no obvious meaning ambiguity can get translated into two or more different verbs in the target language, yet forming a perfect translation conveying the same meaning as in the source language. Take the verb widen in English, seen 32 times in the Penn Treebank (Marcus et al., 1993) – in its Czech translation, 14 different verbs have been found: not only the most direct translation “*rozšířit*”, but also *prohloubit* (lit. *deepen*), *rozrůst se* (lit. *grow [oneself]*), *stoupnout* (lit. *rise*), *zvětšit se* (lit. *enlarge*), *zvyšovat* (lit. *raise*, *get higher*) etc. Immediately, questions arise primarily about synonymy, but also about concrete vs. abstract distinction, relation to valency and argument structure, and more. Different meanings of the same verb, or verb senses, are recorded and described – usually rather implicitly and informally – in both monolingual and bilingual dictionaries and we as humans can understand the sense distinctions well. However, verb senses should be described more precisely and explicitly. How do we know what is the explicit set of senses for any particular verb? Which senses (of different verb lexemes) are synonymous or near synonymous in the broader context of use? It has been shown that if we let different people determine this, even on the same set of examples (i.e., using the same corpus) they inevitably come up with a different set. More precisely, the inter-annotator agreement (Artstein and Poesio, 2008) will be low, regardless of the level of linguistic expertise the annotators might have. Some researchers even go so far as to declare that they “do not believe in word senses” (legendary quote by the lexicographer Sue Atkins (1993)), explained by an article with the same title by Kilgarriff (1997) that it should be interpreted as not believing in pre-determined, fixed set of word senses. Others try to find a sweet spot between a hard-to-agree-on, fine-grained set, represented e.g., by WordNet (Fellbaum, 1998), and a coarse-grained set, which does not provide enough detail–such as VerbNet lexicon (Palmer et al., 2009, Petersen et al. 2016). FrameNet (Baker et al., 1998), an English lexical resource which adds roles and uses semantic frames to group verbs and provide examples of use (based on attested corpus examples) is another well-known resource.

In our CzEngClass project (Urešová et al., 2018), we are exploring (partial) semantic “equivalence” (synonymy or near synonymy) of verbs in a bilingual (Czech-English) setting, heavily using their valency properties. We are thus working towards a multilingual perspective (starting from a bilingual one - Czech–English). In addition, and in accordance with the lifelong view of the late Adam Kilgarriff, we are working “bottom up”, i.e., from the evidence as recorded in our corpora, and not “top-down”, from a predefined set of semantic classes (Kilgarriff, 1997). We believe that using the following existing resources based on the Functional Generative Description theory (FGD; Sgall et al., 1986), and referring also to the new Abstract Meaning Representation (AMR; Banarescu et al., 2013) is helping us proceed in that direction: manually treebanked corpora: PDT and PCEDT (Hajič et al., 2018, Hajič et al., 2012), 1 million tokens with complex annotation each, with part of English PCEDT annotated in AMR; valency lexicons linked to the treebanks: PDT-Vallex (Urešová, 2011), EngVallex (Cinková, 2006), and CzEngVallex (Urešová et al., 2016); and automatically analyzed parallel corpus with 100 mil. tokens (CzEng, Bojar et al., 2012).

We are taking advantage of the aforementioned key resources as well as another FGD-based lexicon VALLEX (Lopatková et al., 2016, 2018) and other available resources, such as VerbNet, PropBank and FrameNet, English and in the future also Czech WordNet (Pala et al., 1994), large monolingual corpora, other parallel corpora such as Interkorp (Čermák and Rosen, 2012) and the NLP tools available for both languages. The result is being stored in the CzEngClass lexicon, schematically depicted in Fig. 1 (Urešová et al., 2018).



Figure 1 Structure of the CzEngClass lexicon and the underlying resources

The new resource can be seen on top of Fig. 1 in the red box: it contains a set of classes, each corresponding to some cognitive concept. Each class contains verbs (more precisely, verb senses defined by their valency frames in the PDT-Vallex and EngVallex lexicons) that can express that particular concept, and that has been found in the bilingual corpora we have used. Most of those verb pairs are also linked to the already existing bilingual valency lexicon CzEngVallex.

Each synonym class has been assigned a set of semantic roles, inspired by either the FrameNet or by the by the formal representation of lexical-semantic conversions as elaborated in (Kettnerová, 2014; Kettnerová et al., 2012) and incorporated into the newer versions of VALLEX. This set of roles is shared by all the verbs in one class. In order to verify a verb membership in the class, its arguments, and sometimes even free modifications, are mapped to this common set of semantic roles. If a role cannot be mapped to an argument, or vice versa, such a verb cannot be a member of that class (minor exceptions aside). Table 1 shows an example of such a mapping for verbs synonymous with the verb *offe*r – *nabídnout*, taken from (Urešová et al., 2018).

|  |  |
| --- | --- |
|  | Roles |
| **Offerer** | **Recipient** | **Entity Offered** | **Entity Received** |
| nabídnout | ACT | ADDR | PAT | EFF |
| offer | ACT | ADDR | PAT | EFF |
| bid | ACT | ADDR | PAT | EFF |
| proffer | ACT | ADDR | PAT | SUBS |
| tender | ACT | ADDR | PAT | SUBS |
| extend | ACT | ADDR | PAT, PAT(RSTR) | SUBS |
| make available | ACT | EFF(BEN) | PAT | SUBS |
| + Restriction: EFF[exchange] |

Table 1 Mappings for the class *offer-nabídnout*

As stated in (Urešová et al, 2018), the CzEngClass lexicon is a contribution to a set of lexical resources important for many NLP tasks, such as event detection and linking, semantic relation extraction, etc. While these tasks often use unlabeled data for training, there is always a need for human-annotated data for evaluation, tuning, etc. Importantly, we believe that such a resource is currently missing in the offerings of lexical resources. In addition, it contains links to all the relevant related resources such as VALLEX, FrameNet, VerbNet, PropBank and WordNet, making it suitable for comparative studies.

1. Bi-lingual and/or multi-layered resource – information structure and linking (deep) syntax and text

One of the main methodological principles of the Prague Linguistic School since its entrance on the linguistic scene was a comparative approach. This approach had a strong effect especially on the studies of Czech anglicists V. Mathesius, J. Vachek and their followers such as J. Firbas and L. Dušková. It is then quite natural that in the era of linguistic studies based on computerized text resources, multilingual corpora, be they parallel (translated) or comparable ones, started to play a significant role. Among the first annotated multilingual corpora there was the parallel English-Czech corpus PCEDT (Hajič et al. 2011; 2012), which is a mostly manually annotated parallel corpus of English and Czech texts with almost 50 000 sentences for each part. The English part contains the Wall Street Journal section of Penn Treebank, along with the original phrase-structure analysis and a newly added dependency-based deep structure syntactic analysis (tectogrammatics). The Czech part are manual translations of the original texts, along with their surface and deep syntactic analyses, automatically parsed and manually checked. Part of the PCEDT (3857 sentences) is annotated also with regard to the basic aspects of the information structure (topic-focus articulation, TFA, in our terms; for the TFA annotation, see Mírovský et al. 2013). With its theoretically well-founded multi-layer and multi-lingual annotation, the PCEDT has served for a formulation of numerous research questions which have led us to a series of experiments carried out on the PCEDT data. All these questions are related to the domain of the topic-focus articulation (TFA) of the sentence and the theoretical framework we work with is the TFA approach as developed in Functional Generative Description proposed by Petr Sgall and elaborated further by his students and followers (Sgall 1967; Sgall et al. 1973, 1980, 1986). In the present contribution we devote our attention to the following general issues:

(i) the „universality“ of information structure,

(ii) the relation between the structure of the sentence and the coherence of text,

(iii) the status of information structure with regard to semantics and pragmatics.

* 1. The „universality“ of information structure

Information structure is considered to be a universal phenomenon, related to the communicative function of the sentences, though the means of realization of this structure may differ from language to language, ranging from word order and prosody to specific morphemes or syntactic structures. It has also been demonstrated that information structure is semantically relevant: sentences that differ in their TFA may achieve a different meaning or may even differ in their truth conditions. A natural assumption then follows that translation from one language to another should preserve the TFA of the original sentence.

To test this hypothesis on the available PCEDT data, we have first looked (Hajičová, Mírovský, Rysová and Rysová, 2019) at the original English (source) sentences with regard to their Czech translations (target) from the point of view of the Focus proper of the sentence, i.e. the element with the highest degree of communicative dynamism, in terms of Firbas (1992). The total number of automatically aligned sentences without coordination of the main predicates was 3857; there were 2514 cases (65.3%) with the same underlying syntactic value (functor) at the last position of both source and target text and 1287 cases where there was a difference in the functor. We have searched for two situations, (i) the Focus proper assignment, and (ii) the presence of the item that is in Focus proper in the source sentence (English) in at least the global Focus in the target (Czech) sentence. For our analysis under (i), we have randomly chosen 120 cases of the set of 1287 sentences and put them under a detailed inspection. To answer the research question (ii), we have put under scrutiny all 171 cases where the element appearing as Focus proper in the source language (English) was not included even in the global Focus of the target (Czech) sentence. After manual filtering of cases where the difference in Focus was due to a difference in the syntactic structure while the target lexical item corresponded to the source one, we have found out that most frequently the difference concerned the mutual position of the main predicate and its modification of time or place (in English, this modification often fulfilled the role of Focus proper) or the position of the predicate itself (in Czech, more frequent position of the predicate was also the Focus proper). As for the “weaker” situation, namely cases when the Focus proper in the source sentence was at least included in the whole global Focus part of the target sentence, it has come out that again, most differences concerned the syntactic relation of temporal or local modification which in the Czech target sentence appeared in the Topic rather than in the global Focus.

These findings have led us to look more systematically at the position of the temporal (TWHEN) and local (LOC) modifications (Hajičová, Mírovský, Rysová, K.., Rysová, M. in prep.). In particular, we have followed the cases where the two languages studied differ in the placement of the modifications TWHEN or LOC in the Topic in one language and in the Focus part of the same sentence in the other. In order to get a richer sample of examples, we have searched in the whole of PCEDT and we have approximated the division into Topic and Focus by the position of these modifications before (Topic) and after (Focus) the main verb (PRED). We had at our disposal the samples in Table 2.

|  |  |  |
| --- | --- | --- |
|  | TWHEN | LOC |
| Before PRED in E., after PRED in Cz. | 233 | 67 |
| After PRED in E., before PRED in Cz. | 765 | 271 |
| Total | 998 | 338 |

Table 2 The position of TWHEN and LOC with respect to the Predicate in English compared to Czech

In order to analyze the position of TWHEN, we have randomly chosen a sample of 100 English sentences with their Czech counterparts from each of the sets (out of 233 and 765 examples, respectively). The following observations seem to hold:

(i) With TWHEN in the post-verbal position, this modification was typically expressed in English by a short adverb (-ly adverb, yesterday, …) and was placed next to the Predicate, In such a case, this post-verbal element may be considered to be a part of Topic also in English (1), or, alternatively, the preverbal position of such an adverb in English may be considered part of the Focus (2).The same holds about cases of TWHEN expressed in English by a short adverb and placed at the end of the sentence, but (presumably) this adverb does not carry the intonation centre (3). There are also apparent cases of differences, due to the post-verbal placement of the TWHEN modification in Czech (in contrast to the source English sentence). There is a tendency in Czech to place the Predicate into the second position of the sentence, which has led to the placement of the TWHEN modification in the Czech sentence after the verb also in case in which it was an indisputable element of the Topic of the sentence (4). All these examples, if analyzed properly with regard to Topic and Focus rather than with regard to its pre- or post-verbal position, do not represent instances of differences between English and Czech in the Topic vs. Focus position.

(1) E.: *In national over-the-counter trading, the company closed yesterday at $23.25 a share.*

 Cz.: *Při celostátním mimoburzovním obchodování společnost včera uzavřela na 23.25.*

(2) E.: *The utility company currently has about 82.1 million shares outstanding.*

 Cz.: *Tento podnik veřejných služeb má v současné době v oběhu 82.1 milionu akcií.*

(3) E.: *Democrats had been negotiating with some Republican congressional leaders on a compromise lately.*

 Cz.: *V poslední době vyjednávali demokraté s některými čelními republikánskými představiteli Kongresu o kompromisu.*

(4) E.: *A year earlier, Nationwide Health earned.*PRED *$2.4 million, or 29 cents a share.*

 Cz.: *Výnosy společnosti Nationwide Health činily.*PRED *v loňském roce 2.4 milionu dolarů, neboli 29 centů na akcii.*

(ii) In English, the position of TWHEN at the end of the sentence (i.e. in the prototypical position of Focus) maybe due to the weight of the element, be it a prepositional phrase or a whole dependent clause (5).

(5) E.: *The shares traded at about A$ 1.50 in March, when the plan to acquire MGM\/UA was announced.*

 Cz.: *V březnu, kdy byl plán na převzetí společnosti MGM/UA oznámen, se akcie obchodovaly kolem 1,50 australského dolaru.*

(iii) Nevertheless, there was a considerable number of ”true“ examples where the English sentence differed from its Czech counterpart in the placement of the TWHEN modification in the Topic vs. the Focus part (6), (7):

(6) E.: *But we're ... going to be in the exact same situation next year.*

 Cz.: *Ale příští rok budeme... v naprosto stejné situaci.*

(7) E.: *Only twice since the 1960s has annual gross domestic product growth here fallen below 5% for two or more consecutive years.*

 Cz.: *Roční nárůst hrubého domácího produktu zde spadl pod 5 % během dvou nebo více po sobě jdoucích let pouze dvakrát od šedesátých let.*

For some of these cases, as (6), the initial position of TWHEN in Czech may be interpreted as a contrastive Topic: it is still (a part of) Topic, the sentence being „about“ it, but the contrastive character of this element makes it comparable with Focus (which, as a choice of alternatives, always has a contrastive character).

For the analysis of the placement of LOC with regard to the Topic vs. Focus position, we have again randomly chosen 100 sentences from the set of 271 English sentences with LOC after PRED and we have analyzed all the 67 pairs of English-Czech sentences in the set of LOC before PRED. The following observations, in several points similar to those listed above for TWHEN, seem to hold:

 (i) The LOC modification may be considered as a part of Topic or alternatively as a part of Focus in case the position of the LOC is close to the predicate (8) or the LOC in the final position need not be a carrier of the intonation centre (9) the prosodic factor being a decisive factor for the identification of Focus. A tendency was also observed to place the Predicate into the second position of the Czech sentence which has led to the post-verbal placement of the LOC modification also in case in which it was an indisputable element of the Topic of the sentence, see (10).

(10) E.: *In an interview, Pemberton Hutchinson, president and chief executive, cited several reasons for the improvement: higher employee productivity and ”good natural conditions'' in the mines, as well as lower costs for materials, administrative overhead and debt interest.*

Cz.: *Prezident a výkonný ředitel Pemberton Hutchinson jmenoval.PRED v rozhovoru několik důvodů zlepšení: vyšší produktivitu zaměstnanců a* „*dobré přírodní podmínky" v dolech, stejně jako nižší cenu materiálu, administrativní režii a úroky z úvěrů*

(ii) A modification is placed at the end of the sentence in English because of its weight rather than because of its appurtenance to Focus proper (11):

(11) E.: *The citation was misstated in Friday's edition.*

 Cz.: *V pátečním vydání byla tato citace uvedena chybně.*

(iii) However, also with the LOC modification there was a considerable number of “true” examples where the original English sentence differed from its Czech equivalent in the placement of the LOC modification in the Topic vs. in the Focus part (11).

(11) E.: *The citation was misstated in Friday's edition.*

 Cz.: *V pátečním vydání byla tato citace uvedena chybně.*

It is often the case that the preceding context helps to identify the Focus, but not necessarily so, as the following example demonstrates (12):

(12) E.: *The year was misstated in Friday's edition.*

 Cz.: *V pátečním vydání byl rok uveden chybně.*

 E. previous context**:** *QUANTUM CHEMICAL Corp.'s plant in Morris, Ill., is expected to resume production in early 1990.*

* 1. The relation between the structure of the sentence and the coherence of text

The interest in the study of the inter-relationships between a composition of the sentence and text-coherence dates back to the psychologically rather than linguistically oriented considerations by Weil (1844, quoted here from the 1978 E. translation), who recognized two types of the „movement of ideas“, namely marche parallèle and progression: „If the initial notion is related to the united notion of the preceding sentence, the march of the two sentences is to some extent parallel; if it is related to the goal of the sentence which precedes, there is a progression in the march of the discourse“ (p. 41). In addition to these basic strategies, Weil recognized also a reverse order called by him ‚pathetic‘, when the speaker under deep emotions „enters into the matter of his discourse at the goal.” (p. 45). In Czech linguistics, Mathesius (1947) with his study on the thematicity and “continuity” of English subject (further analyzed esp. by Dušková 2008; 2010) and Daneš‘ notion of thematic progressions (Daneš 1970; 1974) follow a similar line of thinking.

Having at our disposal the PDT corpus annotated on a deep layer of syntactic structure including the information structure of the sentence and, simultaneously, i.e., on the same documents, and incorporating some basic features of discourse structure, incl. the coreferential relations that go beyond the sentence structure (Hajič et al., 2018), we have decided to test the local coherence properties of Czech as established by links between the thematic (Topic) and rhematic (Focus) parts of sentences in different genres of discourse (Hajičová and Mírovský 2018a). In particular, we wanted to verify if the classical observations valid for English as a language with a grammatically fixed word order, namely that there is a prevalence of “constant” theme, are also valid for a typologically different language, such as Czech, in which the word order is not guided by grammatical rules.

For this purpose, we used the data from the PDT to follow four possible “thematic” relations between neighbouring sentences: (i) (some element of the) Topic of the sentence *n* refers to (some element of the) Topic of the sentence *n-1* (denoted below as Tn-1 – Tn); (ii) (some element of the) Topic of the sentence *n* refers to (some element of the) Focus of the sentence *n-1* (denoted below as Fn-1 – Tn); (iii) (some element of the) Focus of the sentence *n* refers to (some element of the) Focus of the sentence *n-1* (denoted below as Fn-1 – Fn); (iv) (some element of the) Focus of the sentence *n* refers to (some element of the) Topic of the sentence *n-1* (denoted below as Tn-1 – Fn-1). “An element x refers to an element y” means that there is an anaphoric link (be it a proper coreference or a bridging relation) between the referring expressions x and y in adjacent sentences.

After a preliminary test of the methodology applied to 100 sentences randomly chosen from 5 different genres, we have put under scrutiny a larger amount of data from the essay genre and we have complemented our identification of the “global” Topic and Focus by a more detailed analysis of the inner structure of these parts as for the value of contextual boundness within the TFA attribute. We have also paid a more detailed attention to the type of anaphoric relations, to see whether the difference between (pure) coreference and bridging plays some important role.

The smaller sample contained 100 annotated sentences from the genre of essay, 79 of which were linked by links of coreference or bridging relations. In this sample, the Fn-1 – Tn sequences prevailed only slightly (28 cases) followed by the Fn-1 – Fn type (24 cases), the Tn-1 – Tn type (11 cases) and the Tn-1 – Fn type (8). This is to say that the ratio between what we have considered to be typical relations (from Topic in the second sentence of the pair) and the non-typical relations (from the Focus of the second pair) was almost balanced (39 versus 32). Under a more detailed analysis of the 24 cases of the Fn-1 – Fn type relations, it has been confirmed that in most cases, the anaphoric link leads from a contextually bound element of Fn which may serve as a support to distinguish local topics and local foci with the overall Topic and Focus. This finding is in an agreement with the assumption (made explicit in Hajičová, Partee and Sgall 1998) of the theory of TFA we subscribe to that the recursive character of this articulation makes it possible (or even necessary) to distinguish between the “overall” bipartition of the sentence into its Topic and Focus and the local partitioning within these two parts into what may be called “local Topic” and “local Focus”. Another explanation of the unexpected links between elements of the Foci of the adjacent sentences is the fact that 12 out of the 24 Fn-1 – Fn links were bridging relations in which the mentioning in the second sentence has a contrastive character (i.e. the contrast between a whole and a part of the whole, set or subset) or is accompanied by a particle (such as only) with a focusing function which by itself is contrastive.

To confirm our observations we have carried out a third step, in which we have applied the analysis onto the whole subset of the essay genre in the PDT corpus; this sample contains 189 documents with the total of 6858 sentences, among which 4606 adjacent sentences contained a pairwise anaphoric link. The figures obtained confirm even more clearly the picture described above: the Fn-1 – Tn sequences prevailed considerably (1771 cases) over the Tn-1 – Tn type (1278 cases) while the number of the non-prototypical links was much lower (the Fn-1 – Fn type with 1004 cases and the Tn-1 – Fn type with 553 cases).

As the PDT corpus offers the possibility to examine relations present at different layers of annotation rather than to restrict the attention to a single level, we also used our material to look at possible relationships between the anaphoric links and the surface syntactic function of the elements in question. It has come out that contrary to the situation in English, for which the relevant linguistic studies suggest a prevalence of a “constant theme”, our material documents that this is not the case in Czech: in the second sample of 100 sentences, among the Tn-1 – Tn type (i.e. the type relevant for the issue under investigation) , there was no instance of a subject-to-subject link.

Our analysis based on the PDT multi-layered annotation of Czech sentences has demonstrated that among the four possible types of the relations between anaphoric links and the Topic-Focus bipartition of the sentence, the most frequently occurring type is a link between the Topic of the sentence to the Focus of the previous sentence. In case there is an anaphoric link leading from a sentence to the Focus of the next following sentence, this link frequently leads to a contextually bound element of the Focus of the next sentence, which supports the assumption that it is convenient to distinguish between “overall” Topic and Focus and the local Topic and Focus; and/or the anaphoric relation is of the type of bridging, which is often interpreted as a contrast.

* 1. The status of information structure with regard to semantics and pragmatics

The information structure of the sentence can be viewed upon from two standpoints: either as one of the aspects of the (underlying, semantic) structure of the sentence, or, alternatively, as a phenomenon belonging (at least mainly) to the domain of pragmatics. The TFA theory we subscribe to adheres to the former stream claiming that information structure, TFA in our terms, is semantically relevant, and as such its description belongs to the level of linguistic meaning (see Sgall et al. 1986; Hajičová et al. 1998). This standpoint can be illustrated by examples in (13).

(13) Mary called Jim a Republican.

(a) Then he insulted HER.

(b) Then he INSULTED her.

In both (a) and (b). Jim and Mary are (cognitively) ‘known’ since they are referred to in the preceding sentence, but only (b) is linguistically structured as having both of them in its Topic (theme) and the information in the Focus (rheme) is the event of insulting. In (a), Mary (though “known” from the previous sentence) is placed in the Focus. This interpretation is supported by the different intonation patterns indicated by the capitals.

To support this position, we have put under scrutiny (Hajičová and Mírovský 2018b) one of the basic assumptions of the pragmatically oriented studies, namely that the dichotomy of Topic and Focus (theme/rheme, or whatever terms may be used) is based on a cognitively based dichotomy between given and new information.

We start from the hypothesis that a ‘given’ item in a sentence must be somehow anchored in the previous sentence(s) while a ‘new’ item lacks such an anchoring, i.e. we study the Topic/Focus dichotomy vis-a-vis coreference relations (cf. Prince 1981; Kuno 1972). The data we have used, namely the PDT 3.5 (Hajič et al. 2018) has allowed us to follow both the absence and the presence of anaphoric links leading from some element of the Focus of the sentence to the preceding context. In case of an absence of such a link, we may conclude that the Focus of the sentence is ‘new’; a presence of such a link indicates the presence of a ‘given’ element in the Focus.

We have put under scrutiny documents of 10 different genres containing more than 20 sentences each. In 7286 cases (37%), there was no anaphoric link leading from the Focus of the given sentence to some element of the previous sentence(s); there were 12159 cases (63%) where there was a link from an element in the Focus of the sentence to some element occurring in the previous sentence(s), out of which in 6453 cases (53% of 12159) the link led to the Focus and in 5706 cases (47%) the link led to the Topic. These findings have allowed us to conclude that in most cases (63%), the Focus part of the sentence does not represent ‘new’ information. In addition, when we have taken into account also the type of the anaphoric relation, in case the coreferential link led to the Focus of the preceding sentence, the anaphoric relation leading from the Focus of a given is mostly of the type of bridging and is often interpreted as a contrast.

1. Summary

The aim of the present contribution was twofold: first, to demonstrate that in the era of computerized corpora availability, an “armchair linguist” and a “corpus linguist” wherever possible, should exist in the same body while the former may find “a profit in being a consumer of some of the resources that corpus linguists have created“ (Fillmore 1992, p. 35), and second, that the work in this direction may bring an advancement both in natural language processing and in fundamental linguistic research in the field of computational linguistics if the hitherto available corpora are used in collaboration rather than separately. It is our belief that for this second aim the close cooperation between Brno and Prague as successfully initiated by Karel Pala will bring lot of useful fruit also in the future.

Acknowledgements

The work described herein has been supported originally by a large number of various grants and projects by several funding agencies; the most recent support comes from the Ministry of Education, Youth and Sports under the Research Infrastructures program (projects LM2015071 and LM2018101) and the Grant Agency of the Czech Republic (project GA17-07313S).

References

Artstein, R. and M. Poesio: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4), pp. 555-596. (2008)

Atkins, S.: Tools for computer-aided corpus lexicography: the Hector project. In: *Papers in Computational Lexicography*: Complex’93, Budapest. pp. 1-60. Baker, C. F., Fillmore, C. J. and J. B. Lowe (1998): The Berkeley FrameNet project. In: *Proceedings of the COLING-ACL*, pp. 86-90. Montreal, Canada. (1993)

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and N. Schneider: Abstract Meaning Representation for Sembanking. In: *Proceedings of the 7th LAW Workshop*, pp. 178-186. ACL, Sophia. (2013)

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M. and A. Tamchyna: The Joy of Parallelism with CzEng 1.0. In: *Proceedings of the 8th Intl. Conference on Language Resources and Evaluation* (LREC 2012), ELRA, pp. 3921-3928. Istanbul, Turkey (2012)

Cinková, S.: From PropBank to EngValLex: Adapting the PropBank-Lexicon to the Valency Theory of the Functional Generative Description. In: *Proceedings LREC 2006*, pp. 2170-2175. Genova, Italy (2006)

Čermák, F. and A. Rosen: The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics.* Volume 17, No. 3, pp. 411-427, John Benjamins (2012)

Daneš, F.: Zur linguistischen Analyse der Textstruktur. *Folia linguistica*, 4, p. 72-78 (1970).

Daneš, F.: Functional Sentence Perspective and the organization of the text. In: Daneš, Ed. *Papers on Functional Sentence Perspective*. pp. 106-128. Prague, Academia (1974)

Dušková L.: Theme movement in academic discourse. In: M. Procházka and J. Čermák, Eds., *Shakespeare between the Middle Ages and Modernity. From translators art to academic discourse.* pp. 221-247. Prague, FF UK (2008)

Dušková L.: Rozvíjení tématu v akademickíém a narativním textu [The developmen of theme in an academic and narrative text]. In Čmejrková S., Hoffmannová J., Havlová E.: *Užívání a prožívání jazyka*. K 90. narozeninám Františka Daneše. pp. 253-260. Praha, Karolinum (2010)

Fellbaum, C.: *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA (1998)

Fillmore, C. J.: “Corpus linguistics“ or “Computer-aided armchair linguistics“, In*: Directions in Corpus Linguistics*, Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991, ed. Jan Svartvik, pp. 61–77. Mouton De Gruyter, Berlin New York (1992)

Firbas, J.: *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge: Cambridge University Press (1992)

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z. and Z. Žabokrtský: *Prague Czech-English Dependency Treebank 2.0.* Data, UFAL MFF UK, Prague, Czech Republic, http://ufal.mff.cuni.cz/pcedt2.0, <http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4> (2011)

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z. and Z. Žabokrtský: Announcing Prague Czech-English Dependency Treebank 2.0. In: *Proceedings of the LREC 2012*, pp. 3153–3160 (2012)

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š. and Z. Žabokrtský: Prague Dependency Treebank 3.5. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN PID: <http://hdl.handle.net/11234/1-2621> (2018)

Hajičová, E. and J. Mírovský: Discourse Coherence through the Lens of an Annotated Text Corpus: A Case Study. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC’18), Miyazaki, Japan. European Language Resources Association (ELRA) (2018a)

Hajičová, E. and J. Mírovský: Topic-Focus vs. Given-New: Information Structure and Coreference Relations in an Annotated Corpus. In: *Proceedings of SLE 2018*, Tallin (2018b)

Hajičová, E., Mírovský, J., Rysová, K. and M. Rysová: Information Structure in an Annotated Parallel English-Czech Corpus. In: *Proceedings of the SLE 2019 Conference*, Leipzig (2019)

Hajičová, E., Mírovský, J., Rysová, K. and M. Rysová: Ordering of Adverbials of Time and Place in Grammars and in an Annotated English-Czech Parallel Corpus. (in prep.)

Hajičová E., Partee, B. H. and P. Sgall: *Topic-Focus Articulation, Tripartite Structures and Semantic Content*. Dordrecht-Boston: Kluwer Academic Publishers (1998)

Kettnerová, V.: *Lexikálně-sémantické konverze ve valenčním slovníku.* Czech Republic, 280 pp. Karolinum, Prague (2014)

Kettnerová, V., Lopatková, M. and E. Bejček: The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In: *Proceedings of the 15th EURALEX International Congress*, Department of Linguistics and Scandinavian Studies, ISBN 978-82-303-2228-4, pp. 434-443. University of Oslo, Oslo, Norway (2012)

Kilgarriff, A.: I don’t believe in word senses. *Computers and the Humanities*, 31(2):91-113, (1997)

Kuno, S.: Functional sentence perspective. *Linguistic Inquiry* 3, pp. 269-320 (1972)

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Z. Žabokrtský: *Valenční slovník českých sloves VALLEX.* Nakladatelství Karolinum, Praha, Czechia, ISBN 978-80-246-3542-2, 698 pp. (2016)

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., Žabokrtský, Z. and P. Barančíková: VALLEX 3.5 - *Valenční slovník českých sloves*. <http://ufal.mff.cuni.cz/vallex/3.5/> Charles University, Prague (2018)

Marcus, M., Santorini, B. and M. A. Marcinkiewicz: "Building A Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*. 19(2), pp. 313-330 (1993)

Mathesius V.: *Čeština a obecný jazykozpyt* [Czech and General Linguistics], Praha: Melantrich (1947)

Mírovský, J., Rysová, K., Rysová, M. and E. Hajičová: (Pre-)Annotation of Topic-Focus Articulation in Prague Czech-English Dependency Treebank. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing*, ISBN 978-4-9907348-0-0, pp. 55–63. Nagoya, Japan (2013)

Pala, K. and J. Všianský: *Slovník českých synonym*. Lidové Noviny (1994)

Palmer, M.: Semlink: Linking PropBank, VerbNet and FrameNet. In: *Proceedings of the Generative Lexicon Conference*. GenLex-09. Sept. 2009, Pisa, Italy (2009)

Peterson, D., Boyd-Graber, J., Palmer, M. and D. Kawahara: Leveraging VerbNet to build Corpus-Specific Verb Clusters, In: *Proceedings of \*SEM 2016*, ACL 2016, May, Berlin, Germany (2016)

Prince, E.: Toward a taxonomy of given/new information. In Cole, ed. *Radical Pragmatics*, pp. 223–254(1981)

Sgall P.: Functional Sentence Perspective in a Generative Description. In: *Prague Studies in Mathematical Linguistics 2*, 203–225. Prague, Academia (1967)

Sgall, P., Hajičová, E. and E. Benešová: *Topic, Focus and Generative Semantics.* Kronberg/Taunus: Scriptor (1973)

Sgall, P., Hajičová, E. and E. Buráňová: *Aktuální členění věty v češtině*. [Topic-Focus Articulation of the Czech Sentence]. Prague: Academia (1980)

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

Urešová, Z.: *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics 9, 375 pp. Charles University (2011)

Urešová, Z., Fučíková, E. and J. Šindlerová: CzEngVallex: a bilingual Czech-English valency lexicon. In: *The Prague Bulletin of Mathematical Linguistics*, 105, ISSN 0032-6585, pp. 17-50 (2016)

Urešová, Z., Fučíková, E. Hajičová, E. and J. Hajič: Creating a Verb Synonym Lexicon Based on a Parallel Corpus. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC’18), European Language Resources Association (ELRA). Miyazaki, Japan (2018).

Weil, H.: *De l’order des mots dans les langues anciennes comparées aux langues modernes*, Paris: Joubert (1844). Translated by Charles W. Super as *The order of words in the ancient languages compared with that of the modern languages,* Boston: Ginn, 1887, reedited and published by John Benjamins, Amsterdam 1978