PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 87 JUNE 2007 61–70

Identification of Topic and Focus in Czech

Evaluation of Manual Parallel Annotations

Šárka Zikánová, Miroslav Týnovský, Jiří Havelka

Abstract

This paper presents results of a control annotation of the Topic-Focus Articulation of Czech sentences based on the notion of "aboutness". This is one of the steps testing the hypothesis about the relation between contextual boundness and "aboutness". We suppose that the bipartition of the sentence into its Topic and Focus ("aboutness") can be automatically derived from the values of contextual boundness assigned to each node of the dependency tree representing the underlying structure of the sentence. For the testing of this hypothesis, control manual parallel annotations have been carried out. The principles of the controll annotations are described and preliminary results are reported on.

1. Introduction

The topic-focus articulation of a sentence into its Topic and Focus can be looked upon from two points of view: as derived from a primary notion of contextual boundness or in terms of *"aboutness"* (Focus is "about" Topic, i.e. F(T); with a primary reading of a negative sentence non-F(T); see Mathesius, 1947, p. 235; Sgall, Hajičová, and Panevová, 1986; Firbas, 1992). According to *contextual boundness,* elements in sentences are classified as contextually bound (CB; with a subtype of contrastive contextually bound elements, CCB) or contextually non-bound (CN; Sgall, Hajičová and, Buráňová, 1980; Sgall, Hajičová, and Panevová, 1986), cf. the following example:

(1) *(Maruška se obrátila na lesní víly.) Víly*$_{CB}$ *ji*$_{CB}$ *vyslechly* $_{CN}$.
[lit.: Mary turned to forest fairies.] The_fairies$_{CB}$ received$_{CN}$ her$_{CB}$.

The bipartition of the sentence into Topic and Focus is then derived from the CB/CCB/NB features; for our example, the Topic is *Víly*$_{CB}$ *ji*$_{CB}$ [(The) fairies$_{CB}$ her$_{CB}$] and the Focus is *vyslechly*$_{CN}$ [received$_{CN}$].

The criterion of the "aboutness" divides a sentence into two parts: Topic (T, a part expressing what the sentence is about) and Focus (F, a part giving information about the Topic). Thus, if (2)

Please cite this article as: Šárka Zikánová, Miroslav Týnovský, Jiří Havelka, Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations. The Prague Bulletin of Mathematical Linguistics No. 87, 2007, 61–70.

is used in the context of the question "Where are the children? / What are the children doing?", the following assignment of Topic and Focus would hold: the Topic of the sentence is *Děti* [The children] and the Focus is *běhají po ulici* [are running in the street].

(2) *Děti* $_{T}$ *běhají* $_{F}$ *po ulici* $_{F}$.
[lit.: The_children $_{T}$ are running $_{F}$ in the_street $_{F}$.]

It should be noted (cf. Sgall, Hajičová and, Panevová, 1986), that although in principle the CB items belong to the Topic of the sentence and the NB items to the Focus, this is not so when deeply embedded sentence elements are taken into account. See the element *vašeho* [your] in (3) which belongs to the Focus, though it is contextually bound. (The context of the sentence can be "What did you do yesterday?".)

(3) *Včera*$_{CB,\,T}$ *jsem potkal*$_{CN,\,F}$ ***vašeho***$_{\mathbf{CB,\,F}}$ *kolegu*$_{CN,\,F}$.
[lit.: Yesterday$_{CB,\,T}$ I_met$_{CN,\,F}$ **your**$_{\mathbf{CB,\,F}}$ colleague$_{CN,\,F}$.]

2. The framework of the project: from contextual boundness to aboutness

In our project the relations between contextual boundness and aboutness are investigated. According to the underlying hypothesis, the values of aboutness (Topic and Focus) can be derived from the values of contextual boundness Sgall, Hajičová and Panevová (1986). We test this hypothesis on the material from the Prague Dependency Treebank 2.0 (PDT), where approximately 50,000 Czech sentences have been annotated on three levels, one of them being the underlying syntactic level (tectogrammatics). On that level, sentences are represented in a form of dependency trees, in which the nodes represent autosemantic elements of the sentence and the edges represent the types of relations between the governing and the dependent nodes. Every node has been assigned (in addition to other relevant values) one of the values of contextual boundness.

Our study proceeds in the following three steps:

- the formulation of an algoritm transforming the values of contextual boundness into the values of aboutness; implementation of the algoritm on the data from the PDT (Sgall and, Hajičová, 2005; Hajičová, Havelka and, Veselá, 2005);
- manual parallel annotation of the control data according to the aboutness relation (i.e. directly assigning the bipartition of Topic and Focus);
- comparison of the values achieved in the manual annotation with the automatically assigned T-F bipartition and evaluation of the results.

In the present paper, we are concerned with the second point of the overall programme of the project – we describe and evaluate the results of the manual parallel annotations which will later serve as referential data for the evaluation of the automatic recognition of Topic and Focus.

3. The linguistic material

For the control annotation, the texts from the PDT have been used, so that we get data comparable with the results of the automatic procedure. The texts in the PDT come from Czech

newspapers from the beginning of the 1990's; they extend from short remarks to longer essays. The annotators worked with whole texts, since for the correct analysis of the topic-focus articulation, it is necessary to respect the context.

In total, almost 11,000 sentences have been analysed (cf. Tables 1–2). All the annotations have been done in parallel, in order to take into account possible disagreement of the annotators in their interpretation of the topic-focus articulation as well as to be aware of errors by individual annotators. The main part of sentences (almost 10,000 sentences) has been analysed in three parallel annotations; a smaller sample of almost 900 sentences has been annotated in six parallel versions.

4. The method of the annotation

In order not to "spoil" the control data with the hypothesis to be verified, we worked with ten annotators who were not familiar with the previous annotation of the topic-focus articulation in the PDT. If we wanted to get the picture of the common perception of the topic-focus articulation by native speakers, we could not influence annotators with too strict instructions which could be contradictory to their natural intuition. Therefore basic principles of the annotation have been outlined only; later some problematic parts have been discussed in detail (e.g. the analysis of questions, sentences consisting just of one word or sentences with direct speech; cf. Zikánová, 2006).

The annotators worked with a linear (surface shape) form of the texts. There were four possible values, which could be ascribed to individual words in texts:

T	part of the Topic	(what the sentence is about)
F	part of the Focus	(new information about the Topic)
B	Boundary	(a marker of the boundary dividing two structures in which Topics and Focuses should be identified separately, e.g. a conjuction or a punctuation mark within a sentence)
N	Not clear	(problematic words where the annotator is not sure)

The elementary instructions for the annotation included the following points:

- Analyse the structure of the main clause only. Dependent clauses are to be treated as integral elements of the main clause. (Therefore the borderline between the Topic and Focus should not be marked within a dependent clause.)
- Describe the appurtenance of every single word or unit in the main clause to Topic or Focus. It is possible that there is more than one border between these two parts of the sentence, both of these parts can be interrupted with other elements.
- In coordinated clauses, analyse the structure of each main clause separately. (In complex sentences with subordinated clauses, analyse the main clause only.)
- Describe the nominal group as an integrated element (with a preposition, pronoun, adjective or another noun, as the case may be).
- It is possible to assign all the elements of a sentence as belonging to the Focus. (It is not necessary that the sentence contains Topic.)

Generally, when choosing which elements in the linear surface shape of a sentence might have been in the analysis omitted, we have been guided by the principles of the automatic annotation of underlying dependency trees in the PDT as we want to compare these sets of data. The following example presents the way of analyzing sentences in control annotations:

(4) (In the previous context, poor conditions in different world trading zones have been mentioned in general.)
V Indonésii je minimální denní mzda jeden a půl dolaru a někdy za to musí dělníci pracovat 10–12 hod.
[lit.: In Indonesia is minimal daily pay one and half dolar and sometimes for it have workers to_work 10–12 hours.]
[In Indonesia, the minimal daily pay is one and half dolar and sometimes workers have to work for 10–12 hours for it.]

1. There are two coordinated clauses in the sentence; they are to be analysed separately, the conjuction *a* is assigned the value B (Boundary).

2. When setting Topic and Focus, we formulate first a question about the presupposed Topic of the sentence. As for the first clause, we can ask the following questions: *What can we say about Indonesia? What can we say about the minimal daily pay in Indonesia? What can we say about one and half dolar?*

3. Then the analysed sentence is tested as an answer to the formulated question:

(4a) *What can we say about Indonesia? – V Indonésii je minimální denní mzda jeden a půl dolaru.*
In Indonesia, the minimal daily pay is one and half dolar.

(4b) *What can we say about the minimal daily pay in Indonesia? – V Indonésii je minimální denní mzda jeden a půl dolaru.*

(4c) *What can we say about one and half dolar? – V Indonésii je minimální denní mzda jeden a půl dolaru.*

If the answer naturally matches with the question (with respect to the previous context), then the elements repeated from the question are assigned the value T (Topic) and the elements of the part that is the proper answer are assigned the value F (Focus). If the answer does not correspond to the question, the choice of the presupposed Topic is not correct.

In our case, questions (a) and (b) can be answered with the analysed sentence, whereas the question (c) does not correspond to it. Thus, the Topic-Focus values will be assigned in the following way:

(4a') *V Indonésii* $_T$ *je* $_F$ *minimální denní mzda* $_F$ *jeden a půl dolaru* $_F$.

(4b') *V Indonésii* $_T$ *je* $_F$ *minimální denní mzda* $_T$ *jeden a půl dolaru* $_F$.

Since there is a (restricted) variability in matching questions, a certain variability in answers and analyses is admissible, too (4a'–b').

The second clause of the compound sentence is analysed according to the same instructions. The appropriate question to which this clause can be an answer is *What can we say about this daily pay?*

(4d) *Někdy za to musí dělníci pracovat 10–12 hod.*
[lit.: Sometimes workers have to_work for 10–12 hours for it.]

Někdy $_F$ *za to* $_T$ *musí* $_F$ *dělníci* $_F$ *pracovat* $_F$ *10–12 hod.* $_F$.

5. Results and discussion

When evaluating the parallel annotations, we have restricted our attention to certain types of the phenomena observed. With the following elements the assigned value of aboutness has not been taken into account:

- all the words of subordinated clauses except for the verb governing the subordinated (dependent) clause,
- all auxiliary words, which have no corresponding node on the tectogrammatical level of the PDT (functional words such as verbal morphemes, prepositions),
- punctuation marks.

Examining the results, we work with the T/F values of aboutness which have been described above in Sect. 4 and with an additional value "U" – "unannotated" for very sporadic occurrences of words overlooked by mistake by the annotators.

Tables 1 and 2 show the level of agreement among three and six parallel annotations, respectively.

Table 1. Agreement among three parallel annotations

	Occurrence	**Percentage**
Number of sentences	9,825	100.00
Agreement in the annotation of whole sentence	3,553	36.16
Number of words	79,419	100.00
Agreement in the annotation of individual words	60,137	75.72

Table 2. Agreement among six parallel annotations

	Occurrence	**Percentage**
Number of sentences	879	100.00
Agreement in the annotation of whole sentence	232	26.39
Number of words	6,232	100.00
Agreement in the annotation of individual words	4,212	67.59

In Tables 3 and 4, the level of agreement in annotation of individual words is presented in a more detailed way.

Explanations: T and F in the three-letter and six-letter labels refer to the assignment of a word to Topic, or to Focus, respectively, so that e.g. TTT means that a word was considered to be a part of Topic with all the three annotators, or TTT TFF means that a word was considered

to be a part of Topic by four annotators and as belonging to the Focus by two annotators.

Table 3. Types of annotations of individual words in three parallel analyses

	Occurrence	Percentage
FFF	46,099	58.05
TTT	14,036	17.67
TFF	10,575	13.32
TTF	8,287	10.43
TFN	139	0.18
FFN	139	0.18
TTN	67	0.08
Others	77	0.10
Total	**79,419**	**100.00**

Table 4. Types of annotations of individual words in six parallel analyses

	Occurrence	Percentage
FFF FFF	3,332	53.47
TTT TTT	880	14.12
TFF FFF	635	10.19
TTT TTF	367	5.89
TTT FFF	335	5.38
TTF FFF	332	5.33
TTT TFF	288	4.62
FFF FFN	23	0.37
Others	40	0.64
Total	**6,232**	**100.00**

The results of the three parallel annotations in Table 3 as well as of the six parallel annotations in Table 4 indicate that the highest percentage of agreement has been achieved with words belonging to the Focus. The agreement as for the appurtenance of a word to the Topic is not that frequent, nevertheless it is in both annotations the second most common case of agreement. In both annotations, the first two positions in the Tables are occupied by cases of absolute agreement; altogether there are 75.72 % of the absolute agreement in the annotation of individual words at three parallel annotations (cf. Table 1) and 67.59 % of the absolute agreement at six parallel annotations (cf. Table 2). In Table 3, which presents the results of three parallel annotations, the disagreement of annotators is almost the same if the assignment is T or F (lines 3 and 4); with six parallel annotations there is an apparent preference of the annotators to assign F (line 3) rather than T (line 4); actually, this is in accordance with our comments above on lines 1 and 2.

It is interesting to notice that the annotators did not acknowledge much doubt in the assignment of values, although the instructions they received allowed to do so and the reading of some words is not unambiguous: they get much more often in an open disagreement with each other than using the value N (not clear).

The following examples present some results of the parallel annotations. In sentence (5), all the three annotators fully agree in their analysis:

(5) (There is no previous context, the text starts with this sentence.)
Jihlavská radnice *hodlá rázně řešit problém neplatičů nájemného.*
[lit.: **Jihlavian town_council** wants peremptorily to_solve problem of_ bad_payers of_ hire_costs.]

[The town council of Jihlava is about to solve their problem with bad payers of hire costs peremptorily.]

(The parts of Topic are marked with bold characters, the other parts belong to Focus.)

Another example of the full agreement is presented under (6), where all the six annotators analyze the sentence in the same way:

(6) (In the previous context, Edvard Beneš was mentioned as a theme of a recent TV- discussion.)

Edvard Beneš *byl tématem natolik kontroverzním, že přivedl do varu i nejserióznější historiky.*

[lit.: **Edvard Beneš** was theme in_so_far controversial, that he_upset even the_most_ respectable historians.]

[Edvard Beneš was such a controversial theme, that he upset even the most respectable historians.]

In some cases, there are more options in the choice of the test question, and subsequently the solutions differ with individual annotators. The sentence (7) presents an extreme example of disagreement among three annotators, where all the interpretations respect the basic guidelines of the annotation.

(7a) (There is no previous context, the text starts with this sentence.)

Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru *vynesly bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu, titul Fotbalista měsíce dubna v anketě týdeníku Kicker.*

(This analysis corresponds with the question: *What can we say about the following qualities of a football player?)*

[lit.: **Seven goals within nine matches, immense game dynamism and desire to win, as well as readiness quickly to_clear language barrier** have_brought to_the_former player _of_Slavia Pavel Kuka, now acting in German Kaiserlautern, title Footballer of_the_Month April in inquiry of_the_weekly_magazine Kicker.]

[Seven goals within nine matches, the immense dynamism in game and the desire to win, as well as the readiness to clear the language barrier quickly have brought the title The Footballer of the Month April in the inquiry of the weekly magazine Kicker to the former player of Slavia Club Pavel Kuka (now acting in German Kaiserlautern).]

(7b) *Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru vynesly* ***bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu,*** *titul Fotbalista měsíce dubna v anketě týdeníku Kicker.*

(Question: What can we say about the player Pavel Kuka?)

(7c) *Sedm branek v devíti utkáních, obrovská herní výbušnost a vůle po vítězství, stejně jako ochota rychle překonat jazykovou bariéru vynesly bývalému slávistovi Pavlu Kukovi, nyní ve službách německého Kaiserlauternu,* ***titul Fotbalista měsíce dubna v anketě týdeníku Kicker.***

(Question: What can we say about the title The Footballer of the Month April?)

The control subcorpus manually annotated in the way described in our paper is now being compared with the output of the automatic assignment of Topic and Focus to the same subcorpus of texts (annotated on the tectogrammatical level of the Prague Dependency Treebank). The automatic procedure is based on the hypothesis that the bipartition of the sentence into its Topic and Focus can be derived from the values of contextual boundness, while the manual annotation reflects directly the bipartition.

As the results show, the variability of manual solutions must be taken into account in further steps. We should be aware that while we get only a single, unambiguous result from the automatic annotation, more ways of interpretation could be correct. This will be of a great importance in the phase of the comparison between the automatic and manual annotation and its evaluation – it must be reasonably determinated which type of an agreement between automatic and manual annotation is significant and which is not. Also, to achieve a deeper insight into the issue of the position of the boundary between Topic and Focus, it is necessary for the analysis of the cases of disagreement between annotators to take into account the appurtenance of the relevant items to different word classes and the structure of sentences generally (cf. the ambiguous position of nominal groups with rhematizers, of the predicate verb or adverbials in some Czech sentences; see Zikánová, 2006). The evaluation of the results of this comparison will be a useful test of the hypothesis and will enrich our knowledge of the information structure.

Acknowledgements

The research reported in this contribution has been carried out under the grant project of the Ministry of Education, Youth and Sports (Czech Republic) MSM-0021620838 "Modern methods, systems and structures of informatics".

References

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication.* Cambridge University Press, Cambridge.

Hajič, Jan et al. 2006. *Prague Dependency Treebank 2.0.* Linguistic Data Consortium, Philadelphia.

Hajičová, Eva and Petr Sgall. 2004. Degrees of Contrast and the Topic-Focus Articulation. In: Steube, Anita (Ed.), *Information Structure: Theoretical and Empirical Aspects.* de Gruyter, Berlin – New York, pp. 1–13.

Hajičová, Eva and Petr Sgall. *Corpus Annotation As a Test of a Linguistic Theory.* In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, ELRA, Paris, pp. 879–884.

Hajičová, Eva, Jiří Havelka and Kateřina Veselá. 2005. Corpus Evidence of Contextual Boundness and Focus. In: *Proceedings from The Corpus Linguistics Conference Series,* vol. 1, no. 1, 9 pp. Birmingham, ISSN 1747-9398, 2007.04.02 under http://www.corpus.bham.ac.uk/PCLC/birmingham-tex-def-def.doc

Mathesius, Vilém. 1947. O tak zvaném aktuálním členění větném. In Mathesius, Vilém. *Čeština a obecný jazykozpyt.* Melantrich, Prague, pp. 234–242. [About so called functional sentence perspective.]

Mathesius, Vilém. 1982. Aktuální členění větné a sloh. In Mathesius, Vilém. Řeč a sloh, quotation according to Mathesius, Vilém; Macek, Emanuel and Josef Vachek (Eds.): *Jazyk, kultura a slovesnost.* Odeon, Prague, pp. 124–128. [Functional sentence perspective and text composition.]

Sgall, Petr and Eva Hajičová. 2005. The Position of Information Structure in the Core of Language. In: Carlson, Gregory N. and Francis Jeffry Pelletier (Eds.): *Referency and Quantification: The Partee Effect.* CSLI Publications, Stanford (California), pp. 289–302.

Sgall, Petr, Eva Hajičová and Eva Buráňová. 1980. *Aktuální členění věty v češtině.* Academia, Prague. *[Topic-Focus Articulation of Czech Sentences.]*

Sgall, Petr, Eva Hajičová and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspect.* Reidel Publishing Company, Dordrecht and Academia, Prague.

Sgall, Petr. 2002. Moravská a pražská (malostranská) koncepce aktuálního členění. In: Hladká, Zdeňka and Petr Karlík (Eds.), *Čeština – univerzália a specifika, 4.* Nakladatelství Lidové noviny, Prague, pp. 51–58. [Moravian and Praguian (Lesser Town) conception of Topic-Focus Articulation.]

Zikánová, Šárka. 2006. Problematické syntaktické struktury: k rozborům aktuálního členění v Pražském závislostním korpusu. In: *Proceedings of the conference VII. mezinárodní setkání mladých lingvistů,* 15.–17. 5. 2006. Olomouc. [In print since 2006]. [Problematic syntactic structures: to the studies of the topic-focus articulation in the Prague Dependency Treebank.]