Identification of Thematic Discourse Relations on the Data from an Annotated Corpus of Czech

Eva Hajičová and Jiří Mírovský

Charles University, Prague, Czech Republic
Institute of Formal and Applied Linguistics
[hajicova|mirovsky]@ufal.mff.cuni.cz

Abstract. In the present contribution we analyze the data of the Prague Discourse Treebank 2.0 (PDiT 2.0; M. Rysová et al., 2016) as for the text coherence based on the so-called thematic progressions, that is links between sentences with regard to their topic–focus articulation (information structure). For this purpose, we work with two ingredients of the PDiT annotation, namely (i) the annotation of the anaphoric relations ("proper" coreference and some basic types of bridging) between sentence elements (both at short and at long distance), and (ii) the bipartition of the sentence into Topic (T) and Focus (F) based on the annotation of contextual boundness.

Keywords: thematic progressions, topic–focus articulation, anaphoric relations.

1 Related Work

1.1 Centering Theory

One of the most deeply elaborated and best known theory of discourse (local) coherence is the so called *centering theory* (Grosz, Joshi and Weinstein, 1995) based on the model of the local attentional states of speakers and hearers as proposed by Grosz and Sidner (1986). Each utterance in discourse is considered to contain a *backward looking center* which links it with the preceding utterance and a set of entities called *forward looking centers*; these entities are ranked according to language-specific ranking principles stated in terms of syntactic functions of the referring expressions. The *transitions* from one utterance to the following one are then specified by rules that capture their ordering: the most preferred are '*continue*' and '*retain*' (the backward looking center of a given utterance equals the backward looking center of the preceding utterance) followed by '*smooth shift*'and '*rough shift*' (the backward looking center of a given utterance differs from the backward looking center of the preceding utterance). The intuition which is behind this ranking of transitions is very close to those behind the notion of the low cost effort (Fais 2004, p.120).

Interesting experiments investigating the effects of utterance structure and anaphoric reference on discourse comprehension examined in the context of utterance pairs with parallel constituent structure (e.g., *Josh criticized Paul. Then Marie insulted him*) are reported in Chambers (1998). The results reveal several limitations in centering theory and suggest that a more detailed account of utterance structure is necessary to capture how coreference influences the coherence of discourse.

A corpus-based evaluation of the preferences proposed in centering theory is given by Poesio et al. (2000). The study has reached some interesting results. As for the 'shifts' rule stating that (sequences of) continuations are preferred over (sequences of) retains, which are preferred over (sequences of) shifts, the tests revealed that there are more shifts than retains.

1.2 Thematic Progressions

To our knowledge, the first comprehensive treatment of the dynamic development of discourse, though clad in psychological rather than linguistic considerations, was given by Weil (1844, quoted here from the 1978 E. transl.). Weil recognized two types of the "movement of ideas", *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). He also noticed a possibility of a reverse order called 'pathetic': "When the imagination is vividly impressed, or when the sensibilities of the soul are deeply stirred, the speaker enters into the matter of his discourse at the goal." (p. 45.)

In Czech linguistics, this idea is later reflected in Daneš' notion of *thematic progressions* (Daneš 1970; 1974), explicitly referring to the relation between the theme and the rheme of a sentence and the theme or rheme of the next following sentence (a simple linear thematic progression and a thematic progression with a continuous theme), or to a 'global' theme (derived themes) of the (segment of the) discourse.

2 Corpus Based Study

In our present corpus-based analysis we focus our attention on the issue of local coherence as established by links between the thematic (Topic) and rhematic (Focus) parts of sentences in different genres of discourse. For this purpose, we use the data from the Prague Discourse Treebank 2.0, which offers a good testing bed as it provides – in addition to the dependency underlying (deep) syntactic relations – annotation of (i) contextual boundness from which the Topic–Focus bipartition of the sentence can be derived, and (ii) basic anaphoric relations, incl. some types of bridging. Such an annotation has allowed us to follow the occurrence of the two basic types of thematic progressions mentioned above, namely (i) continuous theme (Topic), i.e. the Topic of the given sentence is anaphorically related to the Topic of the previous sentence, and (ii) the "progressive" rheme (Focus), i.e. the Topic of the given sentence is anaphorically related to the Focus of the previous sentence.

2.1 Small Sample

For the first step, in which we wanted to test whether our research methodology and the corpus material available may lead to some interesting and representative results, we have randomly chosen 6 documents of 5 genres with the total of 150 sentences and applied the (already implemented) algorithm for the division of the sentence into Topic and Focus based on the values of the TFA attribute (with values non-contrastive contextually bound, contrastive contextually bound and contextually non-bound).[1] As a result, we had at our disposal the total of 150 dependency trees with marked (binary) division into Topic and Focus and with the annotation of coreference and basic bridging relations between referring expressions of the adjacent sentences.

On this sample, we have followed four possible "thematic" relations between neighbouring sentences (the boundary between Topic and Focus is indicated in our examples by a slash):[2]

(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 $T_{n-1} \leftarrow T_n$):

Myšlenka stručného ústavního zákona, který by prostě stanovil, že výdaje státního rozpočtu mají být kryty příjmy téhož roku, / se vyskytla v řadě zemí. Nejrozsáhlejší diskuse na toto téma / se odehrála v 80. letech ve Spojených státech.

The idea of a concise constitutional law, which would simply state that the state budget expenditures are to be covered by the same year's income, / has occurred in a number of countries. The most extensive discussion on this issue / took place in the 1980s in the United States.

(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 $F_{n-1} \leftarrow T_n$):

Dnes je každý / pod novinářskou diktaturou. Diktatura jest / nehlučná, ale jest.

Today everybody is / under a journalist dictatorship. Dictatorship is / not noisy, but it is.

(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 $F_{n-1} \leftarrow F_n$):

Barevný terčík / usnadňuje nakládání pošty do kontejnerů. Během přepravy barva / zlepšuje přehled o tom, zda se zásilka nezpožďuje.

The coloured disc / makes easier the loading of the mail into containers. During the transport the colour / makes the information easier whether the article is not delayed.

[1] The Topic-Focus bipartition of the sentence has been carried out automatically based on the primary opposition of contextually bound and non-bound items reflected in the PDiT by a manual assignment of one of three values of the attribute of TFA. The distinction of contextual boundness should not be understood in a straightforward etymological way: an *nb* element may be 'known' in a cognitive sense (from the context or on the basis of background knowledge) but structured as non-bound, 'new', in Focus. The overall accuracy of the algorithm, measured on the assignment of tectogrammatical nodes either to Topic or Focus of the sentence, is 0.93 (Rysová et al., 2015).

[2] The examples in this section are original sentences from the PDiT.

(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 $T_{n-1} \leftarrow F_n$).

Novináři jsou / hlídací psi společnosti. Taková je / všeobecně sdílená představa o poslání novinářů.

Journalists are / watching dogs of the society. This is / a generally shared image of the mission of journalists.

“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. As for the genres of the more closely studied documents, in this first step our attention was focussed on the essay and letter genre.

Our starting assumption was that if the sentence is to be “about” something (i.e. about the Topic of the sentence), this “something” has to be somehow established (anchored) in the memory of the addressees. This is why we first examined the types (assumed as prototypical) $T_{n-1} \leftarrow T_n$ and $F_{n-1} \leftarrow T_n$, that is the pairs of sentences in which Topic refers to the Topic of the previous sentence (“continuous Topic”) or in which the Topic refers to the Focus of the previous sentence (“progression of Focus”). This assumption has been confirmed in both genres, but there was a difference which of the two types prevails in which genre: $T_{n-1} \leftarrow T_n$ occurred twice as often than $F_{n-1} \leftarrow T_n$ in the letter document, while in the essay genre, $F_{n-1} \leftarrow T_n$ occurred three times as often than $T_{n-1} \leftarrow T_n$. With the non-prototypical relations, that is with the types $F_{n-1} \leftarrow F_n$ and $T_{n-1} \leftarrow F_n$, both types occurred rather rarely in the letter genre but the type $F_{n-1} \leftarrow F_n$ was surprisingly frequent in the essay type (13 occurrences as compared to 20 of $F_{n-1} \leftarrow T_n$ and 8 of $T_{n-1} \leftarrow T_n$). Under a more detailed inspection, it has been found that in most of these cases the anaphoric relation of an element in F_n leads from a contextually bound element of 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”.

2.2 Large Data

To obtain a more general picture of the distribution of the different types of “thematic” relations as attested in larger data, we applied the analysis onto a collection of 10 genres, namely (i) advice, (ii) comment, (iii) description, (iv) essay, (v) invitation, (vi) letter, (vii) news, (viii) overview, (ix) review and (x) survey. We put under scrutiny documents containing more than 20 sentences and looked for anaphoric chains globally, that is we did not restrict our search to adjacent sentences. Taking into account anaphoric chains consisting of two elements only, the results obtained for all these genres are as follows: as for the relations leading from the Topic of the given sentence to some preceding sentence, the $F_{n-x} \leftarrow T_n$ sequences prevailed considerably (3 436 cases) over

Table 1. Anaphoric chains.

Frequency	Anaphoric chain
3 436	F – T
3 307	F – F
1 863	T – T
1 439	T – F
643	F – T – T
597	F – F – F
432	T – T – T
…	
184	F – F – F – F
…	
36	F – T – T – T– F
…	
9	F – T – T – F – F – T
etc.	

the $T_{n-x} \leftarrow T_n$ type (1 863 cases); the total number of these typical relations was 5 299. This result indicates that continuous topic, i.e. the anaphoric relations between Topics of two sentences, are considerably less frequent than the progression of focus, i.e. anaphoric reference from the Topic of the given sentence to an element in the Focus of (some of) the preceding sentence(s).

2.3 Non-Typical Cases

However, the relations we consider to be non-typical (leading from the Focus of a given sentence to an element in the Topic or in the Focus of (some of) the previous sentence(s)) occurred surprisingly frequently (the total of 4 746 cases, out of which $F_{n-x} \leftarrow F_n$ type was found in 3 307 cases and the type $T_{n-x} \leftarrow F_n$ was found in 1 439 cases). These figures have led us to a deeper analysis of these non-typical cases. For this purpose we have sorted the material obtained in this step according to the length of the coreference chains, i.e. according to the "course" ("progression") of the given anaphoric relation throughout the document. In this way, we obtained a list (and frequences) of two-element chains, three-element chains etc. sorted by the four above mentioned "directions" of anaphoric relations. Table 1 is an illustration of the resulting data, where in the first column there is the frequency of the given relation, and F(ocus) and T(opic) denote the part of the sentence in which there occur the referring expressions linked by the given anaphoric link. (The first four lines of the Table are those mentioned in Sect. 2.2 above.)

We have put under a more detailed scrutiny the cases of what might be called "continuous foci" (i.e. the type F – F – F etc.) to see under which conditions they arise. For this purpose we have analyzed 40 examples in which the length of the "continuous foci" was 4 and more. Here again, in 29 cases the anaphoric link leads from a contextually bound element of F which supports the necessity to distinguish local topics and local foci with the overall Topic and Focus. The rest of the cases include (i) bridging relations rather than proper coreference, (ii) a list in Focus (e.g. list of exhibitions in a locality), (iii) change of speakers of sentences in the Focus of which the referring expression occurs.

The obtained data have allowed us also to follow the distance between the referring expressions in terms of the number of sentences in between them. The starting hypothesis is that the longer the chain, the more probable is the re-occurrence of the referring expression in the Focus of the sentence. A perfunctory look at the collected data indicates that this is an important factor: e.g. in the above mentioned chain, the distance (indicated by numbers of intervening sentences) is as follows: F *-1-* F *-3-* F *-3-* F. One of the points of our future inquiry will be to investigate the dynamism of discourse in terms of the necessity to re-introduce an item in the Focus part of the sentence based on the "distance" and also in terms of the form of the referring expression, e.g. when a reference by a pronoun (or even a zero pronoun) is possible and when it is necessary to refer to some "fading" item by a noun or a nominal group. For the overall framework and hypotheses for such an inquiry, see Hajičová and Vrbová (1982), Hajičová (2003) and Hajičová and Hladká (2008).

3 Conclusions

In the present contribution we have focused on the intersentential relations based on coreferential chains (both proper coreference and some basic types of bridging relations) with regard to the bipartion of the sentences into their Topic and Focus. We first verify the accepted methodology on a small sample of texts from two genres of the annotated texts from the multi-layered Prague Discourse Treebank 2.0, followed by an analysis of a more representative sample of annotated texts from nine genres. We have also taken into account the length of the anaphoric chains and the length of the segments (in terms of the number of sentences) in between two expressions referring to the same item.

The following observations have been reached:

(a) 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; this is in contrast to the assumption of Fais (2004) based on the low cost and Chamber's (1998) assumption of structural parallelism, but in favour of Poesio et al.'s (2004) finding on the prevailance of shifts to retain relation.

(b) If compared with the studies on thematic progressions in English carried out by Czech linguists (see e.g.Dušková 2008), the structural parallelism seems to be valid for English, thanks to the function of English subject in the grammatically fixed word order. Our observations seem not to support such a parallelism for Czech, a language the word order of which is guided by communicative factors rather than by grammatical rules.

(c) In case there is an anaphoric link leading from the Focus of a sentence to the Topic or Focus of the preceding sentence:

(i) this link frequently leads from a contextually bound element of the Focus of the given sentence, which supports the assumption that it is convenient to distinguish between the "overall" Topic and Focus and the local Topic and Focus; and/or

(ii) the anaphoric relation is of the type of bridging, which is often intepreted as a contrast.

Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (projects GA17-03461S and GA17-06123S). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

Chambers, C. (1998). Structural Parallelism and Discourse Coherence: A Test of Centering Theory, *Journal of Memory and Language,* Volume 39, Issue 4, November 1998, Pages 593-608

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

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

Dušková, L. (2008). 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. Prague, FF UK, 221-247.

Fais, L. (2004). Inferable centers, centering transitions, and the notion of coherence. *Computational linguistics 30*, 119-150.

Grosz, B. and C. L. Sidner (1986). Attention, Intentions and the structure of discourse. *Computational Linguistics*, 12, 175-204.

Grosz, B. J., Joshi, A. K. and S. Weinstein (1995). Centering: A Framework for modeling the local coherence of discourse. *Computational Linguistics, 21*, 203-225.

Hajičová, E. (2003). Aspects of Discourse Structure. In: *Natural Language Processing between Linguistic Inquiry and System Engineering* (ed. by W. Menzel and C. Vertan), Iasi, pp. 47-56.

Hajičová, E. and B. Hladká (2008). What does sentence annotation say about discourse? In *18th International Congress of Linguists*, Abstracts , The Linguistic Society of Korea, Seoul, Korea, pp. 125-126

Hajičová, E. and J. Mírovský (in prep.). Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study. Accepted for LREC 2018.

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

Hajičová, E. and J. Vrbová (1982). On the role of the hierarchy of activation in the process of natural language understanding. In: Horecký J., Ed. , *Coling 82 – Proceedings of the Ninth International Congress of Computational Linguistics.* Amsterdam: John Benjamins. 107-113.

Poesio, M., Stevenson, R., Di Eugenio, B.and J. Hitzeman (2004). Centering: a parametric theory and its instantiations. *Computational Linguistics* 30, 309-363

Rysová, K., Mírovský, J. and E. Hajičová (2015). On an apparent freedom of Czech word order. A case study. In: *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015),* IPIPAN, Warszawa, Poland, ISBN 978-83-63159-18-4, pp. 93-105.

Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J. and Š. Zikánová (2016). *Prague Discourse Treebank 2.0.* Data/software, ÚFAL MFF UK, Prague, Czech Republic, http://hdl.handle.net/11234/1-1905, Dec 2016

Weil, H. (1844). *De l'order des mots dans les langues anciennes comparées aux langues modernes*, Paris: Joubert. 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.