Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study

Eva Hajičová, Jiří Mírovský
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Praha 1, Czech Republic
{hajicova, mirovsky}@ufal.mff.cuni.cz

Abstract

A corpus-based study of local coherence as established by anaphoric links between the elements in the thematic (Topic) and the rhematic (Focus) parts of sentences in different genres of discourse. The study uses the Czech data present in the Prague Dependency Treebank and annotated for surface and underlying syntactic relations, the contextual boundness of tree nodes (from which the bi-partition of the sentence into Topic and Focus can be derived) and the coreference and bridging relations. 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 immediately preceding sentence. In case there is an anaphoric link leading from the Focus of one sentence to the Topic or Focus of the immediately preceding sentence, this link frequently leads from a contextually bound element of the Focus, 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 and the relationship is often interpreted as a contrast. As for the relationship between the relations of the Topic-to-Topic type, due to the word order typological difference for Czech and English, these relations in Czech are not at all related to the syntactic function of subject.

Keywords: discourse, coherence, coreference

1. Introduction

One way how to look at discourse is to view it as a sequence of utterances linked by coherence relations (Halliday and Hasan, 1976). There are several possible ways how to account for these relations: e.g. they may be described on the basis of coreferential links between the elements present in the utterances, or on the basis of some discoursive links between segments of the adjacent utterances. In the present contribution, we first characterize some of the hitherto described approaches to this issue passing over to an examination of a possibility to look at the text coherence taking into account both the information structure of the utterances and the anaphoric relations. The proposed approach is based on a sample of Czech text corpus (the Prague Dependency Treebank 3.0) annotated for deep syntactic relations, information structure, and coreference and discourse relations.

2. Related Theories

One of the most deeply elaborated and best known theories of discourse (local) coherence is the so-called *centering theory* (Grosz, Joshi and Weinstein, 1983) 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 highest ranked entity on the list is the so-called *preferred center*, i.e. the most likely link to the next following utterance. The *transitions* from one utterance to the following one are then specified by rules that capture their ordering: the most preferred is '*continue*', which means that the backward looking center of a given utterance equals the backward looking center of the preceding utterance and at the same time is the preferred center of the given utterance, followed by '*retain*' (the backward looking center of a given utterance equals the backward looking center of the preceding utterance but is not the preferred center of the given utterance), '*smooth shift*' (the backward looking center of a given utterance differs from the backward looking center of the preceding utterance but at the same time is the preferred center of the given utterance), and '*rough shift*' (the backward looking center of a given utterance differs from the backward looking center of the preceding utterance and is not the preferred center of the given utterance), in this order. 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): "utterances that 'continue' the 'topic' of a previous sentence in a prominent position impose a lower inferential load, and are thus more coherent, than utterances which relegate the topic to less prominent position or which change the topic".

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). In previous studies of structural parallelism it was shown that an ambiguous pronoun (e.g., *him*) is biased to corefer with an antecedent in the same structural position (e.g., *Paul*). Of interest was whether parallelism can also influence the capacity for a pronoun to facilitate discourse comprehension and whether the centering model of discourse coherence can account for such effects. Most generally, centering predicts that a pronoun will increase coherence when it corefers with the subject of the previous utterance and that a single pronoun is sufficient to optimize local coherence. Three experiments are reported in the study addressing the interpretation of ambiguous pronouns, the comprehension of utterances containing a pronoun whose antecedent occupies either a parallel or nonparallel position, and also evaluating how the presence of multiple anaphoric links facilitates

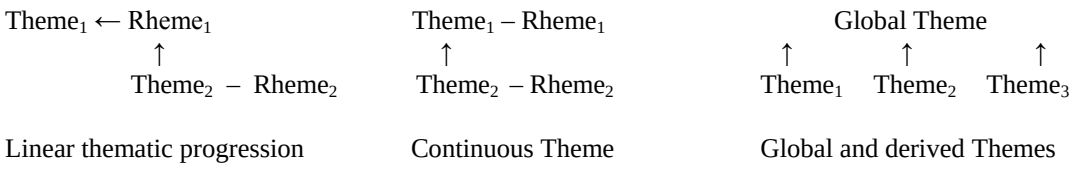

Figure 1: Types of thematic progressions according to Daneš (1970)

comprehension. Overall, 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. (2004). The data used for that work are texts from two of the three domains of the GNOME corpus. The annotation included e.g. the break-up of sentences into clauses and the assignment of grammatical functions and anaphoric relations incl. bridging reference. An automatic script uses this information to compute utterances and the ranking of forward and backward looking centers. The study has reached some interesting results. For example, only the rule stating that if any forward looking center is pronominalized then the backward looking center is also pronominalized has been confirmed. The results concerning the constraint in its strong version that all utterances of a segment except for the first have exactly one backward looking center are especially negative; only if indirect realization is allowed, the constraint holds and is violated by between 20 to 25% of utterances. Another interesting observation is that if ranking is only required to be partial, some utterances end up with more than one backward looking center. 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.

3. Thematic Progressions

Our long-time study of the information structure of the sentence (its Topic-Focus Articulation) has led us to the conviction that this aspect of the sentence structure is a good "bridge" towards a study of (at least one aspect of) the dynamic development of discourse. This, of course, is not a new idea: to our knowledge, its first comprehensive treatment, though clad in psychological rather than linguistic considerations, was given by Weil (1844, quoted here from the 1978 English translation). Weil 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). (It should not be overlooked that Weil also noticed a possibility of a reverse order called by him '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. Schematically, these three types can be captured as follows (see Fig. 1); the arrow denotes the direction of the relation.

In a slightly different but closely related vein, Firbas develops his ideas of the thematic and rhematic layers of a text (1995).

4. Corpus-Based Case Study

4.1 Methodology and its testing on a small sample

In our present corpus-based case study 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 particular, we want to verify if the classical observations valid for English as a language with a grammatically fixed word order, namely that there is a prevailance of "constant theme" (based on Mathesius´ 1947 study on the thematicity and "continuity" of English subject, and further analyzed esp. by Dušková, 2008; 2010), are also valid for a typologically different language, namely Czech, in which the word order is not guided by grammatical rules.

For this purpose, we use the data from the Prague Dependency Treebank (PDT in the sequel, for the most recent version see Hajič et al., 2018), which offers a good testing bed as it provides – in addition to the dependency underlying (deep) syntactic relations – annotation of (i) contextual boundness[1] from which the Topic–Focus

[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 PDT 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; see e.g (1) *John entered the room.* (2) *He first went to the window.* In (2), the window refers to cognitively 'known' object, i.e. known from the preceding context (the window of the room), but the sentence is structured in such a way that this element is contextually non-bound, it belongs to the focus of (2), as documented by placing the pitch on them if the sentence is read aloud.

bipartition of the sentence (TFA) 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) the "progressive" rheme (Focus) in linear thematic progression, i.e. the Topic of the given sentence is anaphorically related to the Focus of the previous sentence, and (ii) continuous theme (Topic), i.e. the Topic of the given sentence is anaphorically related to the Topic of the previous sentence.

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 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, see Sgall, 1979, p. 180; Sgall et al. 1986, pp. 216ff; the original algorithm was later implemented and then tested on the whole of PDT and the results were reported in Hajičová et al., 2005, see also Rysová et al., 2015). 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 $T_{n-1} \leftarrow T_n$ and called above continuous Topic):

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 $F_{n-1} \leftarrow T_n$ and called above progression of Focus):

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.

(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.

The genres of the selected documents were (i) interviews, (ii) plot, (iii) news, (iv) letter, and (v) essay, all the documents in PDT being of a journalistic domain. A first perfunctory look at the annotated data indicated that the interviews are a special kind of text, basically with two speakers, and that anaphoric links to the speakers (identified by pronouns or "dropped" pronouns) prevail, being mostly of the $T_{n-1} \leftarrow T_n$. type. Also the news and the texts marked as 'plot' did not provide an interesting material for the kind of analysis we aimed at, so that our attention was focussed first on the essay and letter genre (but see below Sect. 4.3 for an extended search).

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 anchor often is reflected in the text by an anaphoric reference from the Topic. 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 the 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 other, non-prototypical relations, 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

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

called "local Topic" and "local Focus". An illustrative example are the sentences in (iii) above, repeated here for convenience:

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.

The two expressions connected by an anaphoric link (bridging type) are *pošty* [mail] and *zásilka* [article], both in the Focus part of the sentences they are part of. However, the element *zásilka* [article] in the (global) Focus of the second sentence of the segment carries the TFA value "contextually bound" and as such is in the local Topic of the sentence; the division of the (underlined) embedded clause into its own Topic and Focus is indicated by a double bar.

4.2 Verification of the results on a larger corpus

To confirm our observations and to find a more substantial support for our initial assumption, we carried out a second step in the analysis, in which: (i) we put under scrutiny a larger amount of data from the essay genre, (ii) we complemented our identification of the "overall" 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, and (iii) we 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 new sample contained another 100 annotated sentences from the genre of essay, 79 of which were linked by links of coreference or bridging relations. (It should be noted that we followed only links between two adjacent sentences and did not analyze sentences the links from which pointed to some more distant preceding context.) In this sample, the $F_{n-1} \leftarrow T_n$ sequences prevailed only slightly (28 cases) followed by the $F_{n-1} \leftarrow F_n$ type (24 cases), the $T_{n-1} \leftarrow T_n$ type (11 cases) and the $T_{n-1} \leftarrow F_n$ type (8). This is to say that the ratio between what we consider to be typical relations (from the 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 $F_{n-1} \leftarrow F_n$ type relations, it has been confirmed that in most cases, the anaphoric link leads from a contextually bound element of F_n which again may serve as a support to distinguish local topics and local foci from the overall Topic and 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 $F_{n-1} \leftarrow F_n$ 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 obtain a more general picture of the distribution of the different types of anaphoric relations as attested in larger data, we applied the analysis onto the whole subset of the essay genre in the PDT corpus; this sample contains 189 documents with the total of 6 858 sentences, among which 4 606 adjacent sentences contained a pairwise anaphoric link. The figures obtained confirm even more clearly the picture described above: the $F_{n-1} \leftarrow T_n$ sequences prevailed considerably (1 771 cases) over the $T_{n-1} \leftarrow T_n$ type (1 278 cases) while the number of the non-prototypical links was much lower (the $F_{n-1} \leftarrow F_n$ type with 1 004 cases and the $T_{n-1} \leftarrow F_n$ type with 553 cases).

4.3 Application of the analysis to different genres

As the PDT annotated material allows for a comparison of the obtained results with respect to different genres,[3] 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; we have identified the total of 17 307 anaphoric links and the results obtained for all these genres taken together are as follows: as for the relations leading from the Topic of the given sentence to the preceding sentence, the $F_{n-1} \leftarrow T_n$ sequences again prevail (6 292 cases, i.e. 36%) over the $T_{n-1} \leftarrow T_n$ type (5 029 cases, i.e. 29%); the total number of these typical relations is 11 321 (65%). 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 the preceding sentence. As for the non-typical relations, i.e. relations leading from the Focus of the given sentence to the Topic or Focus of the preceding sentence, they occur only in 5 986 cases (35%); among them, the $F_{n-1} \leftarrow F_n$ sequences prevail (3 665, 21%) over the $T_{n-1} \leftarrow F_n$ type (2 321 cases, 14%), see Table 1 for the distribution of the types of thematic progressions according to the type of anaphoric relation.

	All anaphora		Coreference		Bridging	
F ← T	6 292	36%	5 091	38%	1 201	31%
T ← T	5 029	29%	3 834	29%	777	20%
F ← F	3 665	21%	2 888	21%	1 195	31%
T ← F	2 321	14%	1 629	12%	692	18%

Table 1: Distribution of the types of thematic progressions according to the type of anaphoric relation

Looking at the genres separately, nine of the ten analyzed genres provide a similar distribution of thematic types, the only conspicuous differences being attested in the genre of "overview" with a balanced occurrence of the $F_{n-1} \leftarrow T_n$, $T_{n-1} \leftarrow T_n$ and $F_{n-1} \leftarrow F_n$ sequences (29%, 29% and 31%, respectively) and a considerably lower frequency of the $T_{n-1} \leftarrow F_n$ type (11%, which corresponds to the general situation). So far, we cannot offer any explanation for this phenomenon. Another observation relates to the genre of "letter": there the prevalence of the $F_{n-1} \leftarrow T_n$ is even more perspicuous than with the other types (44% compared to 28% of $T_{n-1} \leftarrow T_n$ for all relations, and 50%

[3] There are 20 different labels for the genre categories assigned to the PDT documents (Zikánová et al. 2015, p. 27f.)

vs. 35% for bridging only). As a matter of fact, looking at the bridging type of anaphoric relations separately, the distribution of the types of thematic relations differs almost in all genres from the distribution of the thematic relations identified by coreferential links. The explanation may concern two points: (i) first and most importantly, the manual annotation of bridging relations is very difficult in general and open for annotation inaccuracy, the more so in our corpus where we recognize only a few basic types of bridging; (ii) with the genre of letters and advice in particular, we may expect a more or less intimate, personal relations between the writer/author and the addressee, which may lead to a more frequent use of (indirect) anaphoric links (bridging), and the interaction has also a more familiar character, which may lead to a higher occurrence of contrast in focus.

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 a possible relationship between the anaphoric links and the surface syntactic function of the elements in question. As mentioned above, the analysis of thematic progressions in the writings on English (e.g. Dušková, 2008; 2010) suggests that there is a predominance of "constant theme", which, for English, means, that there is a "continuity" of subject (see Mathesius, 1947 for the function of subject in English). Our material documents that this is not the case in Czech: in the second sample of 100 sentences, among the $T_{n-1} \leftarrow T_n$ type (i.e. the type relevant for the issue under investigation), there was no instance of a subject-to-subject link.

5. Summary and Future Work

To sum up our observations based on annotated corpus (the PDT multi-layered annotation of Czech sentences), the following results have been reached:

(a) among the four possible types of the relationship 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 predominance of shifts to retain relation;

(b) in case there is an anaphoric link leading from a sentence to the Focus of the next following sentence,

(i) 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

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

(c) as for the relationship between the relations of the Topic-to-Topic type, due to the word order typological difference for Czech and English, these relations in Czech are not at all related to the syntactic function of subject.

We are aware that the observations and results presented in this paper are only first steps in the corpus-based study of the relationships between the Topic-Focus articulation of the sentence and the anaphoric relations. The resource we have at our hands offers a possibility to follow several aspects of this relationship, out of which the following issues are on our future programme:

(i) The multilayered annotation of the PDT allows for a systematic study of the relation between the syntactic structure – both surface (in terms of subject, object, etc.) and deep (in terms of Actor, Patient, Addressee, etc.) – and the frequency of the types of thematic progressions; in this way we may arrive at explanations based on the typological differences between different languages.

(ii) Not only the adjacent sentences but also sentences linked by anaphoric relations "at a distance" should be examined, with the aim to investigate whether the length of the chain of anaphoric relations and the size of the "gap" ("hole") in between the coreferring expressions make a difference.

(iii) In connection with the point (ii), it will be interesting to examine whether the size of the above-mentioned "gap" makes a difference in the use of a particular anaphoric type or a preference of the use of a particular type of surface expression (pronoun, bare noun, nominal group etc.).

(iv) A more complex task consists in the examination of the dynamics of discourse in terms of the activation of elements of the knowledge shared by the speaker/author and the addressee (for the formulation of the task and the first text analysis see Hajičová, 1993; 2003, and Hajičová and Hladká, 2008). The multilayered PDT annotation offers a useful resource by capturing the forms of the referring expressions, their syntactic functions and the values of contextual boundness, all being the factors determining or influencing the discourse flow.

Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of the Czech Republic (projects GA17-03461S and GA17-06123S) and the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071). 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).

Bibliographical 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, 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.

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

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

Firbas, J. (1995): On the thematic and rhematic layers of a text. In: Warvick B. et al., *Organization of Discourse. Proceedings of the Turku Conference 1995*, pp. 59–72.

Grosz, B. (1977). The representation and use of focus in dialog understanding. *Technical Note 15,* Artificial Intelligence Center, SRI International, Menlo Park, California.

Grosz, B. J., Joshi, A. K. and S. Weinstein (1983). Providing a unified account of definite noun phrases in discourse. In: *Proceedings of the 21st Annual meeting of the Association for Computational Linguistics,* Cambridge, Mass., 44–50.

Grosz, B. and Sidner, C. L. (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. (1993). *Issues of sentence structure and discourse patterns*. Charles University, Prague.

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, 47–56.

Hajičová E., Havelka J. and K. Veselá (2005). Corpus Evidence of Contextual Boundness and Focus. In: *Proceedings of the Corpus Linguistics Conference Series*, University of Birmingham, Birmingham, UK, ISSN 1747-9398, pp. 1–9.

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, 125–126.

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.

Halliday, M. A. K. – Hasan, R. (1976): *Cohesion in English*. Longman: London

Mathesius, V. (1947). O funkci podmětu. [On the function of subject.]. In *Čeština a obecný jazykozpyt,* Melantrich, Prague, 100–103.

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)*, Warszawa, Poland, 93–105.

Sgall, P. (1979): Towards a Definition of Focus and Topic. In *Prague Bulletin of Mathematical Linguistics,* 31:3–25; 32:24–32; reprinted in *Prague Studies in Mathematical Linguistics*, 7, 1981, 78:173–198Sgall, P. Hajičová, E. andJ. Panevová (1986): *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. In Mey, J. L. (ed). Dordrecht: Reidel and Prague:Academia.

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.

Zikánová, Š. et al. (2015). *Discourse and Coherence.* Prague: Charles University, MFF ÚFAL, Prague, Czech Republic.

Language Resource References

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 (2018). *Prague Dependency Treebank 3.5.* Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, Prague, Czech Republic, LINDAT/CLARIN PID: http://hdl.handle.net/11234/1-2621.