An Annotated Corpus as a Test Bed for Discourse Structure Analysis

Eva Hajičová, Barbora Hladká and Lucie Kučová
Institute of Formal and Applied Linguistics, Charles University
Malostranské nám. 25
118 00 Prague
Czech Republic
{hajicova,hladka,kucova}@ufal.mff.cuni.cz

Abstract

A dynamic approach to discourse structure is characterized using the notion of degrees of salience of items in the stock of knowledge the speaker assumes s/he shares with the hearer. A preliminary algorithm of salience assignment (based primarily on the appearance of the nodes of the dependency tree representing the underlying structure of the sentence to topic or focus) has been implemented and a visualization of its results has been produced in order to make the implications of proposed discourse analysis more perspicuous.

1 Motivation

Most of the natural language processing systems require context to be taken into account to get adequate results. In general, context is the information directly present in the document processed, i.e. knowledge of words and the relations among them, as well as the information given by a broader context of situation. In our project, we have tried to "read out" as much information as possible from the sentence underlying structure represented in the form of a dependency tree capturing (a) underlying syntactic structure (with functions such as Actor, Objective, Addressee, etc.), (b) the information structure of the sentence (its topic-focus articulation, TFA), (c) coreference relations.

The linguistic data we have used for our experiment are those of the Prague Dependency Treebank [11], the annotation of which is based on the framework of the Functional Generative Description [14]. We work with the notion of the degrees of salience (activation) of the items in the stock of shared knowledge together with the representation of the dynamic development of the discourse by means of changes of these degrees.

We present an automatic procedure - the *salience* algorithm - capturing a dynamic character of the stock of shared knowledge. We first (in Section 2) give a description of theoretical background we work with. Section 3 is devoted to the data, their manual coreference annotation and their automatic Topic/Focus annotation. Section 4 documents the application of the *salience* algorithm on the data and the visualization of its output is reported in Section 5. Section 6 brings a summary and an outlook.

2 Theoretical background

The approach to discourse patterns described in the present contribution is based on the following four assumptions:

1. natural language communication is an interactive process the dynamics of which is guided i.a. by the hierarchy of salience of the discourse elements;

2. a description of discourse patterns is to be firmly rooted in a due account of sentence structure;

3. one of the aspects of sentence structure (specified not only for utterances as sentence occurrences but already for sentences as types, in their underlying structures) most relevant for a description of discourse is the information structure of the sentence (its topic-focus articulation, TFA);

4. a consistent and comprehensive annotation of a text corpus by values of attributes both for TFA and coreference relations is a very good test bed for the tenets of any theory.

In the treatment of TFA we subscribe to the basic opposition between *contextually bound* (*cb*) elements of the sentence, prototypically in Topic, be it a *contrastive (part of) Topic* or not, and *non-bound* (*nb*) element, typically in Focus. The opposition of contextual boundness is understood as a linguistically structured counterpart of the cognitive distinction between ”given” and ”new” information, rather than in a straightforward etymological sense. A basic algorithm was formulated to determine the appurtenance of a lexical occurrence to the Topic (T) or to the Focus (F) of the sentence [14]; for an implementation of the algorithm and its testing on PDT see [6].

It may be assumed that there is a finite mechanism the addressee can use to identify the referents in a discourse. If the backbone of such a mechanism is seen in the hierarchy (partial ordering) of salience, then it can be understood that this hierarchy typically is modified by the flow of discourse in a way that was proposed by [5] and [4]. In the flow of a discourse, prototypically, a new discourse referent emerges as corresponding to a lexical occurrence that carries *nb*; further references to this item carry *cb* (contrastive or not). Their referents are determined by their degrees of salience. It appears to be possible to capture at least certain aspects of this hierarchy by the *salience* algorithm that was designed to capture a dynamic character of the stock of knowledge assumed by the speaker to be shared by her/him and the hearer(s): not only the repertoire of items it includes is changed but also their activation (salience).

Before we start with the *salience* algorithm specifications and the experiments description we review the terminology we use:

- A **referent** is an object referred to in the given discourse.
- An **item** is a mental image of the referent, which is a member of the stock of shared knowledge.
- A **referring expression** is a lexical representation of a referent.
- A **coreference chain** is a list of the item’s referring expressions with the *anaphor* → *antecedent* relation.

3 Data

For our pilot experiment, we randomly selected forty documents from the PDT 2.0 documents annotated morphologically, syntactically and tectogrammatically Such a small amount of data has been selected intentionally mainly because of the two following reasons:

- If the annotated data are not available, the only way to verify any linguistically-based algorithm (including salience algorithm) lies in a manual selection of data of very limited amount and more-or-less manual processing of a given algorithm. Once the annotated data are at hand, the automatization of algorithm brings the significant quantitative step a forward - exactly what we expected from our experiment.
- The coreference annotation present in PDT 2.0 is restricted to those cases in which the anaphors are rendered as pronouns (also zero pronouns in Czech as a pro-drop language). Regarding the *salience* principle the coreference annotation must be broadened to capture also coreference relations in which anaphors are expressed by nouns and noun phrases. We are aware that forty documents cannot present a representative sample to formulate annotation guidelines extension precisely enough. However, they serve quite well to start to formulate the guidlines and refine them especially when only one annotator is working on it, as in our case.

3.1 Coreference annotation

The annotation scheme of PDT 2.0 consists of three layers: morphological, analytical and tectogrammatical. Within this system, coreference annotation is captured at the tectogrammatical layer [7]. This annotation is restricted to those cases of coreference in which the anaphors are rendered as pronouns. Nevertheless, the annotation scheme is rather broad: it covers a large set of phenomena related to grammatical coreference (including relative, reflexive, and reciprocity pronouns, and the control relation both for verbs and for nouns of control), and at the same time it makes it possible to cross the sentence boundaries in the field of textual coreference, and take into account not only the cases where demonstrative and personal pronouns refer to the co-text (incl. clauses, sentences or whole segments of text), but also those situations in which the referent is ”out” of the co-text.

The phenomena of coreference and anaphora (anaphora resolution) have attracted the attention of many researchers all over the world since 1970s, and many approaches have been developed. Considering the fact we use the data of the PDT, we mention only some of the corpus-based approaches: the PDT.2 data are comparable to those produced by the project of University of Stendhal and Xerox Research Centre Europe [17] both in their volume of data and in their focus on pronoun annotation. There is a large set of problems in the annotation of noun anaphora which can be illustrated with the following references to some researches: from the work of the Research group in Computational Linguistics at University of Wolverhampton that focused on direct nominal anaphora [8], through the UCREL annotation scheme [1] or DRAMA annotation scheme [10] which crosses the boundaries of direct nominal anaphora towards indirect (bridging) anaphora to the approaches focused on the annotation of bridging anaphora like [2] or [18]. In the light of the aim of our paper the approaches concerning the correlation between anaphoric expressions and centering, f. ex. [12], [13], [15], [3] are very interesting.

In the presented experiment, the coreference annotation guidelines have been broadened in order to capture also coreference relations in which anaphors are expressed by nouns and noun phrases. For the time being, the annotation scheme covers only the *identity relations* for noun phrases. The following types of relations can be distinguished:

1. pronominalisation (incl. zero pronouns) – the antecedent of the anaphor is a pronoun
2. repetition of the noun – the noun of the antecedent and the anaphor are the same
3. the antecedent of the anaphor is realized by a different noun:
   a) proper nouns (named entities)
   b) synonymous expressions
   c) anaphors which are specified by an ”identifier”. There is no direct relation between the anaphor and its antecedent; their relationship is indicated only by the identifier – demonstrative pronouns and some adjectives (*this*, *that*; *given*, *said*). This type is a transitional point reaching into the larger area of associative anaphora.

The taxonomy just presented may seem to be based on the forms of anaphoric expressions too much, but it should be interpreted in more ways.

Firstly, according to some theories of bridging descriptions, we can regard all the items in 3. as bridging (associative) anaphors. What ”bridges” the distance between the anaphor and its antecedent, be it more (synonymy) or less (proper nouns) grounded in the hierarchy of the lexical system, is the world knowledge. Secondly, we assume that every use of a different noun in the position of anaphoric expression may bring new information – the scale ranges from stylistic nuances to considerable modifications of sense. Sometimes, it is hard to identify whether the referent of the noun is still the same, or whether a new discourse entity emerged. The taxonomy presented above is an example of a hierarchical classification of possible modifications.

Although own annotation is focused on identity relations, some steps have already been undertaken also in the wide range of associative relations. Some element-subset relations as well as cases when a head noun of the antecedent is more specified by its dependents are covered.

We are aware that this is just the first step and that once we take into account also those relations in which the anaphors are expressed by adjectives or verbs, several complex issues should be analyzed, especially those concerning the associative relations.

A sample document of journalistic style consists of nine sentences (1) through (9). The English translation is a literal one, preserving as much as possible the constructions and the word order of the Czech sentences; the words in the brackets have no equivalents in Czech.

(1) Bělorusko$_{1}$: zastavení likvidace$_{2}$ arzenálů$_{3}$. [lit. Belorussia: stopping (the) liquidation (of) arsenal.]

(2) Moskva - [lit. Moscow -]

(3) Běloruský prezident Alexandr Lukašenko$_{5}$ nařídil$_{4}$ pozastavit likvidaci$_{2}$ vojenské techniky$_{3}$ na území republiky$_{1}$. [lit. Belorussian president Alexandr Lukashenko ordered (to) stop (the liquidation) (of) military technology on (the) territory (of-the) republic.]

(4) Oznámil to$_{4}$ v Minsku na čtvrtečním slavnostním večeru k oslavám Dne obránců vlasti. [lit. (He) announced this in Minsk (on) Thursday ceremonial evening (on the) celebration (of) Day (of) Defenders (of) Motherland.]

(5) Opatření$_{2}$ se týká tanků, letadel, obrněných transportérů a$_{3}$ bojových vozidel pěchoty. [lit. (The) measurement Refl. concerns tanks, planes, armored carriers and military vehicles of infantry.]

(6) Podle Lukašenka$_{5}$ prý byl tento krok vyvolán ani ne tak nedostatkem finančních prostředků, jako spíše ”patrným porušováním vyvtořené rovnováhy sil ve světě” [lit. According-to Lukashenko allegedly was this step-Nominative evoked (by) not so (the) lack (of) financial means, as rather (by a) visible breaking (of) created balance (of) powers in (the) world.]

(7) Agentura Interfax soudí, že prezident$_{5}$ měl na mysli přání východoevropských zemí vstoupit$_{8}$ do Severoatlantické aliance, což$_{8}$ by pro Bělorusko$_{1}$ znamenalo bezprostřední sousedství s NATO$_{9}$. [lit. (The) agency Interfax assumes that (the) president had on mind (the) wish (of) East-European countries (to) enter into (the) North-Atlantic alliance, which would for Belorussia mean immediate neighborhood with NATO.]

(8) Lukašenko$_{5}$ také řekl, že má pochybnosti o dosud deklarovaném neutrálním statusu Běloruska$_{1}$, uvedl, že je pro nový systém národní bezpečnosti, a oznámil, že ustavil ”pracovní skupinu pro vytvoření vojenské doktríny Běloruska$_{1}$”. [lit. Lukashenko also said that (he) has doubts about (the) hitherto declared neutral status of Belorussia, (he) stated that (he) is for (a) new system (of) national security, and announced that (he) put-together ”(a) working group for creation (of) military doctrine (of) Belorussia”.]

(9) Přitom ujistil, že v souladu s republikovou legislativou ”žádný běloruský voják nebude bojovat za hranicemi Běloruska$_1$”. [lit. At-the-same-time (he) assured that in accordance with republic’s legislation ”no Belorussian soldier will-not fight outside (the) borders (of) Belorussia”.]

The indices by some of the words indicate that they belong to a particular coreference chain. Table 1 provides a complete list of all nine coreference chains we have taken into account in the sample document. The nine chains represent the following items: [1] *Belarusssia*, [2] *liquidation*, [3] *arsenal*, [4] *to order*, [5] *Lukashenko*, [6] general Addressee of the verb *to order* which in the surface is deleted, but is restored in the underlying representation of sentence (3) as well as lemma *&Cor* for the subject of the embedded infinitive clause with verbs of control, i.e. subject of the verb *to stop*, [7] *country*, [8] *to enter*, [9] *NATO*.

Every coreference chain has a form of a list of triples consisting of a referent’s tectogrammatical lemma, a TFA value and the sentence identification a given referent appears in (`lemma/[tfc]/(id)`). TFA value *f* stands for a node contextually non-bound, value *t* for a node contextually bound and non-contrastive, and *c* for a node contextually bound and contrastive.

[1] Bělorusko/f/(1) republika/f/(3) Bělorusko/c/(7) Bělorusko/t/ Bělorusko/f/(8) Bělorusko/t/(9)
[2] likvidace/f/(1) likvidace/f/(3) opatření/t/(5)
[3] arzenál/f/(1) technika/f/(3) a[tank/f/ letadlo/f/ transportér/f/ vozidlo/f/]///(5)
[4] nařídit/f/(3) ten/t/(4)
[5] Lukašenko/c/(3) #PersPron/t/(4) Lukašenko/t/(6) prezident/t/(7) Lukašenko/t/ #PersPron/t/ #QCor/t/ #PersPron/t/ #PersPron/t/(8) #PersPron/t/(9)
[6] #Gen/t/ #Cor/t/(3)
[7] země/f/ #Cor/t/(7)
[8] vstoupit/f/ co/t/(7)
[9] aliance/f/ NATO/t/(7)

Table 1: Coreference chains in the sample document

Going back to the terminology we listed in Section 2, we can give the examples of the mentioned terms. Let us illustrate them on the [3] coreference chain being the list of the following *anaphor* → *antecedent* relations (with the specified No. of sentence the referring expressions are included in): *arzenál* (1) → *technika* (3)→ *tank, letadlo, transportér, vozidlo* (5). All these expressions refer to an item that can be roughly characterized as a kind of military technical equipment.

3.2 Annotation

The annotation tool TrEd [16] used for the syntactical and tectogrammatical annotation of PDT 2.0 was used for the extendend coreference annotation with the modification so that not only coreference links are visualized as arcs pointing from the anaphor to its antecedent, but the anaphor and its antecedent obtained the same id as the other members of the same coreference chain.

To exemplify the coreference annotation with TrEd, we provide Figure 1 depicting the tectogrammatical representation, coreference and TFA annotation of the sentence (3). Every node of the tree is accompanied by a list of attributes - we display only those attributes that are relevant to our task, namely the tectogrammatical lemma, the TFA value and the coreference chain id. Whereas the tectogrammatical lemmas and TFA values are specified for each node (except the technical one), the coreference identification is added only to nodes participating in some coreference chain (compare the nodes *nařídit*$|f|4|$ and *prezident*$|f|$).

As a supplement to Figure 1, we present Figure 2 displaying the tectogrammatical representation of the sentence (4) which follows the sentence (3) in the sample document. Figure 2 was selected to demonstrate how the coreference chains cross the sentence boundaries. The nodes *#PersPron*$|t|5|$ and *ten*$|t|4|$ in Figure 2 have their antecedents *Lukašenko*$|c|5|$ and *nařídit*$|f|4|$, respectively, in Figure 1.

3.3 Data statistics

Table 2 overviews statistics on the data after the extended coreference annotation. The average number of coreference chains and the average length of them are the most interesting figures. It is quite difficult to make any conclusion whether there is or not some mutual relation between them. In a similar way, this concerns also the relation between the length of the document (i.e. the number of sentences) and the number of coreference chains.

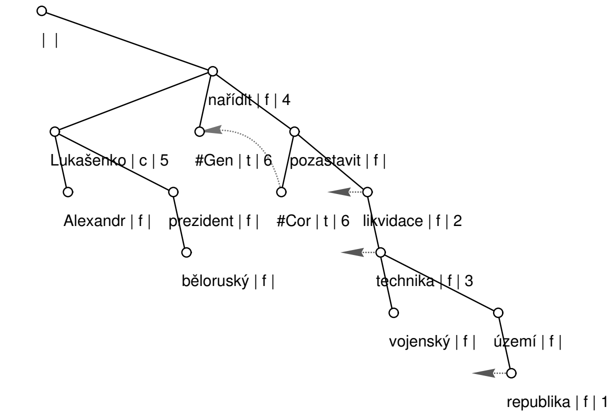

Figure 1: The tectogrammatical tree, TFA and coreference annotation of the sentence (3)

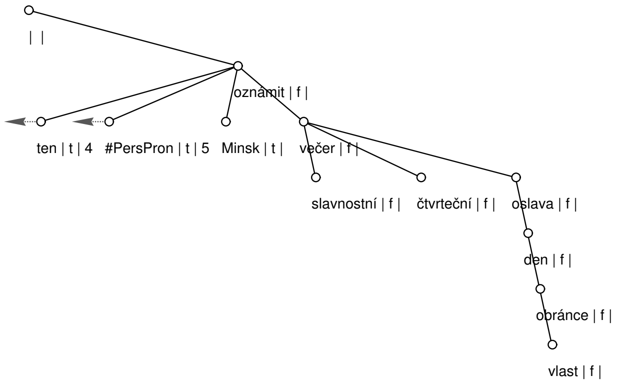

Figure 2: The tectogrammatical tree, TFA and coreference annotation of the sentence (4)

We got these three numbers for all the forty documents, sorted them increasingly according to the number of sentences and present the figures in Table 3. The row in boldface specifies the figures from the sample document: nine sentences, nine coreference chains (of length 6,3,3,2,10,2,2,2,2) and of 3.6 average chain length.

# documents (files)	40
# sentences	509
# word tokens	8,760
average length of sentence	12.72
number of documents on sport	50%
number of documents on politics	50%
average number of coreference chains	11.05
average length of coreference chains	3.225

Table 2: Basic characteristics of the data

3.4 Topic and Focus

An algorithm of the bipartition of the sentence into its (global) Topic (T) and Focus (F) on the basis of the values of the TFA attribute (t, c, or f) has been formulated [6]. We applied it to our data, as visible in Table 4 similar to Table 1 - in addition to three already listed attributes, we provide an attribute specifying if an item is mentioned in the T or in the F. Since an independent manual T/F annotation is not yet completed, we are not able to evaluate the performance of this algorithm. However, the complete input information (`lemma/[tfc]/[TF]/(id)`) needed

#Ss	#Chs	AvLChs	#Ss	#Chs	AvLChs
2	2	2.000	12	19	2.947
2	2	2.000	12	4	5.000
3	5	2.000	12	8	4.000
4	3	4.000	14	15	2.600
4	4	2.750	14	16	4.000
5	5	2.400	15	15	3.600
6	5	2.800	15	18	3.056
6	7	2.714	15	9	3.000
6	8	3.250	16	20	2.850
6	9	3.000	17	11	3.909
8	11	3.091	17	20	3.200
8	6	3.167	18	18	2.833
8	8	3.375	18	7	5.571
8	9	2.556	19	14	2.500
8	9	2.778	19	14	4.429
9	**9**	**3.556**	22	13	3.077
10	10	3.300	25	21	2.810
10	12	2.417	27	19	4.053
10	9	4.333	28	17	3.118
11	9	2.667	40	22	4.273

Table 3: Number of sentences (#Ss), number of coreference chains (#Chs), and average length of coreference chains (AvLChs): statistics from the data

by the *salience* algorithm is available.

[1] Bělorusko/f/F/(1) republika/f/F/(3) Bělorusko/c/F/(7) Bělorusko/t/F/ Bělorusko/f/F/(8) Bělorusko/t/F/(9)
[2] likvidace/f/F/(1) likvidace/f/F/(3) opatření/t/T/(5)
[3] arzenál/f/F/(1) technika/f/F/(3) a[tank/f/F/ letadlo/f/F/ transportér/f/F/ vozidlo/f/F/]///(5)
[4] nařídit/f/F/(3) ten/t/T/(4)
[5] Lukašenko/c/T/(3) #PersPron/t/T/(4) Lukašenko/t/T/(6) prezident/t/F/(7) Lukašenko/t/T/ #PersPron/t/F/ #QCor/t/F/ #PersPron/t/F/ #PersPron/t/F/(8) #PersPron/t/T/(9)
[6] #Gen/t/T/ #Cor/t/F/(3)
[7] země/f/F/ #Cor/t/F/(7)
[8] vstoupit/f/F/ co/t/F/(7)
[9] aliance/f/F/ NATO/t/F/(7)

Table 4: Coreference chains in the sample document with the Topic/Focus and TFA annotation

4 The *salience* algorithm

The knowledge-based *salience* algorithm was formulated to capture the dynamic character of the stock of knowledge assumed by the speaker to be shared by her/him and the hearer(s): not only the repertoire of items it includes is changed but also their activation (salience). The algorithm contains the following four rules, with $dg_x{}^n(r)$ to be read as 'an item x represented by the referent r has the salience degree $dg_x{}^n(r)$ after the n-th sentence of a document is uttered, i.e. salience degree of the item is modified after each sentence starting with sentence in which the item has appeared firstly:

1. $dg_x{}^n(r) = -1$ if r carries TFA value t or c in the n-th sentence.
2. $dg_x{}^n(r) = 0$ if r carries TFA value f in the n-th sentence.

3. $dg_x{}^n(r) = dg_x{}^{n-1}(r) - 2$ if r is not included in the n-th sentence and has been mentioned in the Focus of the last (not necessarily immediately) preceding sentence ($(n-1)$-th through 1-st sentence).

4. $dg_x{}^n(r) = dg_x{}^{n-1}(r) - 1$ if r is not included in n-the sentence and has been mentioned in the Topic of the last (not necessarily immediately) preceding sentence ($(n-1)$-th through 1-st sentence).

5 Discussion of the results

We illustrate the development of salience degrees during a discourse on the sample document. In short, the document informs the reader that president Lukashenko ordered specific measures concerning the liquidation of the military arsenal with regard to NATO's future.

When one is to interpret the numerical data, visualization of them if at all possible can help a lot. We choose the "R" system for statistical environment and graphics [9] to draw plots capturing the development of the salience degree. We choose such an approach that the salience curves of all items included in a document are plotted into one plot with the sentences on the x-axis and with the salience degree on the y-axis. The curves are distinguished by the colors as well as the coreference chains are and the coreference chain id is used in plotting points. Figure 3 displays the visualization of the development of salience degrees in the sample document.

The visualization of the application of the algorithm indicates the ways in which such a dynamic account of discourse structure may be applied. First, a certain segmentation of the texts analyzed is displayed: one can imagine that vertical lines can be drawn between those parts of discourse in which certain items keep a higher degree of activation and do not 'fade away' too far. On the other side, horizontal lines can be imagined to indicate certain thresholds for the possibility of a weaker (pronominal) referential expression to be used, or the necessity for a stronger reference by a noun or a more descriptive noun group. Also the topic of a segment of the discourse can be determined on the basis of the groupings of items on the top of the schema for the given segment.

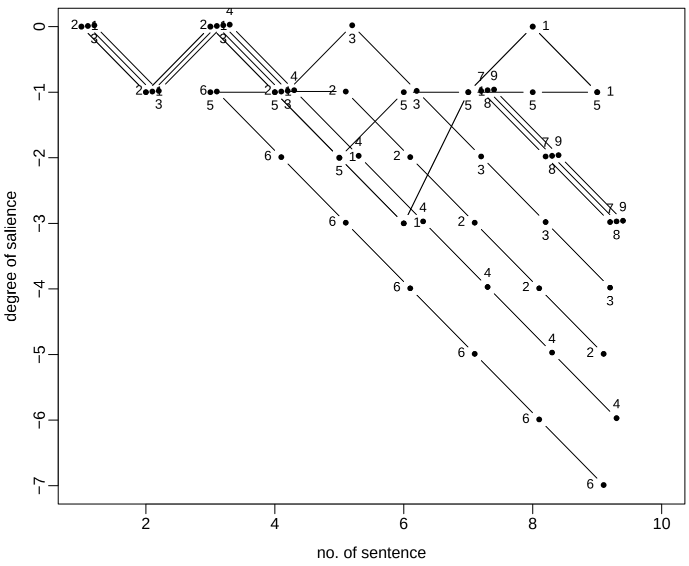

Figure 3: Development of the salience degrees in the sample document: vizualization

6 Conclusion

In the present contribution, we have reported on an algorithm of the assignment of the degrees of salience of items of knowledge assumed by the speaker to be shared by him/her and the hearer. The implementation is based on the data from the annotated corpus of Czech which capture both the syntactic structure of the sentence, its information structure (TFA) and the basic conference relations. The output of the programme is visualized in order to get a more perspicuous picture of the relevant aspects of discourse structure.

7 Acknowledgement

The work reported on in this paper has been carried out under the projects MSM0021620838 and GAUK 350/2005/A-FG/MFF.

References

[1] Fligelstone Steve 1992. Developing a Scheme for Annotating Text to Show Anaphoric Relations. In G. Leitner (ed.), *New Directions in Corpus Linguistics.*. Berlin: Mouton de Gruyter, pp. 153–170.

[2] Gardent Claire, Manuélian Hélne, and Eric Kow. 2003. Which bridges for bridging definite description? In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*. Budapest, Hungary.

[3] Hahn Strube. 1997. Centering-in-the-large: Computing referential discourse segments. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (EACL'97)*. Madrid, Spain.

[4] Hajičová Eva. 1993. *Issues of sentence structure and discourse patterns*. Karolinum-Charles University Press, Prague, Czech Republic.

[5] Hajičová Eva and Jarka Vrbová. 1982. On the role of the hierarchy of activation in the process of natural language understanding. In *Proceedings of the COLING'82*, pp. 107–113.

[6] Hajičová Eva, Jiří Havelka and Kateřina Veselá. 2005. Corpus evidence of contextual boundness and focus. *Prague Bulletin of Mathematical Linguists*, 12(3):pp. 456–789.

[7] Kučová Lucie et al. 2003. Anotování koreference v Pražském závislostním korpusu. *ÚFAL Technical Report 19*, Prague, Czech Republic.

[8] Mitkov Ruslan et al. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium* (DAARC'2000)), Lancaster, UK.

[9] Murrell Paul. 2005. *R Graphics*. Chapman & Hall/CRC.

[10] Passoneau Rebecca and Diane Litman. 1997. Discourse segmentation by human and automated means. In *Computational linguistics* 23 (1), 3139.

[11] PDT 2.0. http://ufal.mff.cuni.cz/pdt2.0 2006. *Prague Dependency Treebank, 2nd edition*.

[12] Poesio Massimo. 2003. Associative Descriptions and Salience: A Preliminary Investigation. In *Proceedings of the 10th ACL Workshop on Anaphora (EACL'03)*. Budapest, Hungary.

[13] Poesio Massimo, Uryupina Olga, Vieira Renata, Alexandrov-Kabadjov, Mijail, and Rodrigo Goulart. 2004. Discourse-new detectors for definite description resolution: A survey and a preliminary proposal. In *Proceedings of the ACL Workshop on Reference Resolution (EACL'04)*. Barcelona, Spain.

[14] Sgall Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht:Reidel / Prague:Academia.

[15] Sidner Candace. 1983. Focusing in the comprehension of definite anaphora. In *Computational models of discourse*. Cambridge, Massachussets: MIT Press. 67330.

[16] TrEd. http://ufal.mff.cuni.cz/ pajas/tred. 2000-2005.

[17] Tutin Agnes, Trouilleux Francois, Clouzot Catherine, Gaussier Eric, Zaenen Annie, Rayot Stephanie, and Georges Antoniadis. 2000. Annotating a large corpus with anaphoric links. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC'2000)*. Lancaster, UK, 2838.

[18] Vieira Renata and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. In *Computational Linguistics*, 26 (4), 525579.