On Scalarity in Information Structure

Eva Hajičová

1.Introduction

The aim of the present paper is to characterize and compare different scales discussed in the relevant linguistic literature that are related in some way or another to the information structure of the sentence, be they more sentence-oriented (Sect. 2) or discourse-oriented (Sect. 3) and to introduce a scale based on a partial ordering of mental images in the stock of knowledge assumed by the speaker to be shared by him and the hearer during the discourse (Sect. 4). This hierarchy has a cognitive background but is reflected in the structure of the sentence, be it its information structure, or the types of referring expressions (such as e.g. definite and indefinite nouns in languages that have explicit forms for this feature, or pronouns vs. full nouns or nominal groups) or in sentence prosody.

2. From a dichotomy to a scale

The original formulations of what is now more generally referred to as the information strucure of the sentence were based on a dichotomy, be it a distinction between psychological subject and psychological predicate, Theme-Rheme, Topic-Comment, Topic-Focus, Presupposition and Focus, Given and New information etc. In structural linguistics, the pioneering (as well as systematic) attention to these issues was paid by Vilém Mathesius, who refers to H. Weil (1844; Mathesius mistakenly referes to him by the date 1855), and to linguists around Zeitschrift für Völkerpsychologie, also Georg v.d. Gabelentz (1868), Hermann Paul (1886), and esp. Ph. Wegener (1885), though criticizing their terms "psychological subject" and "psychological predicate" (Mathesius 1907). Mathesius himself calls this articulation by the Czech term "aktuální členění" because it is determined (guided) by the "actual", that is "topical" situation of the speaker and concerns the way, in which the sentence is incorporated to the factual relation to the situation from which it originated (Mathesius 1939). Mathesius distinguishes "východiště výpovědi" (initial point of the utterance, its basis), which he specifies as "what is in the given situation known or at least is evident and from what the speaker starts" on the one hand and "jádro výpovědi" (nucleus of the utterance), that is "what the speaker utters about or with respect to the starting point of the utterance". Mathesius prefers the above specification rather than the reference to known and unknown. However, already in Mathesius writings a certain inclination to recognize a more articulated scale rather than a mere dichotomy can be traced, when he says that the starting point (basis) may contain more than a single element so that it is possible to speak about the center of the starting point and the accompanying elements (Cz. *jevy průvodni)* which „lead from the center to the nucleus". Referring to the position of the sentence predicate, Mathesius writes that „the predicate is a part of the nucleus but on its edge rather than in its center and represents a transition [*přechod*] between the two parts of the utterance". The notion of a scale is explicitly mentioned in Mathesius (1941) criticism of Trávníček's treatment of Czech word order, when he claims that the word order serves for distinguishing various degrees of importance [*závažnosti , důležitosti*] of the elements of the same sentence (Mathesius 1941).

Mathesius observations inspired the fundamental work of Jan Firbas and his team. As Mathesius' original Czech term *aktuální členění větné* is not directly translatable into English, Firbas – on the advice of Josef Vachek (Firbas 1992, p. xii) and apparently inspired by

Mathesius' (1929) use of the German term *Satzperspektive* – used the term *functional sentence perspective* (FSP in the sequel). Very early in the development of the FSP approach, the binary articulation into theme and rheme was complemented – also in lines with Mathesius ideas mentioned above – by a more structured approach introducing the notions of transition and even a more scalar notion of communicative dynamism (CD). From this point of view, theme was specified by Firbas (1964) as being constituted by an element or elements carrying the lowest degree(s) of CD within a sentence (which was later modified by Firbas 1992 in the sense that theme need not be implemented in every sentence, while in every sentence there must be rheme proper and transition proper.) The concept of communicative dynamism was characterized by Firbas (1971) as a hierarchy of degrees carried by a linguistic element of the sentence, i.e. „the extent to which the element contributes towards the development of communication". The basic ditribution of CD would then reflect what H. Weil called the „movement of the mind". It is interesting to note that Svoboda (2007) says that the degrees of CD can be viewed as degrees of communicative importance ("*sdělná závažnost*") from the point of view of the intention of the speaker.

In FSP, two other scales were introduced, namely those of dynamic semantic functions performed by context-independent elements: (1) the scale of presentation characterized by Setting – Presentation of Phenomenon – Phenomenon presented, and (2) the scale of quality characterized by Setting - Bearer of Quality - Quality– Specification and Further specification. According to Firbas (1992, 67) , these scales are arranged in accordance with a gradual rise in CD.

Almost in parallel with FSP, but partly also as a reaction to it, Petr Sgall and his collaborators in Prague, developed the theory of topic-focus articulation (TFA in the sequel, see e.g.Sgall 1967; Sgall et al. 1973; 1980; Sgall et al. 1986; Hajičová et al. 1998). The TFA theory is an integral part of the formal model of functional generative description of language, namely of the representation of sentences on the underlying (so-called tectogrammatical) sentence structure. These tectogrammatical representations are viewed as dependency trees, with the main verb being the root of the tree. Every node of the tree carries – in addition to other characteristics such as the type of dependency - an index of contextual boundness: a node can be either contextual bound or non-bound. This feature, however, does not necessarily mean that the entity is known from the previous context or new but rather how it is structured as for the information structure. Thus, for example, in the second sentence of the discourse segment *When we walked around the town, we met Paul and his wife. I have immediately recognized HIM, but not HER*. (capitals denoting the intonation center, if the sentences are pronouced), both Paul and his wife are mentioned in the previous context, but the sentence is structured as if they are a piece of non-identifiable information, i.e. marked as contextually non-bound. Contrary to that, the above segment from the information structure point of view can be structured also in a different way, which, in the surface form of the sentence in English, would involve a different placement of the intonation center: *When we walked around the town, we met Paul and his wife. I have immediately RECOGNIZED him*. In this segment, both Paul and his wife are also introduced in the first sentence, but in the second sentence, it is the event of recognizing which is structured as bringing ‚new' (non-identifiable) information, while Paul - being referred to by a non-stressed pronoun – is taken as contextually bound (identifiable). In Czech, the two situations would be expressed by different word order and different forms of the pronoun coresponding to English *him*, namely *jeho* vs. *ho*: *Hned jsem poznal JEHO* versus *Hned jsem ho POZNAL*.

With the help of the bound – non-bound primary opposition, the distinction between the topic

and the focus of the sentence can be made depending in the status of the main verb (i.e. the root) of the sentence: basically, if the verb is contextually bound then the verb and all the nodes depending (immediately or not) on the verb constitute the topic, the rest of the sentence belonging to its focus; if the verb is contextually non-bound, then the verb and all the nodes depending on it to the right constitute the focus, the rest of the sentence belonging to its topic (see the definition of topic and focus by Sgall 1979)..

The left-to-right dimension of the tree serves as the basis for the specification of the scale of communicative dynamism: communicative dynamism is specified as the deep word order, with the least dynamic element standing in the leftmost position and the most dynamic element (the focus proper of the sentence) is the rightmost element of the dependency tree.

3. Cognitively based approaches to hiearchization and scalarity in discourse

Before we pass over to our proposal relating syntactic and information structure of the sentence with a dynamic view of discourse development, we would like to give a brief account of other related approaches going in a similar direction.

Prince (1981) recognizes three levels of givenness (p. 225ff.) in speaker-hearer terms, namely (i) givenness in the sense of predictability/recoverability (referring to Kuno's 1972 old-new information distinction and Halliday's 1967 given-new information distinction but mentioning that the two notions are defined differently by the two authors), (ii) givenness in the sense of saliency (referring to Chafe's notion of given – new information: given information represents „that knowledge which the speaker assumes to be in consciousness of the addressee at the time of the utterance" and new information „what the speaker assumes he is introducing into the addressee's consciousness by what he says" (Chafe 1976, p. 30), and (iii) giveness in the sense of ‚shared knowledge' (referring back to Clark and Haviland's characterization of ‚given' as „information [the speaker] believes the listener already knows and accepts as true" and new as „information [the speaker] believes the listener does not yet know", Clark and Haviland 1977, p. 4). For clarity reasons, Prince uses the term ‚assumed familiarity" rather than the term shared knowledge or givenness and defines a familiarity scale on discourse entities as follows (p. 245):

Evoked Entities (be it situationally or textually, i.e. the entity is already in the discourse-model) > *Unused* (but assumed to be in the hearer's model) > *Non-containing inferrable* > *Containing inferrable* (i.e. what is inferenced of is properly contained within the inferrable NP itself) > *Brand-New Anchored* (i.e. the NP representing the entity is linked, by means of another NP contained in it to some other discourse entity) > *Brand-New Unanchored*.

Six implicationally related cognitive statuses relevant for explicating the use of referring expressions in discourse are proposed in Gundel et al. (1993) and a ‚givenness hieararchy' is postulated intended to contribute to the understanding of how nominal expressions succeed in picking out a speaker/writer's intended referent. The hierarchy is supposed to reflect the assumed cognitive (memory and attention) status of an intended referent/interpretation for the addressee at the point just before the nominal form is encountered; the statuses are in a relation of unidirectional entailment; i.e. they are not mutually exclusive – e.g. anything in focus is, by definition, also activated, etc. The hierarchy is shown below, together with the indication of the appropriate use of a different form or forms (N stands for noun or noun phrase).:

in focus > activated > familiar > uniquely identifiable > referential > type indentifiable

it	*this*	*that* N	*the* N	indef.	*a* N
	that/this N			*this* N	

The givenness hierarchy statuses are mental states (rather than linguistic entities); the linguistic forms that encode these statuses provide procedural information about how to access (a mental representation of) the referent.

Based on Sperber and Wilson‘ notion of accessibility, Ariel (1990) concentrates on the system of accessing NP antecedents but she makes a more general claim: she believes that all context retrievals are governed by Accessibility theory. In her interpretation, Accessibility is a graded notion; the scale is viewed as continuous. In principal, she accepts Givon’s (1983) scale in the syntactic coding of topic accessibility listed here from the most continuous/accessible topic to the most discontinuous/inacccessible topic (DEF NP stands for a definite noun phrase):

zero anaphora > unstressed/bound pronouns or grammatical agreements > stressed/independent pronouns > Right-dislocated DEF NP’s > neutral-ordered DEF/NPs > Left-dislocated DEF NPs > contrastive topicalization > cleft/focus constructions > referential indefinite NPs.

She also follows Givon’s view (most explicitly pronounced in Givon 1983, pp.17ff.) that the Accessibility Marking Scale cannot be taken as universal because it does not cover the full range of referring expressions in all languages, not even capturing all the possibilities in the language she concentrates on, namely English). Givon (1983, p.18) claims that „a better and typologically more relevant predictions can be made by recognizing a number of scales each reflecting some specific means – be those word-order, morphology, intonation or phonological size …“

Lambrecht (1994) speaks about ‚discourse register‘, specified as „the set of representations which a speaker and a hearer may be assumed to share in a given discourse“ (p. 74). In the discussion that follows, he mentions Chafe’s (1976) category of identifiability and bases his analysis on Chafe’s (1987) idea of three activation states: active, semi-active (or accessible) and inactive (p. 93) and relates them to their formal correlates in the structure of sentences. From the point of view of the mental effort necessary to process sentences containing the given topics (cf. also Chafe’s 1987 low cost effort and activation cost in Chafe 1994; cf. also the least processing effort as discussed by Sperber and Wilson 1986), Lambrecht (p. 165) distinguishes the following scale of relative acceptability of ‚topic referents’ (going from the most to the least acceptable, cf. Prince’s familiarity scale quoted above):

active – accessible – unused – brand-new anchored – brand-new unanchored

A slightly different but yet related is the model of the local attentional states of speakers and hearers as proposed by Grosz and Sidner (1986; 1998), which is the basis of the centering theory (Grosz et al. 1983; 1995; Walker et al. 1998). The centering mechanism is supposed to capture intuitions about the flow of discourse as presented by Chafe (1979) and other similar rather intuitive considerations about coherence of discourse. Each utterance in discourse is considered to contain a backward looking center which links it with the preceding utterance (called ‘topic’ and referring mostly to the subject of the sentence rather than being related to the term ‘topic’ in the sense of topic-focus articulation) and a set of entities called forward looking centers; these entities are ranked according to language-specific ranking principles (these principles are 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 transition from one utterance to the next following one is then specified by one of two basic rules, one of which captures the ordering of four possible transitions: 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 (the most likely link to the following 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: according to 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". There are two problems connected with the original formulation of the centering approach: first, it captured only local coherence of adjacent sentences rather than (global) discourse coherence, and, second, the ranking of centers is defined in terms of syntactic relations as specified for the surface shape of the sentence (subject, object etc.) and as such very language-dependent; it might be of advantage to look at the ranking in terms of semantic roles (such as Actor, Patient, Addressee etc.), see our discussion in Kruiff-Korbayová and Hajičová (1997).

Close to the above-mentioned models that understand the distribution of entities in locally coherent texts to exhibit certain regularities is the computationally oriented approach of Barzilay and Lapata (2008). They introduce the notion of an entity-grid and formulate an algorithm which automatically abstracts a text into a set of entity transition sequences and records distributional, syntactic and referential information about discourse entities. The entity grid is a two-dimensional array that captures the distribution of discourse entities across text sentences. The rows of the grid correspond to sentences, the columns correspond to discourse entities (if an entity appears in different linguistic forms that are related in some relation of coreference, it is mapped to a single entity in the grid). The grid cells (boxes) are marked by the indication of syntactic roles of the given entity in the given sentence (subject, object, or neither; the authors consider also the possibility to mark thematic rather than syntactic roles). In addition to coreference links and syntactic relations, they take into account also the parameter of salience: salient entities are indentified on the basis of their frequency. In this respect, they refer to Givon (1987) and Ariel (1988), who assume that frequency of occurrence correlates with discourse prominence. In our approach descibed below, we argue that the notion of salience should be undesrtood in a more complex way and that neither frequency nor the length of the referential chain is a sufficient measure of salience.

4. (Some) heuristics guiding the development of the activation degrees

The studies of the connectivity of text in terms of the relations of topics anf foci of adjacent sentences go actually as far back as to H. Weil's observations about two types of possible "transitions of thoughts", namely a parallel and a progressive march: "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" (Weil 1978, p. 41). This idea is later reflected in Czech linguistics in Daneš' notion of thematic progressions (Daneš 1968; 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. In a slightly different but closely related vein, Firbas develops his ideas of the thematic and rhematic layers of a text (1995).

Our analysis of the development of discourse and its description (the first formulations of the approach in Hajičová and Vrbová 1982 were further developed e.g. by Hajičová 1993; 2003; Hajičová and Hladká 2008) is based on the following underlying hypothesis:

Hypothesis: A finite mechanism exists that enables the addressee to identify the referents on the basis of a partial ordering of the elements in the stock of knowledge shared by the speaker and the addressees (according to the speaker's assumption), based on the degrees of activation (salience) of referents.

The following three basic heuristics (a) through (c) based on the position of the items in question in the topic or in the focus of the sentence (see above in Sect. 2 on the TFA approach), on the means of expression (noun, pronoun) and on the previous state of the activation can be formulated to determine the degrees of salience of the elements of the stock of shared knowledge:

(a) In the flow of communication, a discourse referent enters the discourse, in the prototypical case, first as contextually non-bound, thus getting a high degree salience. A further occurrence of the referent is prototypically contextually bound, the item still has a relatively high degree of salience, but lower than an element referred to in the focus (as *nb*) in the given sentence. **I**t should be added in this connection that the stock of knowledge of the speaker/hearer contains also some permanently salient referents such as: *here, now, Europe, Shakespeare* etc. which can stand in the topic part of the sentence without having been mentioned in the previous co-text.

(b) If an item is not referred to in the given sentence, the degree of salience is lowered; the fading is slower with a referent that had in the previous co-text occurred as contextually bound; this heuristics is based on the assumption that a *cb* item has been 'standing in the foreground' for some time (as a rule, it was introduced in the focus, then used as contextually bound, maybe even several times) and thus its salience is reinforced; it disappears from the set of the highly activated elements of the stock of shared knowledge in a slower pace that an item which has been introduced in the focus but then dropped out, not rementioned (cf. Chafe's 1994 question how long a referent, once having aquired a ‚given' status will retain it). If the referent has faded too far away it has to be re-introduced in the focus of the sentence.

(c) If the difference in the degree of salience of two or more items is very small, then the identification of reference can be done only on the basis of inferencing.

The mentioned three basic heuristics served as a basis for our formulation of several rules for the assignment of the degrees of salience, which we have applied to numerous text segments to check how the determination of these degrees may help reference assignment.

The following basic rules determining the degrees of salience (in a preliminary formulation and taking into account only nominal referents) have been designed, with $dg_i(r)$ indicating the salience degree of the referent r after the *i*-th sentence S_i of a document is uttered:

(i) if r is expressed by a weak pronoun (or zero, i.e. deleted in the surface shape) in a sentence, it retains its salience degree after this sentence is uttered: $dg_i(r) := dg_{i-1}(r)$;
(ii) if r is expressed by a noun (group) carrying nb, then $dg_i(r) = 0$;
(iii) if r is expressed by a noun (group) carrying cb, then $dg_i(r) = 1$;
(iv) $dg_i(q) := dg_i(r) +2$ obtains for every referent q that is not itself referred to in S_i, but is immediately associated with an item present here;
(v) if r neither is included in S_i, nor refers to an associated object, and has been mentioned in the focus of the preceding sentence, then $dg_i(r) := dg_{i-1}(r) +2$.
(vi) if r neither is included in S_i, nor refers to an associated object, and has been mentioned in the topic of the preceding sentence, then $dg_i(r) := dg_{i-1}(r) +1$.

Since the only fixed point is that of maximal salience, our rules technically determine the degree of salience reduction (indicating 0 as the maximal salience). Whenever an entity has a salience distinctly higher than all competing entities which can be referred to by the given expression, this expression may be used as giving the addressee a sufficiently clear indication of the reference specification. It should be emphasized that what matters is the relation higher/lower degree, rather than the absolute numerical value; also the difference of 1 is too small to be relevant (see (c) above).

4. Analysis of a piece of text

One of the analyzed texts illustrating the development of salience degrees during a discourse, as far as determined by such rules, was a sample of literary text (of about 60 sentences) taken from the English translation of Josef Škvorecký's Czech book „Scherzo capriccioso" (published in 1984 by Sixty-Eight Publishers, Toronto; translated by Paul Wilson as „Dvorak in Love", Toronto: Lester and Orpen Dennys Ltd., 1986; the segment analyzed here is on pp. 251-253 and quoted in the Appendix). The episode depicts the author's imagination of the inspiration for Dvořák's opera „Rusalka" (Water Nymph). This analysis (Hajičová 2003b) can be used also for a cross-linguistic comparison, namely that of the English translation and the Czech original text. Therefore, we have applied the same rules to assign the salience degrees to the same referents in the Czech original text, and the result was rather convincing: the development of activation in both texts followed the same lines, with the exception of three points: (i) the sentence boundaries differ (e.g. En.. 53 and 54 is in Cz. ‚merged' into a single sentence), (ii) the grammatically fixed English word order forced the translator to use a different formulation, in order to preserve the TFA of the original Czech sentences (in order to place ‚the two figures' in 1 in the position of focus in En., a pronominal subject ‚they' had to be inserted, while in Cz. the object was placed into focus just by a change of word order; a similar structural change was necessary in 35), and (iii) a possible case of a translator's mistake (sent. 34). In this connection, one can go back to Weil (1844), who remarks: „… in translating from one language to another, if it is not posssible to imitate at the same time the syntax of the original and the order of the words, retain the order of the words and disregard the grammatical relations" (p. 26 of the En. translation in Weil 1978).).

For the purspose of our analysis we selected eight ‚objects' (better to say: mental images of objects): Dvořák's daughter Magda (M), her lover Kovarik (K), the lady (L, referred to also as

Rusalka, waterlilly), the black man (A, referred to also as *banjo, baritone voice*), the buggy (B, also *horses, figures*), the rowboat (R), torch (T) and Dvořák (D). Each of these objects was followed from the point of view of the development of its activation and was assigned numerical values of the reduction of salience to noun/pronoun tokens (or 'zero' for superficially deleted elements, see below) according to the rules quoted above. We do not take into account the rule (iv) for associated items (thus e.g. *banquet* in sentence 12 is associated with L and A); however, we understand the expressions *strawhat, a pair of white shoes*, and *a crumpled white pile* in 13 through 15 to be in such a close association with L that we take them as to directly refer to her and assign the activation degrees accordingly); the same holds true for such a more or less direct reference as is the case of *banjo, baritone voice* which we subsume under A It should be added that we have in mind underlying representations of the sentences in which elements deleted in the surface shape of the sentence are reconstructed, so that we assign a value to the 'zero' in 9 (as if there were the pronoun *they* referring to M and K), in 17, 21, 32 and 33 (as if there were the pronoun *she* referring to the Lady), in 25 and 45 (as if there were the pronoun *he* referring to Kovarik), in 54 (as if there were the pronoun *she* referring to M) and in 57 and 58 (as if there were the pronoun *he* referring to Dvořák).

5. Discussion of implications of the analysis and related problems

As we have pointed out in our previous papers, such an analysis of discourse makes it possible to throw some light on several issues of discourse structure:

(a) Certain patterning can be readily observed: e.g. a more or less regular change of groupings of items on the "top of the stock" if the discourse fluently passes from one group of items talked about to another group, or a cluster of items staying on the top with other items just entering the stage and leaving it very quickly.

(b) The proposed representation of the flow of discourse offers one way of segmenting the discourse more or less distinctly into smaller units according to which items are the most activated ones in these stretches. In our sample, the segmenting can be illustrated as follows: the first segment is characterized by the items Magda, Kovarik, and partly also the lady; in the next segment, from sent. 13, the lady dominates the top of the stock, up to sent. 23 through 25; from this point, Magda and Kovarik "return" on the stage, and in the last segment, from sent. 41, the Master appears on the scene.

(c) An interesting issue for further investigation is that of the identification of possible thresholds (i.e. the placement of vertical lines in the flow charts): thus one can investigate whether and under which conditions such a threshold adds to other prerequisites for the possibility/impossibility of pronominal reference, for the use of a full definite NP in topic, for the necessity to use stronger means for a reintroduction of some already mentioned item, and for the necessity of such a reintroduction to occur in the focus. The presence of 'competitors', of course, is highly relevant for such investigations.

(d) Last but not least, the proposed representation of the flow of discourse can serve as a basis for the identification of 'topics' of the discourse. It is often disputable to determine 'the' topic of a given discourse; however, the discourse topic(s) occur (or at least the items associated with these topic(s), whatever the notion of association may be understood to stand for) most probably among the items staying longer (or more frequently) among the most activated items, i.e. on the top of the stock. A look at our table confirms what apparently the author

wanted the reader to understand: the episode is about Rusalka, i.e. the lady; her activation does not go beyond the degree 7, while all other items cross this point at least once.

Let us turn now in more detail to the relationships between the choice of coreferring expressions (weak pronoun, noun, noun group with simple or complex adjuncts) and the degrees of salience, as illustrated by our sample text. In the first sentence of our segment, the pronoun *they* refers to Dvořák's small daughter Magda and to his guest Kovarik, who take a walk in the surroundings of Dvořák's house in Iowa (during his stay in the US). The following observations can be made:

(i) A weak (or zero) pronoun refers to a highly salient item (see sent. 9, 11, 34); this pronoun expresses a contextually bound item; the use of a strong (long) pronoun is limited to cases when the reference is made to a contextually non-bound item (in the focus of the sentence) or to a contrastive item in the topic (expressed by an (optional) phrasal stress; in the latest version of our analysis of the topic-focus articulation, this item is marked by a special superscript in the underlying representation of the sentence, see Hajičová et al., 1998).

(ii) If two items are close to each other in their degrees of salience, the use of a weak pronoun is limited to cases (a) with relevant grammatical oppositions (gender, number) and (b) with a clear pragmatic basis for inferencing (see e.g. 'they' in 2 and 'them' in 4).

(iii) Otherwise, the coreference to one of the competitors is to be made clear by the use of a noun (see ‘the beauty’ in 16), or even of a noun group with simple or complex adjuncts (see ‘the girl across the river’ in 30).

(iv) Such stronger means have to be used also if the salience has faded away (see ‘the man’, ‘Kovarik’, and ‘the child’ in 19, 24, 26, respectively); if the salience goes beyond a certain threshold of comprehensiveness, the item needs to be reintroduced by a reference in Focus. In our sample, this is illustrated by the different kinds of means referring to the black man: its activation has dropped significantly and thus in both 19, 35 and 46 the reference to him is made by a referring expression in Focus.

This example should make it possible for the reader to check (at least in certain aspects) the general function of the procedure we use, as well as the degree of its empirical adequacy in the points it covers. We are aware of the still tentative character of our analysis, which may and should be enriched in several respects (not to cover only noun groups, to account for possible episodic text segments, for oral speech with the sentence prosody, etc.).

6. Perspectives: integrated annotation of text corpora

In spite of the fact that the mechanism proposed above has been tested on a couple of texts, both from Czech, English and even in Bulgarian and Malayan, a much broader range of texts analyzed as for all the three aspects, namely the underlying sentences structure, the assignment of topic-focus articulation and the coreference relations would be necessary for a deeper evaluation of the analysis. Manual annotation is a very time-consuming and costly task; an opportunity to work with larger data is now offered by the existence of computationally available annotated text corpora.

For Czech, we are now working on a project based on the data available in the Prague

Dependency Treebank (PDT; Hajič et al. 2006; Mikulová et al. 2006). PDT is a corpus of Czech texts comprising 3165 documents (text segments mainly of a journalistic genre) annotated on all levels. The documents consist of 49431 sentences which cover 833195 occurrences of tokens (word forms and punctuation marks) annotated on three levels: (a) morphemic (with detailed part-of-speech tags and rich information on morphological categories), (b) surface shape ("analytical", in the form of dependency-based tree structures with relations labeled by superficial syntactic functions such as Subject, Object, Adverbial, Attribute, etc.), and (c) underlying dependency-based syntactic level (so-called tectogrammatical) with dependency tree structures labeled by functions such as Actor, Patient, Addressee, etc. and including also information on the topic-focus articulation of sentences. In addition, two kinds of information are being added in the latest version of PDT, namely annotation of discourse relations (Mladová et al. 2008) based on the analysis of discourse connectors (inspired by the Pennsylvania Discourse Treebank, see Prasad et al. 2008a, 2008b; Miltasaki et al. 2008) and information on grammatical and on textual intra- and extra-sentential coreference relations.

Thanks to such a richly annotated corpus, we basically have (or will soon have) at our disposal all information we need for an application of our rules for activation assignment: the underlying sentence representation with restored (superficial) deletions as well as with part-of-speech information, the Topic-Focus assignment (via the TFA attribute with values contextually-bound and contextually non-bound) and coreferential chains for nominal and pronominal realization of referential expressions. The task we face now is to implement the activation algorithm on (selected but full) documents, to visualize the 'activation' diagram, and, most importantly, to evaluate the results to see whether we achieve what we have envisaged.

Czech original

1. Na druhém břehu se objevil oheň a u něho dvě postavy.
2. Když přišli blíž,
3. zabělala se na pozadí temných keřů dvě bílá koňská těla.
4. Potom jej poznal.
5. Bledě modrý kočárek.
6. – 7.Před dvěma hodinami v něm seděla kráska z Chicaga a černoch v livreji jí šel pro pivo ke Kapinosům.
8. Zastavili se …
9. A hleděli na druhý břeh.
10. Mladá dáma v bílých šatech okusovala kuřecí stehýnko. …
11. Pohlédl na Magdu.
12. Dětské oči, vyvalené úžasem, zíraly k druhému břehu, kde se konala hostina z pohádky.…
13. Pohlédl na slamák.
14. Ano, vedle něho se v trávě povalovaly překocené bílé střevíčky a
15. zmuchlaná bílá hromádka. …
16. Kráska vvstala a
17. odhodila napůl okousané stehýnko do ohně.
18. Protáhla se,
19. řekla něco černochovi. …
20. Kráska si vyhrnula sukně
21. a vysoko našlapujíc v trávě,

22. vydala se proti proudu říčky.
23. Hlava se jí proměnila v chladně zářící pochodeň.
24. Omámeně vykročil
25. a jal se tiše sledovat pouť krásného přeludu. …
26. Dítě ťapkalo mlčky vzadu. …
27. Dítě zašeptalo:
28. „Vona je rusálka!“ …
29. Zatajil se mu dech.
30. Dívka na druhém břehu řeky si rozepjala živůtek
31. a … zvedla sukně nad hlavu
32. a soukala se z nich.
33. Za okamžik zůstala jen v bělostných kalhotkách po kolona a v letním polokorzetu.…
34. (Spatřil, co v životě jestě neviděl.) …
35. Zdola, po proudu řeky, zaznělo banjo.
36. Příjemný baryton zpíval: …
37. Kráska spustila ruce …
38. Přistoupila k vodě …
39. Za nízkým křovím u jejich břehu cosi zapraskalo.
40. Pohlédl tam a teprve nyní rozeznal mezi listím obrysy loďky.
41 V ní se právě vztyčoval nějaký člověk
42 a na hlavu mu dopadlo měsíční světlo. Divoký vous, zježené vlasy.
43. Mistr! …
44. Rychle pohlédl přes řeku
45. a spatřil rusálku už po pás ve vodě. „Borne like a vapor …“
46. Mistrova hlava v profilu natočeném směrem k sametovému barytonu.
47. Nevidí,
48. jenom slyší,
49 napadlo ho.
50. Sám viděl.
51. Rusálka se pomalu nořila do vody …
52. Konečně zůstal na hladině jen hořící leknín.
53. -54. A do toho dítě zaječelo:
55. „Tati! …
56. Mistr sebou škubl,
57. rozhlédl se,
58 teprve teď spatřil.

Acknowledgement

This paper was written under the support of the grant of the Czech Republic Grant Agency P406/12/0658 and has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References:

ARIEL, M. (1988): Referring and accessibility. Journal of Linguistics 24, pp. 65-87.

AEIEL, M. (1990): Accessing Noun-Phrase Antecedents. Routledge, London.

BARZILAY, E. and M. Lapata (2008): Modeling local coherence: An entity/based approach. Computational Linguistics 34, pp. 1-34.

CHAFE, W. (1976): Givenness, contrastiveness, definiteness, subjects, topics, and point of view, In Li., ed. (1976), pp. 25-55

CHAFE, W. (1979): The flow of thought and the flow of language. In: Givon, T. ed., Syntax and Semantics: Discourse and Syntax, Vol. 12. Academic Press, New York, pp. 159-182.

CHAFE, W. (1987): Cognitive constraints on information flow, In: Tomlin, R. ed. pp. 21-52

CHAFE, W. (1994): Discourse, Consciousness, and Time. (The Flow and Displacement of Conscious Experience in Speaking and Writing.) The University of Chicago Press, Chicago/London

CLARK, H. H. and S. E. Haviland (1977): Comprehension and the given-new contract. In: R. Freedle (ed.), Discourse Production and Comprehension. Ablex, New Jersey, pp. 1-40.

DANEŠ, F. (1968): Typy tematických posloupností v textu (Types of thematic progressions in text), Slovo a slovesnost 29, pp. 125-141.

DANEŠ, F. (1974): Functional Sentence Perspective and the organization of the text, in Daneš F.ed., Papers on FSP, Academia, Prague, pp.106-128.

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

FIRBAS, J. (1964): On defining the theme in functional sentence perspective, Travaux Linguistique de Prague 1, Academia, Prague, pp. 267-280.

FIRBAS, J. (1992): Functional sentence perspective in written and spoken communication. Cambridge University Press,London: Cambridge.

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.

GABELENTZ, G. von der (1868): Ideen zu einer vergleichenden Syntax – Wort- und Satzstellung. Zeitschrift fűr Vőlkerpsychologie und Sprachwissenschaft 6, pp. 376-384.

GIVON, T. (1983): Topic continuity in discourse: An Introduction. In: Givon, ed. (1983), pp. 1-41.

GIVON, T. ed. (1983): Topic Continuity in Discourse: A Quantitative Cross-Language Study. John Benjamins, Amsterdam.

GIVON, T (1987): Beyond forground and background. In Tomlin R., ed., pp. 175-188.

GROSZ, B., JOSHI, A. K. and S. WEINSTE|IN (1983): Providing a unified account of

definite noun phrases in discourse. In: Proceedings of the Annual Meeting of the Association for Computational linguistics 21, pp. 44-50.

GROSZ, B., JOSHI, A. K. and S. WEINSTEIN (1995): Centering: A Framework for modeling the local coherence of discourse. Computational Linguistics 21, pp. 203-225.

GROSZ, B. J. and C. L. SIDNER (1986): Attention, intentions, and the structure of discourse. Computational Linguistics 12, pp. 175-204.

GROSZ, B. J. and C. L. SIDNER (1998): Lost intuitions and forgotten intentions. In: Walker et al., eds. 1998, pp. 39-51.

GUNDEL, J. K., HEDBERG, N. and R. ZACHARSKI (1993): Cognitive status and the form of referring expressions in discourse, Language 69, pp. 274-307.

HAJIČ, J. et al. (2006): Prague Dependency Treebank 2.0. CD-ROM. Linguistic Data Consortium, Philadelphia, PA, USA. LDC Catalog No. LDC2006T01 URL<http://ufal.mff.cuni.c z/pdt2.0/>

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, 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., PARTEE, B. and P. SGALL (1998): Topic-Focus Articulation, Tripartite Structures and Semantic Content, Kluwer Academic Publishers, Dordrecht.

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. (1982), Coling 82 – Proceedings of the Ninth International Congress of Computational Linguistics, John Benjamins, Amsterdam

KRUIFF-KORBAYOVÁ, I. and E. HAJIČOVÁ (1997): Topics and Centers, A Comparison of the Salience-Based Approach and the Centering Theory, Prague Bulletin of Mathematical Linguistics 67, pp. 25-50.

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

LAMBRECHT, K. (1994): Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents. Cambridge University Press, Cambridge, UK.

LI, Ch. (ed.) (1976): Subject and Topic. Academic Press, New York.

MATHESIUS, V. (1907): Studie k dějinám anglického slovosledu [Studies on the

development of English word-order], Věstník České Akademie XIV, pp. 261ff.

MATHESIUS, V. (1929): Zur Satzperspektive im modernen Englisch. Archiv für das Studium der neueren Sprachen und Literaturen 155, pp. 202-210.

MATHESIUS, V. (1939): O tak zvaném aktuálním členění větném. Slovo a slovesnost 5, pp. 171-174; translated as On information-bearing structure of the sentence; in: S. Kuno (ed.): Harvard Studies in Syntax and Semantics, 1975, pp. 467-480.

MATHESIUS, V. (1941): Základní funkce pořádku slov v češtině [Basic functions of the Czech word order], Slovo a slovesnost 7, pp. 169-180.

MIKULOVÁ M. et al. (2006): Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Tech. Report 30 ÚFAL MFF UK. Prague.

MILTSAKAKI, E., ROBALDO, L., LEE, A. and A. JOSHI (2008): Sense Annotation in the Penn Discourse Treebank. In: A. Gelbukh, ed., Computational Linguistics and Intelligent Text Processing, Springer, Berlin/Heidelberg, pp. 275–286.

MLADOVÁ, L., ZIKÁNOVÁ, Š. and E. HAJIČOVÁ, E. (2008): From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008*)*. Marrakech, Morocco: European Language Resources Association, pp. 1--7.

PAUL, H. (1886): Prinzipien der Sprachgeschichte. 2nd edition. Freiburg i.B.

PRASAD, R. et al. (2008a): Penn Discourse Treebank Version 2.0. Philadelphia: Linguistic Data Consortium.

PRASAD, R. et al. (2008b): The Penn Discourse Treebank 2.0. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

PRINCE, E. (1981): Toward a taxonomy of given/new information. In P. Cole, ed. Radical Pragmatics, Academic Press, New York, pp. 223-254.

SGALL, P. (1967): Functional Sentence Perspective in a generative description of language. Prague Studies in Mathematical Linguistics 2, Academia, Prague, pp. 203-225.

SGALL, P. (1979): Towards a definition of Focus and Topic. Prague Bulletin of Mathematical Linguistics 31, 3-25; 32, 1980, pp. 24-32; reprinted in Prague Studies in Mathematical Linguistics 78, 1981, pp. 173-198.

SGALL, P., HAJIČOVÁ, E. and E. BENEŠOVÁ (1973): Topic, Focus, and Generative Semantics. Skriptor, Kronberg/Taunus.

SGALL, P., HAJIČOVÁ, E. and E. BURÁŇOVÁ (1980): Aktuální členění v češtině [Topic-Focus Articulation of Czech Sentences)]. Academia, Prague.

SGALL, P., HAJIČOVÁ, E. and J. PANEVOVÁ (1986): The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Academia, Prague and Reidel, Dordrecht..

SPERBER, D. and D. WILSON (1986): Relevance: Communication and Cognition. Harvard

University Press Cambridge, MA: and Blackwell, Oxford.

SVOBODA, A. (2007): Brněnská škola funkční větné perspektivy v pojmech a příkladech. [Brno school of functional sentence perspective in notions and examples]. Ostrava University, Ostrava.

TOMLIN, R. ed. (1987): Coherence and Grounding in Discourse. John Benjamins, Amsterdam.

WALKER, M. A., JOSHI, A. and E. PRINCE, eds. (1998): Centering theory in discourse. Clarendon, Oxford..

WALKER, M. A., JOSHI, A. and E. PRINCE (1998), Centering in naturally occurring discourse: An overview. In: Walker et al. eds. (1998), pp. 1-28.

WEGENER, P. (1885): Untersuchungen űber die Grundfragen des Sprachlebens. Halle/S.

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

WEIL, H. (1978) : The Order of Words in the Ancient Languages Compared with That of the Modern Languages, Boston, 1887, reedited Amsterdam 1978; English translation of Weil (1844).

Appendix

1. Across the river they could now see a fire with two figures beside it.
2. When they moved closer,
3. they could make out two white horses against the background of the dark bushes.
4. Then he recognized them.
5. The pale blue buggy.
6. Two hours ago, the beauty from Chicago had sat on the seat.
7. While the black man in livery had gone into Kapino's for beer.
8. They stopped …
9. and looked across the river.
10. The young lady in the white dress was biting into a chicken leg. …
11. He looked at Magda.
12. The child's eyes, wide in amazement, stared across the river at this fairytale banquet. …
13. He looked at the straw hat.
14. Yes, beside it in the grass a pair of white shoes had been casually tossed
15. and beside them lay a crumpled white pile. …
16. The beauty stood up
17. and threw the half-eaten leg into the fire.
18. She stretched,
19. She said something to the man …
20. She lifted up her skirts
21. and, stepping gingerly through the grass,
22. she began walking upstream.
23. her head became a cooly glowing torch.

24. Intoxitated, Kovarik stepped forward
25. and silently followed the beautiful phantom's pilgrimage.
…
26. The child padded silently behind him. …
27. The child whispered.
28. „She's a Rusalka! A water nymph!"
29. He caught his breath.
30. The girl across the river unlaced her bodice
31. and … she had lifted the skirt over her head,
32. slipped out of it
33. and stood there in nothing but white knee-length knickers …
34. He couldn't také his eyes off her. …
35. From downstream they could hear a banjo playing.
36. A pleasant baritone voice sang: „…"
37. The girl let her hands drop…
38 Cautiously, she stepped into the water.
39. On their side of the river, …, something creaked.
40. Looking towards the sound, he could barely distinguish the outline of a small rowboat
41. and, in it, someone's dark silhuette.
42. The moonlight fell on the head, the white whiskers, the hair in disarray.
43. The Master!
44. He looked quickly across the stream
45. and saw the Rusalka up to her waist in the water. „Borne like a vapour …"
46. The Master's head turned in profile towards the velvet baritone.
47. He doesn't see;
48. he only hears,
49. he thought.
50. He himself saw. …
51. The Rusalka was slowly lowering herself into the water, …
52. Finally, all that remained on the water was a burning waterlilly.
53. Suddenly the child saw too
54. and shrieked,
55. „Papa!"
56. The Master started,
57. looked around
58. and then saw.