

Hierarchy of Saliency and Discourse Analysis and Production

Eva Hajičová, Petr Kuboň and Vladislav Kuboň

KAM - linguistics

Charles University

Malostranské nám.25

CS-118 00 Prague

Abstract

The hierarchy of saliency of the items of the knowledge assumed by the speaker to be shared by him and by the hearer constitutes one aspect of a dynamic account of discourse (Sect. 1). It is claimed that a representation of this hierarchy is a good support for discourse analysis (reference assignment, Sect. 2) and for discourse production (pronominalization, definite description, Sect. 3).

1.1 In studying communication, it must be distinguished between the speaker's own image, the hearer's image and the assumptions the speaker has made about the hearer's image of the world. In the very process of discourse, the image of the world undergoes changes of different kinds: new objects, relations, etc. are added to the repertoires on the basis of the content of what has just been said, the universe of discourse may be restricted in that a certain of phenomena of a particular kind is marked as relevant for further discourse whereas the other elements are disregarded, or the saliency (activation, foregrounding) of the items is changed in the sense of being easily accessible in memory (see Sgall, Hajičová and Pánevová, 1986, p. 54f). It has been shown (Hajičová and Vrbová, 1982; Hajičová, 1987) that the changes of saliency are dependent to a great extent on the topic/focus articulation of the utterance. As a matter of fact, most algorithms for anaphora resolution work with the notion of saliency (cf., e.g., Hobbs, 1976; Sidner, 1979; Brennan, Friedman and Pollard, 1989); however, while in most of these approaches the degrees of saliency are given only syntactically, the hierarchy of activation in our model is also determined by the topic/focus

articulation of the sentence.

Leaving aside the distinction between the contextually bound and non-bound elements of the utterance within its topic and focus parts (for the relevance of contextual boundness in this respect, see Hajičová, Hoskovec and Sgall, in press), these relationships can be summarized as follows:

(i) the items referred to in the focus of the utterance be it by a noun phrase or by a stressed pronoun receive the highest degree of saliency;

(ii) the items referred to by a noun phrase in the topic part of the utterance are activated one degree less than the items referred to in the focus part;

(iii) a pronominal reference to an item in the topic part of the utterance keeps the activation unchanged;

(iv) the activation of the items not mentioned in the given utterance fades away; the fading is steeper if the given item was the most activated item after the preceding utterance and less steep if the given item preserved high saliency for some of the previous utterances, being mentioned in its topic.

We do not attempt to cover VP-anaphora since in the model of the stock of knowledge assumed by the speaker to be shared by him and the hearer (SSK) we work - for the time being - only with mental images of objects, rather than with those of events.

1.2 Several thresholds can be established on the hierarchical structure of the activated part of SSK; at least two of them are important in the context of the present paper. One threshold characterizes those items of the SSK that are activated to such an extent that they can be referred to in the topic part of the following

utterance; this is to say that the salience of these items is large enough for the hearer to identify easily their referents. The second (higher) threshold delimits that part of SSK the items of which can be referred to by pronouns; their salience is assumed by the hearer to be large enough for the speaker to assign the reference in a straightforward way.

1.3 The representation of the discourse in terms of the hierarchy of activation of the elements of SSK suggests itself to be used for a split up of the discourse into segments; the segments correspond to those parts of the discourse for which there is a characteristic grouping of most activated items. These most activated items in each segment can then be regarded as the 'topic' of the given segment; items which may be understood as the 'topic(s)' of the discourse can then be computed on the base of the 'topic(s)' of the segments.

The ideas outlined in Sect. 2 and 3 will be illustrated by an analysis of multifarious examples; the results of those sections will serve as a theoretical base for further practical applications in various systems.

2.1 SSK can help to solve the referent assignment in discourse analysis. If we want to show a most suitable way of application of SSK in the context of other usual methods of solving this problem, we should remind first of all the assignment based on syntactic relations:

In Czech coordinated clauses the subject of the second (or third, fourth etc.) clause is usually deleted. Then it is (more or less) unambiguously understood to be the same as the subject of the first clause. The same holds for two successive sentences, e.g.:

"Digitalizace je velmi populární trend. (Ø) Stává se symbolem kvalitního záznamu pro posluchače."

"Digitalization is a very popular trend. It becomes the symbol of the recording quality for listeners."

There are two competitors in the first sentence, both NP's. The antecedent of the ellipsis in the second sentence is the subject of the first sentence.

The antecedents of relative pronouns are easy to compute, too. By a thorough investigation of a large amount of technical texts we found that relative pronouns almost certainly (about 90-95 %) refer to the head of the closest preceding NP which has appropriate morphematic categories (gender, number, etc.):

"Používáme disk z polykarbonátu, který jsme předem očistili."

"We use a disque of a polycarbonate which we've cleaned before."

An important role is played also by a tendency to keep the syntactic dependency hierarchy in referring - the antecedent of a pronoun in the subordinate clause is to be found on the higher or equal layer of the hierarchy:

"Program, užívaný systémem s rutinou, která mu pomáhá, může..."

"The program used by the system with the routine which helps it, can..."

Here, the pronoun "it" refers to "program" rather than to "system".

2.2 When working within the framework of the functional generative description (see Sgall, Hajičová and Panevová, 1986), the solution of anaphora can be supported by the topic-focus articulation and the hierarchy of activation of the items of the SSK.

A good help for finding the pronoun's antecedent is the form of the pronoun used in the text. The strong form of a pronoun (ten, tento = this; sebe = himself; ...) refers - in technical texts almost unambiguously - to the focus of the preceding sentence, the weak (unstressed) form implies referring preferably to the topic:

"Nejslabším článkem v celém řetězku je vstup. Ten je příčinou mnoha problémů."

"The poorest member in the whole chain is the input. This causes a lot of problems." /the strong form "this (ten)" refers to "input"/;

"Systém vyvolává rekursivní program. Můžete ho užít,..."

"The system calls a recursive program. You can use it,..."

/ "it" refers to "system" in primary case/;

These strategies are relatively reliable (80-85%) and can be used in discourse analysis.

2.3 When we take into account also other aspects of the role of SSK in discourse analysis, we can base the algorithm of reference assignment on the following strategies:

(1) if the subject of the sentence has a null form, the subject of the preceding clause is referred to, as long as the grammatical agreement is preserved;

(2) in case of a relative pronoun we try to find the head of the closest preceding noun phrase as the antecedent;

(3) if the referring expression is a weak pronoun, we look for the antecedent in the topic of the preceding clause, in case of a strong pronoun (or "adjective pronoun" in the noun phrase as "this man") we investigate the focus;

(4) if there are more competitors after step (3) or if none of the steps (1) through (3) can be used, we apply SSK in the form of a list of NP's from the preceding text (from the beginning of the actual paragraph) with their respective degrees of activity and choose the most activated item with the congruent morphological categories. If we cannot find an item "activated enough" (the concrete value, or the difference of values, is to be determined independently on the way the activation is evaluated), we prefer leaving the anaphora unresolved in order to prevent wrong solutions of the "global" references (to a preceding clause, sentence, an action identified by a verb, a coordination of items etc.) or references which cannot be solved without the use of semantics, e.g.:

"Přehrávače DAT jsou mnohem dražší než CD přehrávače. Toto nové zařízení ještě výrobci nebylo přijato."

"The DAT players are much more expensive than CD players. This new device is not yet accepted by producers."

The antecedent is not "DAT players", which can be computed only on the basis of factual knowledge - if you know DAT's are newer than CD's.

3. The discourse production has more freedom than analysis, because the speaker can choose the means while describing his ideas. Of course, he has to take care of the hearer to enable him to interpret the text easily and, if possible, unambiguously; at the same time, he should not repeat unnecessarily definite NP's. The main criterion in the speaker's choice between the use of a pronoun and a definite NP may be the actual state of SSK. We deal with technical texts only but we believe the basic ideas hold for other types of texts as well.

3.1 When producing a sentence of a continuous text, the speaker can use three types of referring expressions - weak pronouns, strong pronouns (including the demonstrative and relative ones) and more or less complex definite expressions (compare "John" with "the boy who played with a ball yesterday as I have told you..."). Depending on the actual state of SSK he chooses the relatively "weakest" means (from a weak pronoun to a complex description) the use of which enables the hearer to find the referent correctly. Two aspects of SSK are important in this choice:

(a) the degree of activation $/da(0)$ of the object (referent) in SSK - an important role is played by the minimal degree of activation (MIN), i.e., the threshold below which it is not possible to refer to objects by pronouns (see 1.2);

(b) the existence of "competitors" - i.e. objects differing in activation only by degree 1 (see Hajičová and Vrbová, 1982) and having the same morphological categories.

3.2 We claim that on the background of these two aspects we can find the following four cases involved in discourse production:

(i) $da(0) < \text{MIN}$ (as a special case this holds for "new objects"):

In a technical text, the speaker prefers to use a definite NP. The degree of its complexity depends on the presence of possible competitors.

"Vstupní data (0) se mění pomocí programu D-TYPE. Zpočátku vyvolává subrutinu D-START, která ukládá data (0) do paměti."

"The input data (0) are changed by the D-TYPE program. In the beginning it calls the D-START subroutine which loads the data (0) into the memory.";

- (ii) $da(0) > MIN$ and the object 0 has no competitor or the competitor is "far enough":

A weak pronoun can be used in this case.

"Vstupní data (0) se mění pomocí programu D-TYPE. Mění je (0) na speciální typ."

"The input data (0) are changed by the D-TYPE program. It transforms them (0) into a special type.";

- (iii) $da(0_1) > MIN$, the object 0_1 has a competitor 0_2 , none of them having the maximum degree of activation (MAX):

In this case a pronoun does not help. A definite NP (at least for one object) has to be used.

"Oba systémy (0_1) se liší v užitích (0_2). Rozdíl se projeví, pokud se pokusíme systém(0_1)/utility(0_2) odstartovat."

"Both systems (0_1) differ in utilities (0_2). The difference will take place if we try to start the systems(0_1)/utilities(0_2).";

- (iv) $da(0_1) = MAX$, $da(0_2) = MAX-1$ and 0_1 competes with 0_2 :

This is the most difficult situation. We can divide it into three subcases by the way referring expressions are used in the following clause (sentence) C:

- (a) the expression referring to 0_2 has the position of subject in C:

Here we face the "subject-preserving tendency", which is very common in continuous texts. This helps to avoid the possible ambiguity between competitors so that weak pronouns can refer to both objects ($0_1, 0_2$).

"Oba systémy (0_2) se liší v užitích (0_1). Both systems (0_2) differ in utilities (0_1).

(1) ... \emptyset (0_2) Nemají stejnou souborovou strukturu."

... They (0_2) haven't the same file structure."

(2) ... \emptyset (0_2) Používají je (0_1) k ..."

... They (0_2) use them (0_1) for...";

- (b) the expression referring to 0_1 has the position of subject in C:

When referring to 0_1 , a strong pronoun has to be used, in case of 0_2 a weak pronoun will do.

"Oba systémy (0_2) sdílejí některé soubory (0_1). Ty (0_1) jim (0_2) pomáhají k ..."

"Both systems (0_2) share some files (0_1). Those (0_1) help them (0_2) to ...";

- (c) cases (a),(b) do not hold:

In this situation the competition cannot be "solved" by syntactic means. The solution of the problem is the same as in (iii).

"Oba systémy (0_2) sdílejí některé soubory (0_1). Programátoři užívají systémy(0_2)/soubory(0_1) k ..."

"Both systems (0_2) share some files (0_1). The programmers use the systems(0_2)/files (0_1) to...".

3.3 As we have already stated, our study is the first step on the way to a complex account of the impact of SSK in discourse production. To handle the interplay between pronouns and definite NP's in all details, one has to state the relevant differences in the activation of competitors (in various types of sentences), to consider the possibility of the marked use of strong pronouns and definite NP's and many other problems.

References

- Brennan S.E., Friedman M.W. and Pollard C.J. (1989), A Centering Approach to Pronouns, manuscript
Carter D.M. (1985), Common Sense Inference in a Focus-guided Anaphor Resolver. In: Journal of Semantics 4, 237-246.

- Hajičová E. (1987), Focussing - A Meeting Point of Linguistics and Artificial Intelligence. In: Artificial Intelligence II, ed. by Ph.Jorrand and V.Sgurev
- Hajičová E., Hoskovec T. and Sgall P. (in press), Discourse Modelling Based on Hierarchy of Salience; to appear in Prague Studies in Mathematical Linguistics 11
- Hajičová E. and Vrbová J. (1982), On the Role of the Hierarchy of Activation in the Process of Natural Language Understanding. In: Coling 82, ed. by J.Horecký, Amsterdam:North Holland, 107-113
- Hobbs J.R. (1976), Pronoun Resolution. In: Techn.Rep. 76-1, Dept. of Computer Science, City College, CUNY
- Sgall P., Hajičová E. and Panevová J. (1986), The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Prague:Academia - Dordrecht:Reidel
- Sidner C.L. (1979), Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. In: TR-537, M.I.T Artificial Intelligence Laboratory
- Walker M.A. (1989), Evaluating Discourse Processing Algorithms, manuscript