

Conceptual and Linguistic Decisions in Generation

Laurence DANLOS
LADL (CNRS)
Université de Paris 7
2, Place Jussieu
75005 Paris, France

ABSTRACT

Generation of texts in natural language requires making conceptual and linguistic decisions. This paper shows first that these decisions involve the use of a discourse grammar, secondly that they are all dependent on one another but that there is *a priori* no reason to give priority to one decision rather than another. As a consequence, a generation algorithm must not be modularized in components that make these decisions in a fixed order.

1. Introduction

To express in natural language the information given in a semantic representation, at least two kinds of decisions have to be made: "conceptual decisions" and "linguistic decisions". Conceptual decisions are concerned with questions such as: in what order must the information appear in the text? which information must be expressed explicitly and what can be left implicit? Linguistic decisions deal with questions such as: which lexical items to choose? which syntactic constructions to choose? how to cut the text into paragraphs and sentences?

The purpose of this paper is to show that conceptual decisions and linguistic decisions cannot be made independently of one another, and therefore, that a generation system must be based on procedures that promote intimate interaction between conceptual and linguistic decisions. In particular, our claim is that a generation process cannot be modularized into a "conceptualizer" module making conceptual decisions regardless of any linguistic considerations, passing its output to a "dictionary" module which would figure out the lexical items to use accordingly, which would then in turn forward its results to a "grammar", where the appropriate syntactic constructions are chosen and then developed into sentences by a "syntactic component". In such generation systems (cf. (McDonald 1983) and (McKeown 1982)), it is assumed that the conceptualizer is language-free, i.e., need have no linguistic knowledge. This assumption is questionable, as we are going to show. Furthermore, in such modularized systems, the linguistic decisions must, clearly, be made so as to respect the conceptual ones. This consequence would be acceptable if the best lexical choices, i.e., the most precise, concise, evocative terms that can be chosen, always agree with the conceptual decisions. However, there exist cases in

which the best lexical choices and the conceptual decisions are in conflict.

To prove our theoretical points, we will take as an example the generation of situations involving a result causation, i.e., a new STATE which arises because of one (or several) prior ACTs (Schank 1975). An illustration of a result causation is given in the following semantic representation

```
(A) CRIME : ACT =: SHOOTING
          ACTOR --> HUMO =: John
          SHOOTING:AT --> HUM1 =: Mary
          BODY-PART =: HEAD

==> STATE =: DEAD
     OBJECT --> HUM1
```

which is intended to describe a crime committed by a person named John against a person named Mary, consisting of John's shooting Mary in the head, causing Mary's death.

2. Conceptual decisions and lexical choice

Given a result causation, one decision that a language-free conceptualizer might well need to make would be whether to express the STATE first and then the ACT, or to choose the opposite order. If these decisions were passed on to a dictionary, the synthesis of (A) above would be texts like

Mary is dead because John shot her in the head.

John shot Mary in the head. She is dead.

made up of one phrase expressing the STATE and one expressing the ACT. But it seems more satisfactory to produce texts such as

(1) *Mary was killed by John. He shot her in the head.*

(2) *John shot Mary in the head, killing her.*

built around *to kill*. Such texts don't follow conceptual decisions dissociating the STATE and its cause: *to kill* (in the construction $N_0 V N_1 =: John killed Mary$) expresses in the same time the death of N_1 and the fact that this death is due to an action (not specified) of N_0 (McCawley 1971). We showed in (Danlos 1984) that a formulation embodying a verb with a causal semantics such as *to kill* to describe the RESULT, and another verb to describe the ACT is, in most of the cases, preferable to a formulation composed of a phrase

for the STATE and another one for the ACT. This result indicates that conceptual decisions should not be made without taking into account the possibilities provided by the language, in the present case, the existence of verbs with a causal semantics such as *to kill*. This attitude is also imperative if a generator is to produce frozen phrases. The meaning of a frozen sentence being not calculable from the meaning of its constituents, frozen phrases cannot be generated from a language-free conceptualizer forwarding its decisions to a dictionary

3. Conceptual decisions, segmentation into sentences and syntactic constructions

Let us suppose that a result causation is to be generated by means of two verbs, one with a causal semantics such as *to kill* for the RESULT, and one for the ACT, and let us look at the ways to form a text embodying these two verbs. The options available are the following:

- **order of the information.** There are two possibilities. Either the phrase expressing the RESULT or the phrase expressing the ACT occurs first.
- **number of sentences.** There are two possibilities. Either combine the phrases expressing the RESULT and the ACT into a complex sentence, as in (2) (*John shot Mary in the head, killing her.*), or form a text made up of two sentences, one describing the ACT, one describing the RESULT, as in (1) (*Mary was killed by John. He shot her in the head.*).
- **choice of syntactic constructions.** We will restrict ourselves to the active construction and to the passive one. For the latter, there is the choice between passive with an agent and passive without an agent. On the whole, for each of the two verbs involved, there are three possibilities.

The combination of these 3 options gives 36 possibilities, but it turns out that only 15 of them are feasible. For example, texts composed of two sentences, one in a passive form with an agent, the other in a passive form without an agent, are appropriate to express a result causation only if the RESULT precedes the ACT, or if the agent is in the first sentence expressing the ACT:

- (3a) *Mary was killed by John. She was shot.*
- (3b) *Mary was killed. She was shot by John.*
- (3c) *Mary was shot by John. She was killed.*
- (3d) **Mary was shot. She was killed by John.*¹

1. A star (*) indicates that a text is awkward but it does not necessarily mean that it is ungrammatical or uninterpretable.

As another example, it is possible to combine the phrases expressing the ACT and the RESULT into a complex sentence if they are both in an active form

John shot Mary, killing her.

John killed Mary by shooting her.

but it is impossible if they are both in a passive form: the following formulations are awkward

**Mary was killed by being shot by John.*

**Mary was killed by John by being shot.*²

and the only other conceivable possibilities are to use a subordination conjunction such as *because*, *when* or *as*, but the resulting texts are clumsy:

**Mary was killed (because + when + as) she was shot by John.*

**Mary was shot by John and, because of that, she was killed.*

A generation system must know for each combination whether it is feasible or not. Either this knowledge is calculable from other data, or it constitutes data that must be provided to the generator. We are going to see that the second solution is better. First, on a semantic level, one can seek to verbalize the intuitions that can be drawn, for example, from paradigm (3), but this activity can be only descriptive and not explicative. In other words, the inacceptability of (3d) is a fact of language that cannot be explained by semantic computations of more general import. So the list of the 15 feasible combinations must be part of the data of the generator. Now the following question arises: is it possible to determine the structures of the texts corresponding to the 15 elements of this list. The answer is affirmative when the number of sentences is 2, and negative when it is 1. The combinations with two sentences involve only one type of linearization: juxtaposition. On the other hand, the combinations with one sentence involve

- a **present participle** if the ACT and RESULT are both expressed in an active form and if the ACT precedes the RESULT, as in *John shot Mary, killing her*
- a **gerundif** if the ACT and the RESULT are both expressed in an active form and if the RESULT precedes the ACT, as in *John killed Mary by shooting her*

2. The deletion of the agent leads to a formulation which is correct *Mary was killed by being shot* but which does not express the author of the crime.

- a relative clause if the RESULT is expressed in a passive form with an agent and precedes the ACT, this being expressed in an active form, as in *Mary was killed by John who shot her in the head*

- etc.

These types of linearization are not predictable. As a consequence, they must be provided to the generator. This one must embody in its data the structures of the texts corresponding to the 15 feasible combinations. These structures constitute a real discourse grammar for result causations. The formulation of result causations must be modelled on one of the 15 discourse structures³. Generating a result causation thus entails selecting one of these discourse structures.

4. Selection of a discourse structure

The fact that only 15 discourse structures out of 36 possibilities are feasible shows that it is not possible to make decisions about order of information, segmentation into sentences and syntactic constructions independently of one another. To do so could potentially result in awkward texts more than half the time.

Furthermore, lexical choice and selection of a discourse structure cannot be made independently of one another. A discourse structure leads to an acceptable text if and only if the formulations of the ACT and the RESULT present the syntactic properties required by the structure. For example, some causal verbs such as *to assassinate* cannot occur after a phrase describing the ACT:

**John shot the Pope in the head assassinating him.*

**John shot the Pope in the head. He assassinated him*⁴.

So, if the verb *to assassinate* is to be used, all of the

discourse structures in which the RESULT appears after the ACT are inappropriate. On the other hand, if a discourse structure where the RESULT occurs after the ACT is selected, the use of *to assassinate* is forbidden.

At this point, we have shown that decisions about lexical choice, order of the information, segmentation into sentences and syntactic constructions are all dependent on one another. This result is fundamental in generation since it has an immediate consequence: ordering these decisions amounts to giving them an order of priority.

5. Priorities in decisions

There is no general rule stating to which decisions priority must be given. It can vary from one case to another. For example, if a semantic representation describes a suicide, it is obviously appropriate to use *to commit suicide*. To do so, priority must be given to the lexical choice and not to the order of the information. If the order ACT-RESULT has been selected, it precludes the use of *to commit a suicide* which cannot occur after the description of the act performed to accomplish the suicide:

**John shot himself, committing suicide.*

**John shot himself. He committed suicide.*

On the other hand, if a result causation is part of a bigger story, and if strictly chronological order has been chosen to generate the whole story, then the result causation should be generated in the order ACT-RESULT. In other words, the order of the information should be given priority. In other situations, there is no clear evidence for giving priority to one decision over another one. As an illustration, let us take the case of a result causation which occurs in the context of a crime. It can be stated that the result DEAD must be expressed by:

- *to assassinate* as a first choice, *to kill* as a second choice, if the target is famous

- *to murder* as a first choice, *to kill* as a second choice, if the target is not famous

Moreover, the most appropriate order is, in general, RESULT-ACT if the target is famous, and ACT-RESULT otherwise. In the case of a famous target, the use of *to assassinate* is not in contradiction with the decision about the order of the information. But in the case of a non-famous target, the use of *to murder* doesn't fit the order ACT-RESULT, for this verb cannot occur after a description of the ACT:

**John shot Mary in the head, murdering her.*

**John shot Mary in the head. He murdered her.*

Therefore, either the decision about the order of the information or the decision to use *to murder* has to be

3. This point is akin to an assumption supported by (McKeown 1982), except that ours discourse structures contain linguistic information contrarily to hers which indicate only the order in which the information must appear.

4. These forms become acceptable if they are added adverbial phrases:
John shot the Pope in the head, thereby assassinating him in a spectacular way.
John shot the Pope in the head. Thereby he assassinated him in a spectacular way.

forsaken. The former solution would yield to texts such as

John murdered Mary by shooting her in the head.

John murdered Mary. He shot her in the head.

where the order of the information is RESULT-ACT, and the latter one to texts such as

John shot Mary in the head, killing her.

John shot Mary in the head. He killed her.

using the verb *to kill* instead of *to murder*. At the current time, the choice between these two solutions can be based only on intuitions that are not sufficiently operational to be integrated in a generation system.

Conclusion and future research

We have shown that decisions about lexical choice, determination of the order of the information, segmentation into sentences and choice of syntactic construction are all dependent one another, the last three amounting to the selection of a discourse structure by means of a discourse grammar. As a consequence, a generation system must be based on a complete interaction between these decisions. In this work, we have been concerned only with the task of expressing into natural language a set of information. In other words, we have only dealt with the generation problem of "How to say it?", and not with the problem "What to say?".

Some authors (cf. (McGuire 1980) and (Appelt 1982)) have rejected the separation between "What to say" and "How to say it" on the basis that the issue of "What to say" is not independent from the lexical choice. Thus, they have argued for generation systems involving interactions between conceptual decisions and linguistic ones. This point is akin to ours, and therefore, our model of generation could be extended so as to treat issues such as generating different texts according to the hearer and what it is supposed that he wants and/or needs to hear.

ACKNOWLEDGEMENTS

I would like to thank Lawrence Birnbaum for many valuable discussions and suggestions on this paper.

REFERENCES

- Appelt, D.E., 1982, *Planning Natural-Language Utterances to satisfy Multiple Goals*, Technical Note 259, SRI International, Menlo Park, California.
- Danlos, L., 1984, *Génération automatique de textes en langues naturelles*, Thèse d'Etat, Université de Paris 7.
- McCawley, J. D., 1971, "Prelexical Syntax" in *Report of the 22nd annual round table meeting on Linguistics and Language Studies*, O'Brien éd., Georgetown University Press.
- McDonald, D., 1983, "Natural Language Generation as a Computational Problem: an introduction", in *Computational Models of Discourse*, Brady et Berwick éd., MIT Press, Cambridge, Massachusetts.
- McGuire, R., 1980, "Political primaries and words of pain", unpublished manuscript, Yale University.
- McKeown, K. R., 1982, *Generating Natural Language Text in response to Questions about database structure*, PhD Dissertation, University of Pennsylvania.
- Schank, R.C., 1975, *Conceptual Information Processing*, North Holland, Amsterdam.