Chapter 23

INCREMENTAL SENTENCE PRODUCTION, SELF-CORRECTION AND COORDINATION

Koenraad De Smedt
Gerard Kempen

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Stages of processing

Since Garrett's (1975, 1980) seminal work on speech error phenomena, it has become customary to distinguish four levels of representation within the sentence production process: a message level, a functional level, a positional level, and a phonetic level (see also Bock, this volume). Garrett's model has been further elaborated and modified by Kempen (Kempen & Hoenkamp, in press; Van Wijk & Kempen, in press) who proposes the global sentence production model depicted in Figure 1. The four modules listed have the following functions:

1. The *conceptual* module forms a conceptual (semantic) representation of the message which the speaker wishes to communicate. The nature of the semantic structures output by this component need not concern us here.
2. The *lexico-syntactic* module constructs an ordered tree structure consisting of constituents and their functional relations. The terminal nodes of syntactic trees (both content and function words) are instances of abstract (not phonologically specified) lexical items called *lemmas* which are retrieved from the lexicon. While Garrett assigns the tasks of inserting function words and computing word order to a later module (the positional stage), Kempen assigns them to this one.

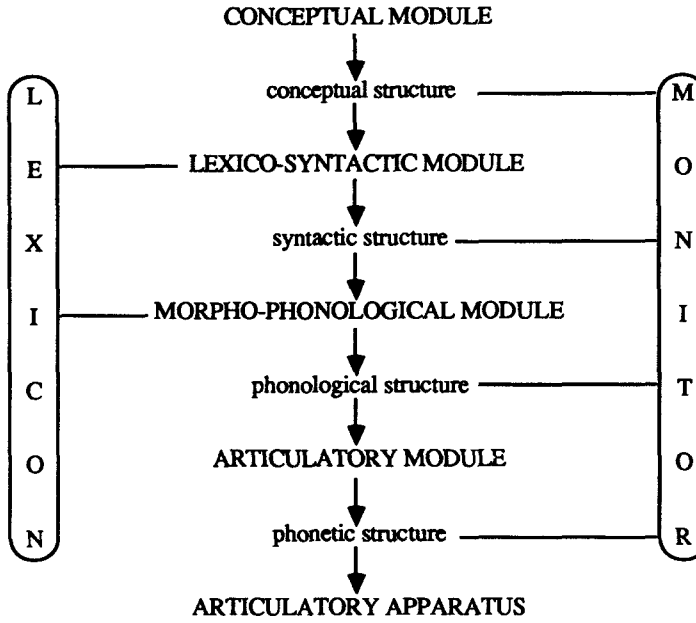

Figure 1. A global model of the sentence production process

3. The *morpho-phonological* module computes the word form of all lemmas by retrieving their phonological specifications (*lexemes*) from the lexicon and making various morphological and phonological adjustments.
4. The *articulatory* module produces a phonetic specification which is used to control the articulatory apparatus.

The intermediate results, which are passed from one module to another, are inspected by a monitor. If the monitor notices that the output of one of the modules is inappropriate or detects a violation of some prevailing constraint, any ongoing activity may be interrupted and backtracking to an earlier point in the production process may be forced. This course of events may give rise to self-corrections.

1.2 Incremental production

The four sequential modules of Figure 1 need not necessarily operate on input structures which correspond to whole sentences. If the modules did operate in this fashion, hesitations *during* the pronunciation of a sentence could not have a nonarticulatory (i.e. a conceptual, syntactic, or lexical) origin. Since this is both counterintuitive and counterfactual, we favor the view that the modules can work on different parts of the final utterance simultaneously, as depicted in Figure 2. We call this piecemeal mode of production *incremental production* (Kempen, 1978).

Although the modules involved in sentence production may work in parallel, each fragment of an utterance still goes through the different stages sequentially. The communication channel between the modules operating in this incremental fashion can be modeled in terms of *streams* (cf. Hoenkamp, 1983, pp. 114-117). For instance, we hypothesize that conceptual fragments are entered at one end of a stream and 'consumed' by the lexico-syntactic module at the other end, as shown in Figure 3.

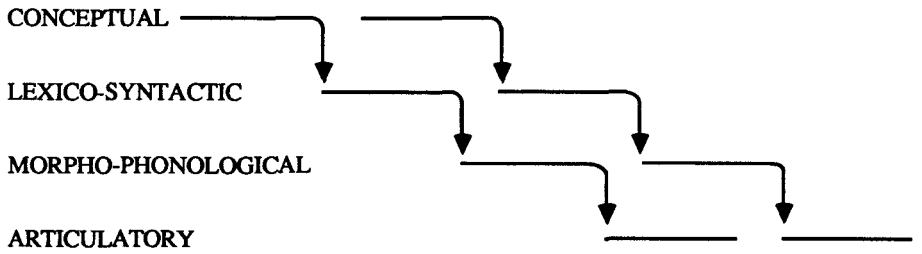

Figure 2. Incremental processing of two fragments

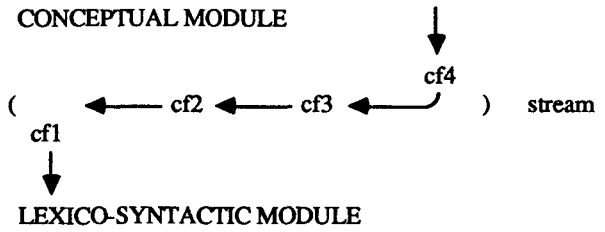

Figure 3. Stream of conceptual fragments

While the lexico-syntactic module is processing elements from the stream, the conceptual module can run simultaneously and add more elements to the end of the stream. We assume that the flow of information between modules is downward; in other words, each module passes intermediate structures to the next module but receives no information back from that module. All upward information, we assume, is a result of monitoring. The conceptual fragments contain markers which indicate their relationship to fragments earlier in the stream.

Such a framework can easily accommodate the fact that hesitations may occur within the sentence as well as between sentences. Also, it can account for *syntactic deadlock*, i.e., the fact that people sometimes 'talk themselves into a corner' when they have produced a partial sentence which they cannot continue in any way they consider appropriate or meaningful, because of lexico-syntactic restrictions. In such circumstances, self-corrections may be triggered. Moreover, the framework allows for 'changes of mind', i.e., decisions by the speaker to revise the conceptual content which has already been expressed. This is represented in the stream as a conceptual fragment marked as a substitute for an earlier fragment.

We will now discuss incremental sentence production and self-correction from the point of view of their origins: conceptual modifications and monitoring. Then some lexico-syntactic mechanisms for dealing with these events will be proposed. We conclude with some comparisons with related research.

2. CAUSES OF INCREMENTATION AND CORRECTION

2.1 Conceptual modifications

We distinguish three basic kinds of modification to a conceptual structure which will affect the shape of an utterance: *deletion*, *replacement* and *addition* of conceptual elements. Deletion and replacement will both give rise to a self-correction, which is often signaled by a pause or a

correction term such as *uh, no, sorry*. Some examples of deletion are (1) and (2). Examples of replacement are (3) and (4).

(1) John and Mary ...uh... only John went to a party last week.
(2) John bought a new bicycle for ...uh... a bicycle for his son.
(3) John ...uh sorry... Mary went to the party.
(4) The runner with the beard ...no... with the glasses is leading now.

Conceptual replacement may also lead to a non-retracing repair. The result is ungrammatical but contains no correction marker and is uttered without hesitation. The examples for English (5) and for Dutch (6) show how a constituent can be replaced without retracing. One or more constituents which have already been uttered are used as a hook to attach a new sentence pattern with a different word order.

(5) That's the only thing he does is fight.
(6) Willemse heeft gisteren heeft de dokter nog gezegd dat het mag.
(Willemse has yesterday has the doctor said it is allowed.)

While conceptual deletion and replacement seem to be relatively infrequent as causes of incrementation or correction, addition is frequent. We assume that conceptual processing, just like syntactic processing, takes place in a piecemeal way, so that the continual addition of conceptual fragments to existing ones will be quite normal. Addition can be of two kinds. The first kind is an addition of a conceptual fragment which is to be in *conjunction* or *disjunction* with an existing fragment and thus leads to a syntactic coordination, as in (7) and (8).

(7) John ... and Mary went to a party.
(8) John ... or Mary went to a party.

The second kind is the addition of a new conceptual fragment in any other relationship than conjunction or disjunction. This may give rise to an *expansion*, i.e., the current utterance is continued with a syntactic fragment which is not a member of a coordination but has some syntactic relation (such as subject, direct object, modifier, etc.) to the current utterance or part of it. Simple examples are (9) and (10).

(9) John and Mary ... went to a party.
(10) John and Mary went ... to a party.

2.2 Monitoring

After a conceptual addition, it may not always be syntactically possible to continue a partially uttered sentence. The lexico-syntactic restrictions imposed by what has already been uttered may severely limit the possible ways of expanding the syntactic structure or finding an appropriate word order. In English, for example, it seems impossible to expand (11) to express a conceptual increment corresponding to *likes to*, as in (12).

(11) John comes ...
(12) John likes to come.

By contrast, an equivalent downward expansion is possible in Dutch, where the meaning underlying *likes to* can be expressed by means of an adverbial phrase as in (13).

(13) Jan komt ... graag.

The difference between the English and the Dutch example shows that the restrictions are lexico-syntactic in nature. In circumstances where expansion is impossible, the monitor will

receive no output from the lexico-syntactic component. A syntactic deadlock will thus be detected and a self-correction will be triggered by causing the conceptual structure to re-enter the lexico-syntactic module and thus to be reformulated, as in (14).

(14) John comes ...uh... likes to come to the party.

Another example of an impossible expansion in English is the expansion of (15) to (16). However, the apposition in (17) or the relative clause in (18) offer alternatives. There may be a covert self-correction during the formulation of these sentences, marked by a pause.

(15) The man ...
(16) The bald man ...
(17) The man ... the bald one that is, ...
(18) The man ... who is bald, ...

Syntactic deadlock is of course but one possible cause of self-correction. Other types of error which are detected by the monitor and which may thus result in a self-correction include the choice of wrong lexical material, fusion errors, and articulation errors. It is often unclear whether in a particular utterance, e.g. (3), the cause of the correction is a conceptual modification or the detection of a lexical error. A discussion of these phenomena is beyond the scope of this chapter. The question of how much conceptual material re-enters the stream to produce a self-correction is an interesting one, but it will likewise not be discussed here (see Van Wijk & Kempen, in press, for some relevant findings and ideas). Our present aim is to show the global picture of the relations between incremental conceptualization and self-correction.

2.3 Overview

Figure 4 gives a schematic overview of the conceptual and monitoring processes discussed in this section. The process flow is downward. Non-retracing repairs and normal incrementation are grouped together in this overview.

In the following section, the three types of lexico-syntactic mechanisms involved (expansion, coordination and correction) will be discussed in more detail.

3. SYNTACTIC MECHANISMS

3.1 Expansion

We distinguish three kinds of expansion, depending on the location in the tree where a new syntactic fragment is added. *Upward* expansion causes the tree to grow upward, i.e., the original root node is no longer the root node of the expanded tree. Other cases we term *downward* expansion, when new branches are added below an existing node. Finally there is a special case called *insertion*, when syntactic material is inserted between existing nodes. Figure 5 shows roughly how the various kinds of expansions affect a syntactic tree. The utterance depicted is (19).

(19) John and Mary are at the party ... seem to be at the party.

Insertion does not necessarily lead to a self-correction, as was the case in (19). An example where insertion leads to the continuation of a fragment which has already been uttered is the Dutch sentence (20). The English translation contains a correction, but the Dutch original does not. The insertion is depicted in Figure 6.

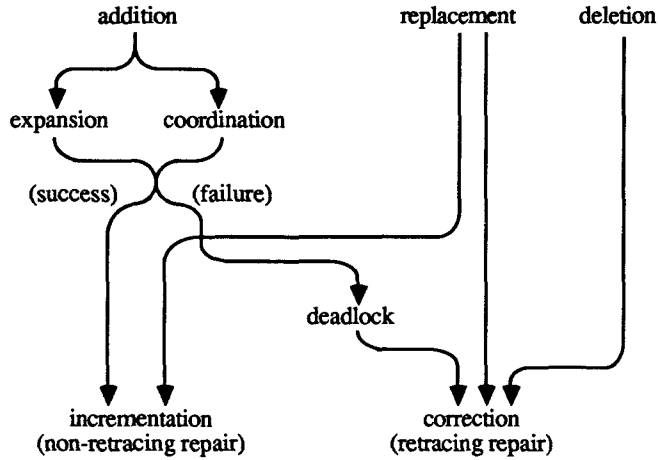

Figure 4. Conceptual modifications, monitoring and lexico-syntactic consequences

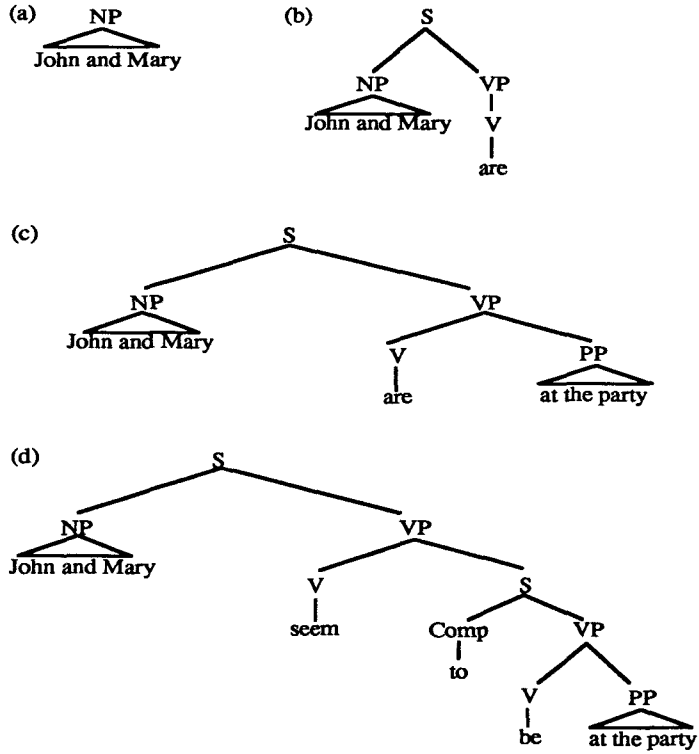

Figure 5. Upward (b) and downward (c) expansion, insertion (d)

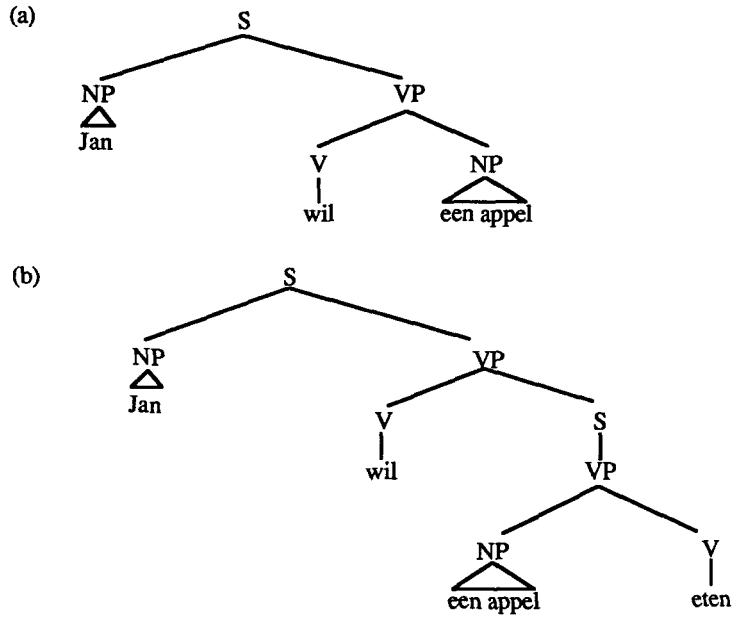

Figure 6. Insertion in utterance (20)

(20) Jan wil een appel ... eten.
(John wants an apple ... wants to eat an apple.)

If upward expansion is allowed, then one must also allow situations where an initial conceptual fragment does not lead to the construction of a main clause. Instead, an isolated noun phrase may be uttered, as in (21).

(21) He ...

Such an initial constituent is 'unattached' in the sense that it does not have a syntactic relation to a mother node. Although a subsequent conceptual fragment may cause the construction of a mother node, it would be a handicap if uttering the initial constituent had to be postponed until the constituent was assigned a relation to it. However, how should the lexico-syntactic module make decisions which depend on such a syntactic relation, for example choosing the surface case marking (*he*, *him*, *his*) while that relation has not yet been specified? One possible solution consists in carrying out one or more *provisional* upward expansions until a sentence node has been created. Subsequent conceptual fragments may lead to syntactic fragments which are *actual* upward expansions. The system then attempts to combine the actual syntactic nodes with the provisional ones. If this *unification* (Kay, 1979) is successful, the nodes are merged. This leads to a successful expansion, as in (22). The unification will fail when the syntactic functions in the provisional and the actual expansions of nodes are different. In that case, either a *restart* using a different syntactic structure (an *anacoluthon*) may take place (23), or lexico-syntactic alternatives may be explored which might lead to a successful expansion, for example, by means of passivization (24).

(22) He ... left.
(23) He ... They invited him.

(24) He ... was invited.

What heuristics or preferences does the lexico-syntactic module use when choosing between alternative possibilities for provisional upward expansions? A partial answer is provided by Bock and Warren (1985). They establish a relationship between *conceptual accessibility* (the ease of retrieving conceptual information from memory) and the *hierarchy of grammatical relations* which plays a role in various cross-linguistic and within-language phenomena (Keenan & Comrie, 1977):

subj. > dir. obj. > indir. obj. > oblique > genitive > obj. of comparison

Similar results were obtained in a sentence recall experiment performed by Keenan & Hawkins (1987). Our hypothesis is that this (or a similar) hierarchy plays a role as a preference scale in the incremental production of sentences. The first constituent which is to be in a syntactic relation with a sentence will have a higher probability of being realized as a subject than as direct object, etc., according to the hierarchy. Subsequent fragments may find the relations higher in the hierarchy already occupied by previously created constituents and will be assigned a function lower in the hierarchy. Since the hierarchy is correlated with word order, it thus serves to guide the sentence formulation process toward maximally fluent incremental sentence production.

Other factors may complement the use of the relational hierarchy in incremental sentence production. We will only point out one other factor here. There is probably some direct interaction between the assignment of a syntactic function and the type of the conceptual fragment. For example, the preferred function assigned to a time-indicating NP such as (25) may not be subject but some lower member of the hierarchy such as sentence modifier (oblique).

(25) Monday morning, ...

3.2 Coordination

Coordination is viewed as an iteration of the lexico-syntactic process on several conceptual fragments which are linked to each other as members of a conjunction or disjunction. The result of lexicalizing and formulating these is a special phrase called a *coordination* which has a number of conjuncts as its immediate constituents.

Often coordinations come about during an incremental process because the speaker may keep adding conjuncts, even after some have been uttered. Therefore, the list of conceptual fragments to be coordinated is viewed as a stream (cf. Figure 3). The stream is buffered to allow detection of the end of the stream. Conjuncts are often realized with 'comma intonation' as long as there is at least one further element in the stream. If it is the final element, it is added after insertion of a conjunction like *and*. This treatment of coordination as an incremental process accounts for sentences in which 'afterthoughts' may give rise to multiple occurrences of the conjunction word (26) or even *dislocations* (27).

(26) John, Peter and Mary ... and Anne came home.
(27) John, Peter and Mary came home ... and Anne.

Utterances like these are not unusual in spoken language. We account for them by assuming that new descriptions have entered the stream after it had been emptied.

3.3 Self-correction

Self-corrections are governed by rules which determine how much of the original utterance needs to be repeated. For example, (28) is not well-formed because in the self-correction all

constituents to the right of the replaced main verb should be reformulated, as in (29). Likewise, (30) is not grammatical because the entire NP should be reformulated (31).

(28) * You should have sent that letter ...uh... handed over.
(29) You should have sent that letter ...uh... handed it over.
(30) * Tony is baking a cake ... sugar-free.
(31) Tony is baking a cake ... a sugar-free cake.

Levelt (1983, p. 78) has observed that the rule which speakers obey when deciding how far they should retrace is similar to the retracing rule for coordinations. He then stated a *well-formedness rule* for repairs in terms of the grammaticality of coordinations, linking the ill/well-formedness of (28) and (29) to that of (32) and (33) respectively.

(32) * You should have sent that letter or handed over.
(33) You should have sent that letter or handed it over.

Following Levelt's rule, we propose a mechanism for generating self-corrections which has the same underlying principles as the mechanism for coordination. If an error has been detected by means of monitoring and its cause has been diagnosed (deadlock, conceptual replacement, lexicalization error, etc.), a conceptual fragment marked as the correction of some earlier fragment is inserted at the end of the stream. This correction fragment may include 'old' concepts that have already been linguistically realized. The correction marker is treated by the lexico-syntactic module in much the same way as the conjunction marker, the only difference being that it is realized as a pause or as a correction term (such as *uh*), rather than as comma intonation or a conjunction (*and, or*). Thus Levelt's observation is fully accounted for.

However, Van Wijk & Kempen (in press), who have verified Levelt's well-formedness rule, found that it covers only one type of self-corrections, which they call *reformulation*. Self-corrections of another type, which they call *lemma substitution*, e.g. (34), do not need computing a new syntactic structure, because simply replacing a lemma in the existing structure suffices.

(34) Dou you really want to buy that record ...uh... compact disc?

Other self-corrections are really restarts, i.e., instead of carrying out a repair, a whole utterance is rejected and the speaker starts all over (cf. (35)).

(35) Did the student ...uh... Did you ask the student anything?

Although restarts could be seen as a special case of reformulation, perhaps they should be handled by a mechanism which is different from that for repairs because the relationship between reparation and reparandum is a different one.

The choice between correction strategies made by the lexico-syntactic module seems to be partially dependent on the origin of the correction. Van Wijk & Kempen found that conceptual addition often leads to reformulation while replacement and deletion often trigger lemma substitution. In addition, we would like to suggest a causal relation between syntactic deadlock and restart.

Example (36) shows that self-correction and coordination can occur in one and the same constituent. In addition, examples (36) and (37) illustrate that the ambiguity of certain self-corrections is similar to that of corresponding coordinations, which again suggests that they should be treated in a similar way.

(36) Peter and Mary ...uh... John left the house.
(37) Peter and Mary or John left the house.

3.4 Control structure

Because the process of deleting, replacing and adding conceptual material may occur repeatedly and independently of each other, the various lexico-syntactic mechanisms involved, namely self-correction, coordination and expansion, may occur in one utterance and even embedded in one another. For example, a conceptual addition may cause a coordination; within one of the conjuncts, a conceptual addition may lead to an attempt at expansion, which, if unsuccessful, will cause a correction to occur, etc. An annotated example of such a sequence is (38).

(38) Peter ...
and a woman ... (conjunction)
who sleeps ... (downward expansion)
who never sleeps more than five hours a night ... (downward expansion with retracing)
or even less ... (disjunction)
came early to my party. (upward expansion)

Consequently, the lexico-syntactic module will need a control structure where the processes can complement each other. We propose a control structure with *nested* iteration loops on the output of the conceptual module. One loop is expansion, which may cause the addition of mother or daughter nodes in the syntactic tree. The other loop combines correction and coordination. It may iterate within each constituent, where it causes the addition of coordinates or corrections. Each of the two loops may be nested within the other one.

4. RELATED RESEARCH

Most natural language generation systems in the literature have not been designed for the simulation of spontaneous speech but for the construction of carefully planned sentences and texts. Hence it will not be surprising that in most systems the conceptual and lexico-syntactic stages are ordered strictly serially for a complete sentence. However, some attention has been given to incremental production in at least two other systems: MUMBLE and KAMP.

In MUMBLE (McDonald & Pustejovsky, 1985a), a conceptual 'planner' and a linguistic module call each other recursively. A surface structure of the sentence is extended in the process. Predefined 'attachment points' in that surface structure determine where and how it can be extended. These extensions seem to be limited to downward expansions and possibly conjuncts: 'another adjective added to a certain noun phrase, a temporal adjunct added to a clause...'(p. 189).

McDonald & Pustejovsky (1985a, 1985b) point out that there is a similarity between their 'attachment' and the grammar formalism in Tree Adjoining Grammars (TAGs; Vijay-Shankar & Joshi, 1985; see also Joshi, this volume). This suggests that TAGs are formalisms which may be especially suitable for incremental generation. They seem capable of simulating a variety of expansions, although the integration with other modules involved in sentence production remains to be worked out. McDonald & Pustejovsky's (1985b) discussion of TAGs is limited to insertions. The example they work out concerns the expansion of (39) to (40). In our treatment, (40) would be realized as a self-correction (41), once the initial sequence (39) has been uttered. However, McDonald & Pustejovsky's system does not seem to start uttering a sentence until it is complete, thereby obviating the need for self-correction.

(39) The ships were hit.
(40) The ships were reported to be hit.
(41) The ships were hit ...uh... were reported to be hit.

In the KAMP system (Appelt, 1983), there is a component called TELEGRAM which couples the processes of conceptualization and formulation in an incremental architecture. In Functional Unification Grammar (Kay, 1979), a sentence can be produced by the *unification* of two *functional descriptions* (FDs). One of these represents a partially specified utterance and

possibly includes some conceptual information. The other one is the grammar of the language. Instead of doing a single unification between a completely specified FD for the sentence as a whole (the 'text' FD) and the grammar (the 'grammar' FD), the TELEGRAM planner works by gradual refinement. Initially, a high-level, incomplete text FD is produced by the planner and unified with the grammar FD. Subsequent planning produces more FDs, which are unified with the grammar FD and incorporated into the text FD. However, the system plans hierarchically, and the resulting enrichments of the text FD seem to be limited to downward expansion and possibly coordination.

There seem to be no natural language generation systems which produce incrementally in such a way that every now and then the system 'talks itself into a corner' and has to backtrack for a self-correction. Existing systems are only partially incremental: Even if they allow the conceptual input to be modified while syntactic sentence construction is already on its way, the uttering of the sentence is delayed until its surface structure is complete. Thus the need for self-corrections is avoided.

One could argue that there is no practical need for artificial language generation systems which can generate truly incrementally and that the risk of an occasional self-correction is only a nuisance. Als long as the systems generate printed output, we agree. But in the case of spoken output, the situation is different. Human listeners hardly have any trouble with corrections and retracings in speech. Therefore, in order to prevent unnaturally long pauses between successive sentences, the system could profitably resort to an incremental production strategy.

In theoretical linguistics, formal grammars seem to be biased toward one or the other kind of expansion, upward or downward. While phrase structure grammars present rules in a manner which is suitable only for downward expansion, categorial grammars specify rules for upward expansion. TAGs seem to suffer less from this bias because they use insertion as a basic mechanism. However, it is not clear whether they could handle cases such as Figure 5a-b, where an isolated NP is attached to an S as a daughter node.

We conclude that a new type of grammar is needed which can generate not only complete grammatical sentences and their structural trees but also sequences of incomplete trees which may arise during the planning of a full sentence. (For an initial proposal concerning such a grammar, see Kempen, 1987).

5. SUMMARY

We have seen how incremental production and self-corrections can be accounted for by allowing increments and other modifications to the conceptual input after the syntactic formulation process has already started. We assume that different modules which are involved in sentence production (i.e. conceptualization, formulation and articulation) can run in parallel. Three types of conceptual modifications may occur while the formulation is already on the way: deletion, replacement and addition. Deletion and replacement of a conceptual fragment which is already being formulated typically give rise to a self-correction. Addition may give rise to a coordination or an expansion. Of the latter there are three types: upward and downward expansion and a mixed case called insertion.

A monitor inspects the results of the production process, which allows the detection of errors. One such error, deadlock, occurs when it is impossible to continue a syntactic fragment with the desired increment. Upon the detection of errors, self-corrections may be triggered. To our knowledge, there is at present no formalism which can generate truly incrementally.

ACKNOWLEDGEMENTS

We would like to thank all people who provided helpful comments on earlier versions of this chapter, in particular Willem Levelt, Carel van Wijk and Anthony Jameson.

REFERENCES

Appelt, D. (1983) TELEGRAM: a grammar formalism for language planning. In: *Proceedings of the Eighth IJCAI*, Karlsruhe.

Bock, J. & Warren, R. (1985) Conceptual accessibility and syntactic structure in sentence formulation. *Cognition, 21*, 47-67.

Garrett, M. (1975) The analysis of sentence production. In: G. Bower (Ed.) *The psychology of learning and motivation (Vol. 9)*. New York: Academic Press.

Garrett, M. (1980) Levels of processing in sentence production. In: B. Butterworth (Ed.) *Language production (Vol. 1: Speech and Talk)*. New York: Academic Press.

Hoenkamp, E. (1983) *Een computermodel van de spreker: psychologische en linguistische aspecten*. Ph.D. Dissertation, University of Nijmegen, The Netherlands.

Kay, M. (1979) Functional grammar. In: *Proceedings of the fifth annual meeting of the Berkeley Linguistic Society*.

Keenan, E. & Comrie, B. (1977) Noun phrase accessibility and universal grammar. *Linguistic Inquiry, 8*, 63-99.

Keenan, E. & Hawkins, S. (1987) The psychological validity of the accessibility hierarchy. In: E. Keenan, *Universal Grammar*: 15 Essays. London: Croom Helm.

Kempen, G. (1978) Sentence construction by a psychologically plausible formulator. In: R. Campbell & P. Smith (Eds.) *Recent advances in the psychology of language (Vol. 2: formal and experimental approaches)*. New York: Plenum Press.

Kempen, G. (1987) A framework for incremental syntactic tree formation. In: *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)*, Milan.

Kempen, G. & Hoenkamp, E. (in press) An incremental procedural grammar for sentence formulation. To appear in *Cognitive Science*.

Levelt, W. (1983) Monitoring and self-repair in speech. *Cognition, 14*, 41-104.

McDonald, D. & Pustejovsky, J. (1985a) A computational theory of prose style for natural language generation. In: *Proceedings of the second conference of the European Chapter of the Association for Computational Linguistics*, Geneva.

McDonald, D. & Pustejovsky, J. (1985b) TAG's as a grammatical formalism for generation. In: *Proceedings of the 23rd annual meeting of the Association for Computational Linguistics*, Chicago.

Vijay-Shankar, K. & Joshi, A. (1985) Some computational properties of Tree Adjoining Grammars. In: *Proceedings of the 23rd annual meeting of the Association for Computational Linguistics*, Chicago.

Van Wijk, C. & Kempen, G. (in press) A dual system for producing self-repairs in spontaneous speech: evidence from experimentally elicited corrections. To appear in *Cognitive Psychology*.