

# Generating Coherent Paragraphs

*Greg Lessard<sup>1</sup> and Michael Levison<sup>2</sup>*

*Department of French Studies, Queen's University, Kingston, Ontario, Canada, K7L 3N6*

KEYWORDS: generation, language learning

AFFILIATION: <sup>1</sup>French Studies, Queen's University, <sup>2</sup>Computing and Information Science, Queen's University

E-MAIL: lessard@francais.queensu.ca  
levison@qucis.queensu.ca

FAX NUMBER: (613) 545 6522

PHONE NUMBER: (613) 545 2088

Recent years have seen something of a divorce between models of computer-aided instruction and theories of second language learning. For example, it has come to be recognized that a satisfactory model of language teaching must incorporate the notion of 'communicative competence', that is, not only the production of grammatically correct utterances, but also the appropriate use of utterances according to situations. At the same time, it has been recognized that linguistic competence goes beyond sentence-length utterances to include textual relations such as anaphora, inference, logical connectives and the like. Finally, it has been suggested that decontextualized utterances with little or no reference to reality do not necessarily serve the interests of language learners; rather, they should be exposed to 'authentic' texts which contain information about the cultural context of the language being taught. Such limitations affect not only instructional systems, but also subject testing environments aimed at capturing in finer detail the nuances of linguistic performance.

One solution to this problem lies in the construction of finely crafted environments in which complex linguistic, discourse and pragmatic relations may be represented. On a basic level, there exists a broad range of authoring systems (Calis, from Duke University, to name but one) which allow the manual entry and tagging of authentic texts to produce comprehension exercises. At a more sophisticated level, one finds the use of programmed 'microworlds' which may be navigated by learners. For example, Hamburger (1994) proposes a language learning environment based on a 'Kitchen World' in which learners may manipulate objects (turning on a faucet, for example), read descriptions and produce utterances ordering that

actions be performed.

The development of more 'authentic' examples of the sort will depend crucially on progress in dealing with textual inference and incorporating encyclopedic knowledge. Work of the sort has been done. For example, the BORIS program (Lehnert 1983) is capable of understanding a complex narrative concerning divorce. On a broader scale, the CYC project (Lenat 1990) involves the attempt to capture a broad range of encyclopedic information and related inferences.

Despite their promise, all of these approaches have a number of weaknesses. BORIS is capable of understanding relatively complex texts dealing with divorce and shows an impressive ability to extract implicit information from such texts. However, the underpinnings of this information are buried in the complex code of the program, with the result that extensions to other domains are difficult. Similarly, extension of Hamburger's microworld would involve significant reprogramming. Despite its broad coverage, the CYC project is not directly applicable to language learning. Finally, traditional hand coding of authentic texts provides precise control and textual authenticity, but at the cost of time and effort, since the author must make explicit all the implicit components of a text. Also, such authentic texts tend to 'date' quite rapidly, so that what was current one year may be outmoded a few years later.

In this paper we investigate an intermediate path: a simple metalanguage capable of generating at least basic paragraphs which deal with real-world phenomena, with some facilities for generalization to a range of examples. In this way, one could go beyond simple decontextualized sentences, while retaining control over the structure of paragraph-length utterances.

The VINCI environment was selected for the experiment, given its generalized power of expression, its multilingual capacities and its use of a linguist-friendly metalanguage. The VINCI system has been described elsewhere in a number of papers (Levison and Lessard 1992, for example). In essence, it is composed of a number of formalisms for describing the syntax, semantics, lexicon and morphology of some subset of a language and for generating utterances according to the description.

A first attempt at producing coherent and contextualized utterances was made in French. An existing French lexicon was extended to include a range of encyclopedic information, including names of French authors, their dates of birth and death, and titles of their major works. The links between these pieces of information are specified by 'lexical pointers' as the following examples illustrate:

"Benjamin Constant"|ENCYC|>humain.Fonction|#1...|  
 naissance:"1767";mort:"1830"|  
 "Chateaubriand"|ENCYC|>humain.Fonction|#1...|  
 naissance:"1768";mort:"1848"|  
 ""Adolphe""|ENCYC|>écrit.Fonction|#1...|auteur:"Benjamin  
 Constant"|  
 ""René""|ENCYC|>écrit.Fonction|#1...|auteur:  
 "Chateaubriand"|  
 ""Mémoires d'outre-tombe""|ENCYC|>écrit.Fonction|#1...|  
 auteur:"Chateaubriand"|  
 "1767"|ENCYC|>année|#1...|siècle:"18ième siècle"|  
 "1768"|ENCYC|>année|#1...|siècle:"18ième siècle"|

Here we see that “Chateaubriand” points at a birthdate (naissance) of “1767”. This date is itself a lexical entry, which points at the century to which it belongs “18ième siècle”. Similarly, the two novels “René” and “Mémoires d’outre-tombe” point at their author.

Using this information, with appropriate syntactic mechanisms, VINCI constructs simple dialogues such as the following, where ‘Q’ represents the computer’s question written to the screen and ‘A’ the expected answers which it stores in a hidden file for comparison with user input.

Q: Qui a écrit ‘René’?

A: Chateaubriand

Q: Quand est-ce que Benjamin Constant est né?

A: Il est né en 1767.

A: Il est né au 18ième siècle.

A: Il est né dans le 18ième siècle.

This framework was used for subject testing in an attempt to tease out more detailed data on the use of “dans” as opposed to “au” for temporal reference by anglophone learners of French (the third answer in the second example above). An indication of the extent to which the environment successfully hid its grammatical agenda is provided by the remark made by several subjects that they were ashamed of their poor knowledge of French literature and planned to enrol straightaway in a course!

While this framework illustrates the possibilities of including encyclopedic information in a generative system, as well as the possibilities of dialogue, links between individual utterances continue to be purely random. Any attempt to address this problem must take account of the grammar of texts and of a range of discourse processes. We will provide two examples of such constraints.

(1) Consider first the level of the sentence. There has been considerable discussion of the role of ‘thematic hierarchies’ in the presentation of textu-

al information. For example, Allen (1987) argues that the ordering of sentence elements reflects the influence of a series of overlapping hierarchies whose overall effect is to place familiar information before new, animates before inanimates, and so on. Corpus work on Preferred Argument Structure appears to provide empirical support for a number of grammatical constraints of a similar kind. Thus Ashby and Bentivoglio (1993) found that in two-argument verbs in French, there is a strong tendency for initial elements to be pronouns rather than full noun phrases.

(2) At the paragraph level, a variety of approaches have been suggested to account for the structure of text, ranging from schemas (McKeown) to frameworks such as Rhetorical Structure Theory (RST) (Mann and Thompson 1987) which provides a typology of paragraph types linking argument structures and formal paragraph patterns, to mixtures of the two (Hovy 1988). There has been a great deal of research on such questions in the area of Natural Language Generation. However, we are not aware of attempts to apply it to second language performance research or language teaching. In particular, there is need for empirical research on the ability of second language learners to conceptualize textual structures while handicapped by limited lexical and syntactic resources. In an attempt to work with this problem, we have proposed mini-grammars of paragraph structure. Somewhat simplified, these contain the following elements:

– *lexical items* tagged with a rich set of lexical pointers to encyclopedically related items. Thus, as the following example shows, the item “apple” includes pointers to positive (p) and negative (n) evaluations of taste, texture and colour, as well as pointers to kinds of apples.

```
"apple"|N|>indef,>apple.Fonction,Attitude,Number||$1|||
  ptaste:"sweet"/ADJ; ntaste:"sour"/ADJ;
  ptexture:"crunchy"/ADJ, "crisp"/ADJ, "juicy"/ADJ;
  ntexture:"mushy"/ADJ, "pulpy"/ADJ, "soft"/ADJ;
  pcolour:"red"/ADJ;
  ncolour:"green"/ADJ;
  kind:"MacIntosh", "Spy";|||
```

– *syntactic rules* (phrase structure and transformations) capable of operating on initial elements in order to generate coherent paragraphs, as in the following example, which defines a paragraph presenting an initial value judgement followed by supporting evidence and finally additional detail.

PARA1 = ( MAKESUBJECTIVE | MAKEOBJECTIVE )

(  
MAKEPRONSUBJ MAKEPTASTE  
MAKEPRONSUBJ MAKEPCOLOUR  
MAKEPRONSUBJ MAKEPTEXTURE  
|  
MAKEPRONSUBJ MAKEPTASTE  
MAKEPCOLOUR  
MAKEPTEXTURE  
)

MAKEHEALTH  
MAKELIKEKIND  
%

In essence, this rule states that the structure PARA1 is composed of a combination of metavariables (in capital letters) each of which defines a particular syntactic operation on a base form. Thus, MAKESUBJECTIVE defines a sentence which expresses a subjective opinion with respect to the base form and the attitude chosen (for example: "I like x"). MAKEOBJECTIVE expresses the same attitude as an objective statement, as in "x are nice". The vertical bar causes a random choice to be made between the two. Similarly, metavariables like MAKEPTASTE use pointer information from the base form to construct utterances which attribute an appropriate positive description of the taste, texture or colour of the base form. MAKEHEALTH produces a judgement on the healthy properties of the base form while MAKELIKEKIND uses pointer information on the base to talk about subsidiary kinds.

Application of this 'paragraph grammar' to particular lexical items produces simple paragraphs, as the following examples (generated from the previous grammar) illustrate:

I like apples. They are sweet, red and crunchy. Apples are good for you. I particularly like MacIntosh.

Oranges are nice. They are sweet. They are orange. They are firm. They are good for you. I especially like navel oranges.

Similar structures provide examples of justification (I like apples because...) or exceptions (I like apples even though...). Additional operations allow the formulation of questions based on the original paragraph, for comprehension testing and the like.

Clearly, the model shown here represents only a first step in a long process. In particular, despite the fact that they generate authentic-looking paragraphs, the models used here do not yet embody historically contingent information, as for example the fact that over the past half-century the genetic diversity of apple stocks has been seriously reduced by the growing and marketing of only

a few species such as MacIntosh. The optimal representation of such information will be a challenge for our model. Similarly, the range of logical relations dealt with is still small.

Despite this, the approach shown here has the advantage of allowing relatively easy manipulation of discourse elements in a controlled fashion. While the two examples shown are quite simple, the principles for producing much more complex structures are already present.

In the presentation, we will describe use of structures of the sort in a subject-testing environment in which paragraphs are shown on the screen, then caused to disappear, to be replaced by comprehension and other questions.

## References

- Hamburger, H. (1994) Foreign Language Immersion: Science, Practice and a System. *Journal of Artificial Intelligence in Education* 5/4:429–454.
- Hovy, D. (1991) Approaches to the Planning of Coherent Text. In Paris, Swartout and Mann, (eds.) *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Boston: Kluwer.
- Lehnert, W. et al. (1983) BORIS – An experiment in in-depth understanding of narratives, *Artificial Intelligence* 20:15–62.
- Lenat, D. (1990) Building large knowledge-based systems: representation and inference in the CYC project. Reading, Mass.: Addison-Wesley.
- Levison, M., Lessard, G. (1992) A System for Natural Language Generation, *Computers and the Humanities* 26:43–58.
- Mann, W.C. and Thompson, S.A. (1987) Rhetorical Structure Theory: Description and Construction of Text Structures. In Kempen, G. (ed.) *Natural Language Generation*. Dordrecht: Martinus Nijhoff.