

# **Interlingua for Bilingual Statistical Reports**

**By**

Benoit Lavoie

CoGenTex Inc.

840 Hanshaw Road, Suite 5

Ithaca, NY, USA, 14850

Tel: 1-607-266-0363

Email : benoit@cogentex.com

## **Abstract**

This paper describes the interlingua used in a text generation system applied to the domain of bilingual statistical reports (LFS; [Iordanskaja et al. 92]). This interlingua is a sentential conceptual interlingua ([Polguère 91]) that is transformed directly into English and French semantic representations ([Mel'cuk 81]) via declarative rules stored in an interlingual dictionary. However, this domain contains problematic cases illustrating that a direct transformation of the interlingua into semantic representations may not always be appropriate. Namely, some bilingual examples from the corpus show important differences at the level of the conceptual content or at the level of thematic structure. As an initial solution to these problems, I propose to consider the stylistic preferences associated with each target language in order to paraphrase (adjust) the interlingua for these languages.

## **Keywords**

Interlingua, Bilingual Statistical Reports, Conceptual Representation, Semantic Representation

## **1. Introduction**

This paper describes the interlingua used in the LFS prototype system ([Iordanskaja et al. 92, Kittredge 92, Lavoie 94]), a text generation system applied to the domain of bilingual

statistical reports. The use of a sentential conceptual interlingua ([Polguère 91]) allows the generation of reports which are generally similar in both content and style to manually written reports. This interlingua is transformed directly into an English and a French semantic representation via a declarative interlingual dictionary that allows the core text planner and sentence planner to be linguistically independent.

Other systems have already been developed for the domain of statistical reports: Ana/Frana ([Kukich 83]/[Contant 85]) and SemTex ([Rösner 87]). Like LFS, all these systems use a sentential conceptual interlingua. However, in these systems, contrary to LFS, no distinction is made between conceptual representation and semantic representation: the interlingua is transformed directly into syntactic representations. In LFS, the use of a semantic representation allows bilingual texts to be generated that can differ not only syntactically, but also semantically.

The structure of this paper is as follows. In section 2, I present the application domain. In section 3, I discuss the motivations regarding the choice of the interlingua. In section 4, I present the architecture of the system, specifying the role played by the interlingua in particular. In section 5, I show how the interlingua is transformed into linguistic representations. In section 6, I discuss some problematic cases associated with the interlingua in the LFS system. In section 7, I propose a partial solution for these problematic cases. Finally, section 8 contains the conclusions of the paper.

## **2. The Application Domain**

The overall application domain covered by the LFS system consists of bilingual summaries of Canadian statistical data. One of these statistical domains concerns reports on the labour force published monthly by Statistics Canada. These reports contain bilingual summaries concerning changes of indices such as employment, unemployment, unemployment rate, etc., among population groups, provinces and industries.

These summaries are characterized by a rigid macro-structure (cf. [van Dijk 1977]) allowing a uniform analysis in terms of sections and sub-sections concerning similar topics from one report to another or from one language to another. Moreover, the sentential organization of the report is generally the same in both languages.

The LFS system also covers two other domains of bilingual statistical reports, Retail Trade and Consumer Price Index, that have many similarities with the labour force domain. However, we will not discuss these domains in this paper.

### 3. Choice of the Interlingua

An interlingua can be roughly classified according to the following two parameters ([Polguère 91]):

- It is either textual or sentential; i.e. it can represent more than one sentence (textual interlingua) or a single sentence (sentential interlingua).
- It is either conceptual, semantic or syntactic.

From a theoretical point of view, the ideal interlingua is a textual conceptual interlingua, since it is the most general one; this interlingua allows the consideration of discourse and linguistic variations in the texts generated in different languages.

In practice however, the choice of the interlingua is often motivated by the application domain; for example, the FoG system ([Bourbeau et al. 90]), currently in use in industry, has a syntactic sentential interlingua that has until now appeared to be sufficient for its domain of weather forecasts. In the LFS system, a sentential conceptual interlingua is used, for the reasons explained below.

#### 3.1. Sentential Interlingua

The LFS system uses a sentential interlingua instead of a textual interlingua. Although this approach may be less general, there are several reasons for this choice:

- As in section 2, the discourse segments in the domain of bilingual statistical reports are generally characterized by similar sentential and discourse clusters in both languages.
- For the rare exceptions found in the corpus where text segments have different sentential organizations (cf. section 6), it is generally possible to paraphrase these text segments so that the organization will be the same.
- The linguistic framework used in the LFS system is the Meaning-Text Theory (MTT; [Mel'cuk 81]) which currently only considers sentential phenomena.

#### 3.2. Conceptual Interlingua

The LFS system uses a conceptual interlingua. From this interlingua, semantic representations for each target language are derived. The choice motivating the use of a conceptual representation as an interlingua is motivated by cases found in the corpus such as the following, reported in [Iordanskaja et al. 92]:

(1a) *The level of employment remained virtually unchanged*

(1b) *Le niveau d'emploi a peu varié.* [The level of employment changed little]

In this case, the sentences (1a) and (1b) are different both syntactically and semantically. Another example where the semantic representations differ is the following:

(2a) (...) *persons aged 25 and over*

(2b) (...) *personnes de 25 **ans** et plus* [(...) persons aged 25 **years** old and over]

In this case, the meaning of 'year' is present in the French expression but not in the English expression. In this last case, this meaning is recoverable only through the domain/contextual knowledge and not from the linguistic knowledge.

### 3.3. Formalism of the Interlingua

The formalism used to represent the interlingua in LFS is similar to that used to represent the conceptual representations in the Gossip system ([Carcagno and Iordanskaja 89]). The interlingua is a conceptual structure annotated with thematic information. The conceptual structure is a network of concepts linked with conceptual relations such as *agent*, *patient*, *object*, etc. By convention, the labels used for the concepts are derived from English expressions.

Figure 1 shows the interlingua and the semantic representations corresponding to the sentences (1a) and (1b) presented in section 3.2. Some explanations of the transformation of the interlingua into semantic representations are given in the next two sections.

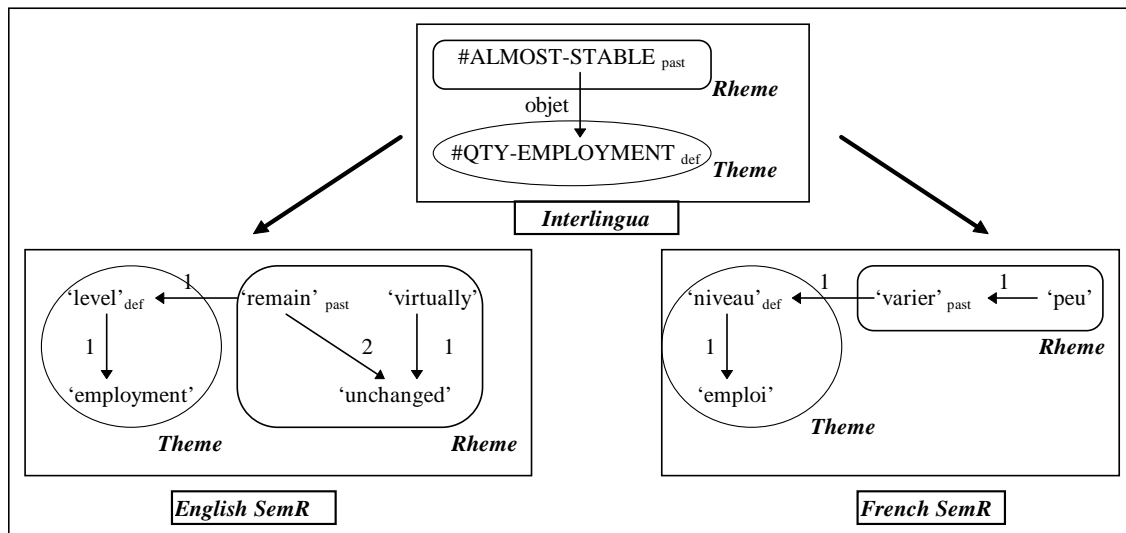
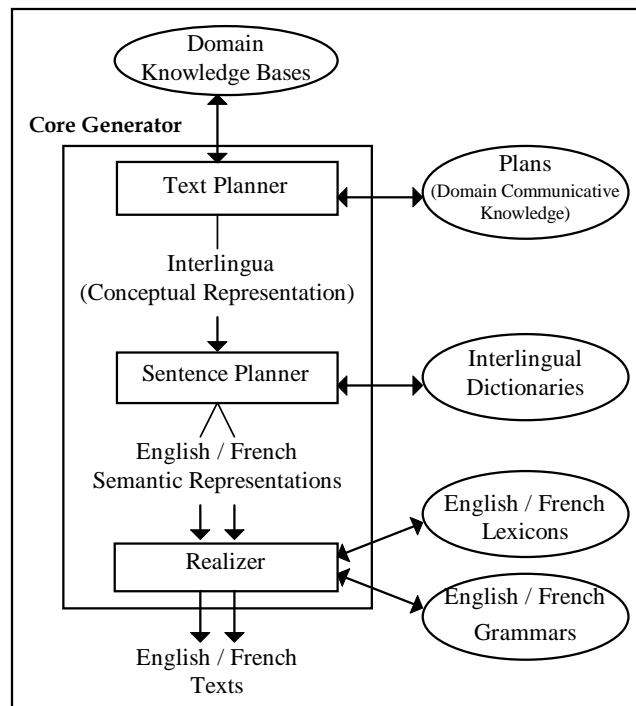


Figure 1: Example of Interlingua and corresponding Semantic Representations.

## 4. System Design

The design of the LFS system is shown in figure 2. It is similar to the design used by the Joyce system ([Rambow and Korelsky 92]) except for the following details:

- Joyce is a monolingual system, while LFS is a multilingual system.
- In Joyce, the conceptual representations produced by the text planner are translated directly into deep-syntactic representations, and not into semantic representations as in the LFS system.



**Figure 2: Global Architecture of the LFS System**

This type of design is widespread in the domain of text generation ([Reiter 94]). It is characterized by two features that facilitate maintenance and portability:

- A modular architecture.
- Knowledge resources represented declaratively and separated from the core generator.

For multilingual generation, the addition of a new target language in such a system implies modifications mainly to the language-specific knowledge bases: namely, the interlingual dictionary, the lexicon, and the grammars.

## 5. Transforming the Interlingua into Linguistic Representation

The transformation of the interlingua into semantic representations is done at the level of sentence planning, after the grouping of elementary conceptual representations into sentential conceptual representations. This transformation is done by recursive application of the mapping rules stored in the interlingual dictionaries.

### 5.1. Mapping Rules

The conceptual-semantic mapping rules describe how to transform a conceptual representation into semantic representations for the target languages. During these transformations, the concepts are generally replaced by one or more semes (semantic units), and the features of concepts (thematic, number, definiteness) are transferred. Figure 3a and 3b give some examples of mapping rules used to generate the sentences (1a) and (1b). In these rules, **Comm** represent thematic features, and **Gram** represents the features associated with the verb tense, number and the definiteness. Default values are associated with these features.

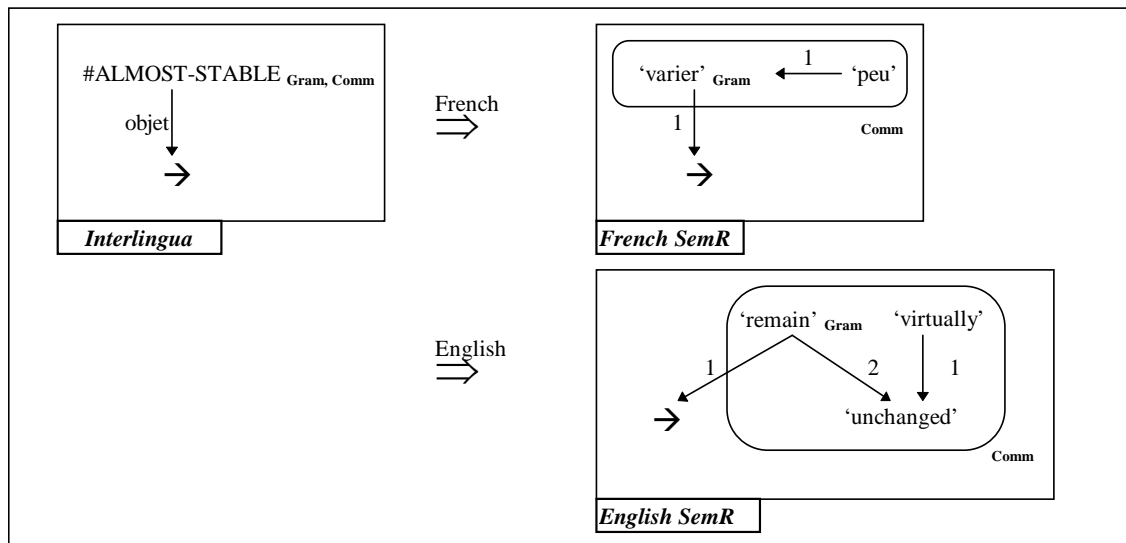


Figure 3a: Example1 of conceptual-semantic mapping rule

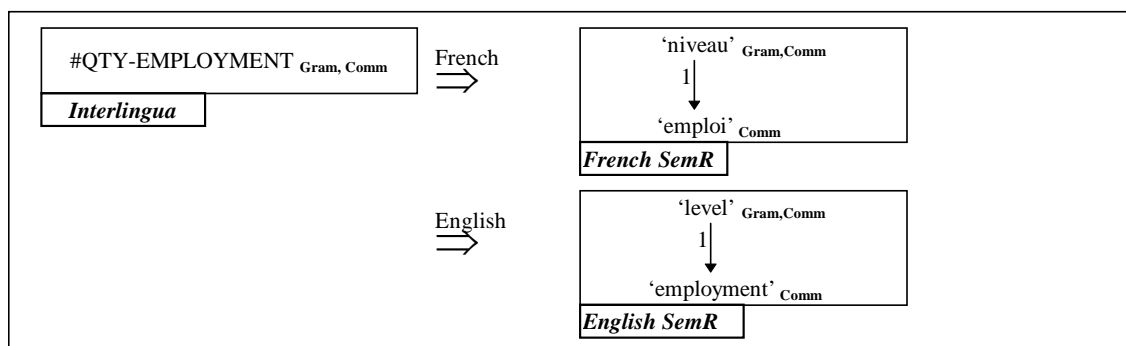


Figure 3b: Example2 of conceptual-semantic mapping rule

## 6. Problematic Cases

The sentential conceptual interlingua described in the previous sections allows the production of most of the sentences in the domain of bilingual statistical reports. However, there are some problematic cases.

**Case of discourse variation:**

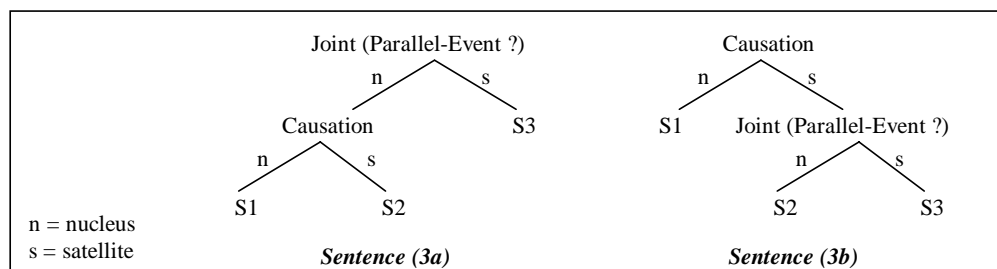
Consider the following excerpts:

(3a) (...) A similar increase in the number of persons entering the labour force S1 resulted in little overall change in unemployment S2. The unemployment rate remained unchanged at 7.6 S3.

(3b) (...) En raison d'une forte augmentation semblable du nombre de personnes joignant la population active S1, le nombre de chômeurs a peu varié S2 et le taux de chômage s'est maintenu à 7.6 S3.

(...) Due to a similar significant increase in the number of persons entering the labour force S1, the number of unemployed changed little S2 and the unemployment rate remained unchanged at 7.6 S3.

These excerpts, taken from the same bilingual report, illustrate a case that cannot be handled with a sentential interlingua, since the sentence scopes are different. Moreover, while the content of these texts is similar, these rhetorical structures appear to be different. In the French text (3b), the use of a conjunction can indicate a relation of causation between segment s1 on the one hand and both segment s2 and segment s3 on the other. In the English text(3a), there is a relation of causation only between the segment s1 and the segment s2. These two possible analyses are given in figure 4. (In these analyses, we are not sure about the nature of the joint relations. However, this does not affect the point of this discussion.)



**Figure 4: Two Possible Rhetorical Analyses for Sentences (3a) and (3b)**

Note that neither of the above-mentioned rhetorical analyses is a good representation of the reality! The numeric value of the *Unemployment Rate* is determined by the following percentage ratio: *Unemployment /Labour Force*. According to this equation, the real (discourse-independent) cause of the event mentioned in the segments S3 of the sentences (3a) and (3b) is the conjunction of the events mentioned in the segments S1 and S2 of these sentences.

**Case of conceptual variation:**

Consider the two following excerpts:

- (4a) Estimates from Statistics Canada's Labour Force Survey show little significant overall change in the labour market between August and September 1990. S1 While the employment level increased slightly, the number of persons who were unemployed also rose S3 and the unemployment rate edged up 0.1 to 8.4.
- (4b) Les estimations tirées de l'enquête sur la population active de Statistique Canada indiquent peu de changements sur le marché du travail en septembre 1990. S1 Alors que le niveau de l'emploi a légèrement augmenté, le niveau de chômage s'est accru lui aussi S3 et le taux de chômage a monté de 0.1 pour s'établir à 8.4.

The English text (4a) is conceptually distinct from the French text (4b) in two ways:

- The text segment S1 in (4a) specifies the complete period of variation (**between August and September 1990**), while the text segment S1 in (4b) specifies only a part of this period (**en septembre 1990**).
- The text segment S3 in (4a) uses the definition of the concept *unemployment-level* (**number of persons who were unemployed**), while this is not the case for the segment S3 in (4b).

The text segments (4a) and (4b) are conceptual paraphrases; they mean the same thing **in the domain of the labour force**. However, they are not semantic (linguistic) paraphrases. Hence, a direct transformation of the conceptual interlingua into French and English semantic representations is inappropriate in this case.

#### **Case of thematic variation:**

Consider the two following excerpts:

- (5a) The estimated level of employment for persons aged 15 to 24<sup>Theme1</sup> rose by 18,000<sup>Rheme1</sup> and their employment / population ratio<sup>Theme2</sup> advanced 0.5 to 59.8.<sup>Rheme2</sup>
- (5b) L'estimation de l'emploi<sup>Theme1</sup> s'accroît de 18,000 chez les personnes de 15 à 24 ans<sup>Rheme1</sup> et leur rapport emploi-population<sup>Theme2</sup> avance de 0.5 pour s'établir à 59.8.<sup>Rheme2</sup>

In this case, it is the thematic structure that is different; in (5a), the complement **for persons aged 15 to 24** is used as part of the theme, while the equivalent expression in (5b), **chez les personnes âgées de 15 à 24 ans**, is used as part of the rheme. However in both languages, both thematic positions are allowed.

## 7. Possible Solution

A possible solution to the problems mentioned in the previous section would be to adjust the conceptual interlingua according to the preferences associated with each target language by performing some paraphrasing transformations. Each resulting language-specific conceptual representation would then be transformed into the corresponding semantic representation. This is illustrated in figure 5. Note that the conceptual interlingua and the conceptual paraphrases adjusted for each target language are all found at the conceptual level; we do not intend to introduce a new level of representation between the conceptual level and the semantic level.

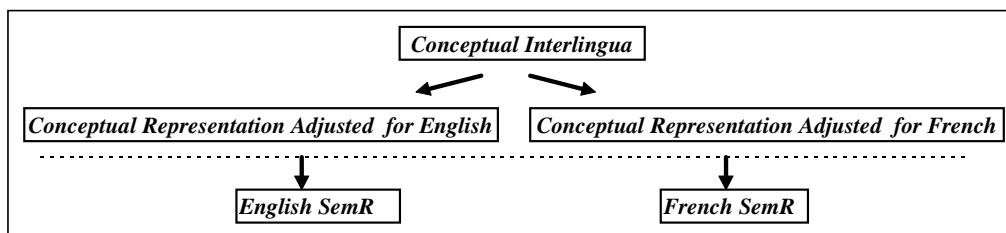


Figure 5: Use of Language-Specific Paraphrases of the Interlingua

The conceptual paraphrasing could consist in modifying the conceptual content (c.f. (4a) and (4b)) or in transforming the thematic structures (c.f. (5a) and (5b)). In the case where no preferences are specified for a target language, as is actually the case in the current implementation of LFS, a direct mapping of the interlingua to the corresponding linguistic representation could be done.

## 8. Conclusion

In this paper I have presented the interlingua used in a text generator for bilingual statistical reports. The direct mapping of the same sentential conceptual interlingua into the French and English semantic representations allows one to handle most of the cases of this domain. However, some cases are found in the corpus where this direct mapping is inappropriate, since the English and French texts have important differences either in conceptual content or in thematic structure. In order to handle these cases, a possible solution could be to use an intermediate language-specific level of conceptual representation, reflecting specific stylistic preferences in particular target languages.

## Acknowledgements

I thank David Caldwell, Richard Kittredge, Tanya Korelsky, Daryl McCullough, Owen Rambow, Ehud Reiter and Mike White for their comments on this work.

## References

- [Bateman et al. 93]  
Bateman, J., Degand, L. and Teich, E. (1993) "Towards multilingual textuality: some experiences from multilingual text generation", *Proceedings of the Fourth European Workshop on Natural Language Generation*, Pisa, pp.5-17.
- [Bourbeau et al. 90]  
Bourbeau, L., Carcagno, D., Goldberg, R., Kittredge, R. and Polguère, A. (1990) "Bilingual Generation of Weather Forecasts in an Operations Environment", *Proceedings of the 13th International Conference on Computational Linguistics*, vol.3, pp.318-320.
- [Carcagno and Iordanskaja 89]  
Carcagno, D. and Iordanskaja, L. (1989) "Content Determination and Text Structuring in Gossip", *Proceedings of the 2nd European Workshop on Text Generation*, Edinburgh, U.K., pp.15-21.
- [Contant 85]  
Contant, C. (1985) "*Génération automatique de texte: application au sous-langage boursier français*", Mémoire de maîtrise, Département de linguistique et philologie, Université de Montréal.
- [Iordanskaja et al. 92]  
Iordanskaja, L., Kim, M., Kittredge, R., Lavoie, B. and Polguère, A. (1992) "Generation of Extended Bilingual Statistical Reports", *Proceedings of the 15th International Conference on Computational Linguistics*, Nantes, France, pp.1019-1023.
- [Kittredge 92]  
Kittredge, R. (1992) "Bilingual Report Generation: Experience with Interlinguæ", *Aspects of Automated Natural Language Generation*, (R.Dale, E.Hovy, D.Roesner and O.Stock, eds.), Springer-Verlag, pp.297-299.
- [Kittredge and Polguère 91]  
Kittredge, R. and Polguère, A. (1991) "Dependency Grammars for Bilingual Text Generation: Inside FoG's Stratificational Models", *Proceedings of the International Conference on Current Issues in Computational Linguistics*, Penang, Malaysia, pp.318-330.
- [Kukich 83]  
Kukich, K. (1983) *Knowledge-Based Report Generation : A Knowledge-Engineering Approach to Natural Language Report Generation*, Ph.D.Thesis, Department of Information Science, University of Pittsburgh.
- [Lavoie 94]  
Lavoie, B. (1994) *Étude sur la planification de texte: application au sous-domaine de la population active*. Mémoire de maîtrise, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal.
- [Mel'cuk 81]  
Mel'cuk, I. (1981) "Meaning-Text Models", *Annual Review of Anthropology*, vol.10, pp.27-62.

[Polguère 91]

Polguère, A. (1991) “Everything has not been said about interlinguæ: the case of multi-lingual text generation system”, *Proceedings of Natural Language Processing Pacific Rim Symposium*, Penang, pp.318-330.

[Rambow and Korelsky 92]

Rambow, O. and Korelsky, T. (1992) “Applied Text Generation”, *Proceedings of the 6th International Workshop on Natural Language Generation*, Trento, Italy, pp.40-47.

[Reiter 94]

Reiter, E. (1994) “Has a Consensus NL Generation Architecture Appeared, and is it Psycholinguistically Plausible?”, *Proceedings of the 7th International Workshop on Natural Language Generation*, Maine, pp.163-170.

[Rösner 87]

Rösner, D. (1987) “The Automated News Agency: SEMTEX — a text generator for German”, *Natural Language Generation*, (G. Kempen, ed.), Martinus Nijhoff Publishers, Boston, pp.133-148.

[van Dijk 77]

van Dijk, T. (1977) *Texts and Contexts : Exploration in the semantics and pragmatics of discourse*, Longmans, London.