Interlingual Modelling: An Applications Perspective

Richard I. Kittredge

Applications in both natural language generation (NLG) and machine translation (MT) provide an opportunity to use linguistic representations which simplify comparisons between languages to essential (usually, semantic) aspects. Meaning-Text models (MTMs) provide a range of choices for interlingua during NLG, and for transfer level during MT, but require some extensions. We review a few advantages and limitations of MTMs as they have been used in recent application systems. While not always advantageous in practice, MTMs can become an increasingly interesting option in more applications through the use of better tools and methods for acquisition and maintenance of linguistic resources.

1 INTRODUCTION

The past two decades have seen the emergence of several applications of natural language generation, including recent commercial ones, in specialized domains ranging from stock market summaries to project status reports. Some debate has taken place in the computational linguistics community, and even more among users, over the need for detailed linguistic modelling in these applications. On the one hand, early systems for stock market reporting (Kukich, 1983) and weather forecasting (Kittredge et al., 1986) used phrasal techniques with minimal "shallow" linguistic representations to generate acceptable texts from input data. In contrast, later systems for weather texts such as FoG (Kittredge & Polguère, 1991; Goldberg et al., 1994) and for statistical summaries such as LFS (Iordanskaja et al., 1992) have used full-blown language models, incorporating syntactic and semantic representations in the course of synthesizing each output sentence. While research questions have sparked the interest in sophisticated langage models, the application of these models to practical problems is not always justified from the economic perspective.

First of all, it has become clear that some contexts for language generation (e.g., small domain, limited language variation, English-only output) can be handled adequately with shallow linguistic models. Even Exclass, a prototype generator for English and French job descriptions (Caldwell & Korelsky, 1994) managed with a categorial grammar language model because the English and French sentence fragments in the application domain were limited in structure, and

grammatical agreement problems (in French) were few. Shallow models, when adequate for the application, have also allowed extremely fast response times from the generator, as in the web-based ProjectReporter (www.cogentex.com/products/reporter). A third and very important consideration favoring shallow generation has been the knowledge barrier posed by the need to hand-craft rules for building full linguistic representations, and the incompleteness of the language models themselves. We return to this point in Section 3. Some of the arguments in choosing between shallow template-based NLG and the use of linguistic models have been summarized by Reiter (1995).

Despite the obstacles to using full-fledged language models in generation, these models appear to be our best long-term hope, when the application sublanguage is large (i.e., has long or complex sentences in great number and variety), requires multiple ways of expressing the same meaning, and necessitates bilingual or multilingual output. Moreover, the future of NLG will include an important role for speech synthesis output, and it is widely accepted that proper speech prosody (pitch, accent and timing) can only be generated with reference to a language model which has significant semantic and syntactic information.

Natural language generators have implemented a variety of linguistic models, but three have been widely discussed and tested in the NLG community: Systemic functional grammar, Tree-adjoining grammar and Meaning-Text models based on Meaning-Text Theory (Mel'cuk, 1988). The remainder of this paper is devoted to a brief review of the uses of MTMs in language generation, and the strengths and weaknesses of these models for NLG applications at the present time. (We will assume some familiarity with MTMs, based on other papers in this volume). Section 2 summarizes some strengths of MTMs for multi-lingual NLG and machine translation, with reference to some North American work over the past decade. Section 3 discusses some limitations of MTMs for NLG in particular and language processing in general, based on applications developed by the author and colleagues at CoGenTex. Section 4 surveys some possible directions for future work involving MTMs that would be of use in application systems.

2 SOME MEANING-TEXT ADVANTAGES FOR NLG AND MT

The usefulness of a Meaning-Text language model for language processing tasks has been known for some time (cf. Melcuk, 1988; Melcuk and Polguère, 1987). Some of most important features are:

- a stratified model which can simultaneously represent sentences at several levels: semantic, syntactic (deep and surface), morphological (deep and surface), phonological, and phonetic; (for some purposes and languages certain levels may be merged);

- a richly structured lexicon, the ECD, including lexical functions for interrelating lexemes,
- an explicit modelling of the human paraprase capability using mappings between the levels and the ECD, providing excellent linguistic coverage in a systematic fashion;
- an explicit representation of communicative structure (theme vs. rheme, etc.), giving control over phenomena involving lexicalization, word order, intersentential anaphora, etc.
- an explicit representation of features affecting speech, including prosody

We will illustrate some of these features below.

2.1 DsyntR used for Interlingua during English-French generation – FoG

MTMs were applied to bilingual NLG for the first time in the FoG system (Bourbeau et al., 1990; Kittredge & Polguère, 1991; Goldberg et al., 1994). FoG produces public and marine weather forecasts in both English and French, using time series of weather forecast data as the only input. The text generator design takes advantage of strong stylistic similarities between the parallel weather sublanguages, both in the way texts are segmented into sentences, and in the similar syntax for each sentence. For example, sentences (1) and (2), which are characteristic of this domain, have slightly different surface syntactic structures (SsyntRs), but can be given isomorphic deep syntactic representations (DsyntRs). Virtually all English and French "translation twin" sentences in the application domain have isomorphic DSyntRs, which allows FoG to use a common abstract DSyntR structure as interlingua for the output of the text planning stage. In the rare cases where English lexemes do not map one-to-one to French lexemes, the relationship is still a simple one (e.g., two English verbs map to the same French verb), so that no semantic representation (SemR) need be used in the generation process. Figures 1 and 2 show the respective DsyntRs for the two sentences.

(1) *Winds southwest 15 to 20 knots diminishing to light late this evening.*
(2) *Vents du sud-ouest de 15 à 20 noeuds diminuant à faibles tard ce soir.*

FoG determines sentence content and plans output sentences as if there were only one output language. A common interlingual representation for each output sentence is then mapped to DsyntRs with language-specific lexemes which are given to separate realizers to produce the final sentences for the two languages.

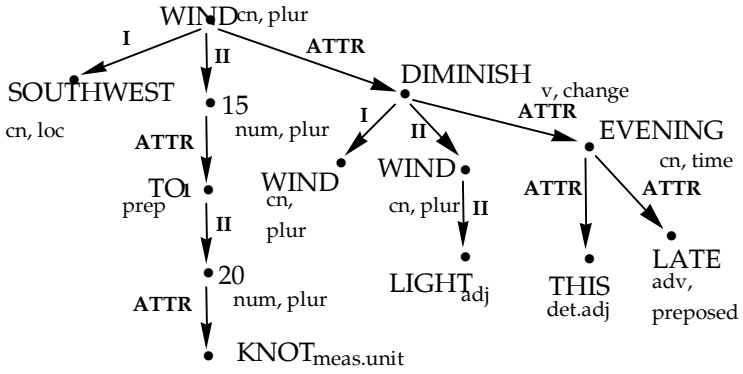

Figure 1 (English) DSyntR for (1)

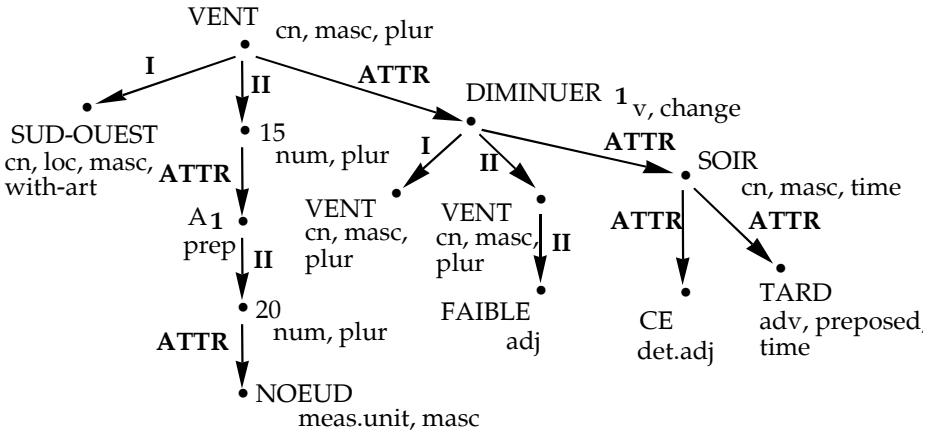

Figure 2 (French) DSyntR for (2)

2.2 Bilingual Realization from SemRs in LFS

In contrast to FoG's semantic simplicity, the LFS system (Iordanskaja et al., 1992) uses a much deeper semantic representation in the course of generating English and French statistical summaries from tabular data. In the great majority of cases, English and French translation twin sentences in this sublanguage can be represented by isomorphic SemR networks. However, in a few cases, the semantic content of the best human translation appears to be different from the source. Sentences (3) and (4) below illustrate a classic case,

for which the corresponding (simplified) SemRs are given in Figures 3 and 4 respectively. The divergence between English and French even on the semantic level in such cases has led to the use of a simplified conceptual interlingua for this sublanguage in the LFS system. Some application-specific mapping rules create the two SemRs from the conceptual interlingual structure, and each SemR is then mapped through the several representation levels into the output sentence using a MTM realizer for that language.

(3) *Employment remained virtually unchanged.*
(4) *L'emploi a peu varié.*

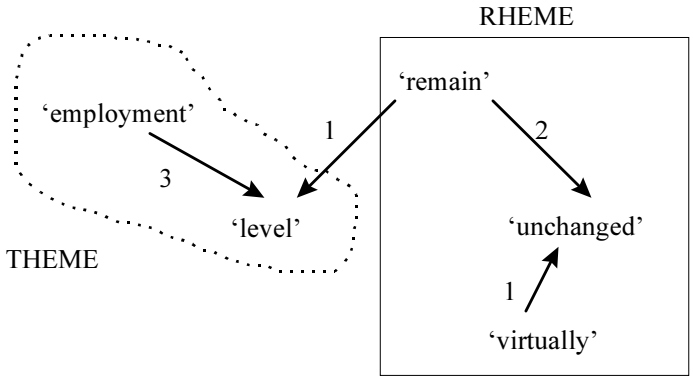

Figure 3. SemR for sentence (3)

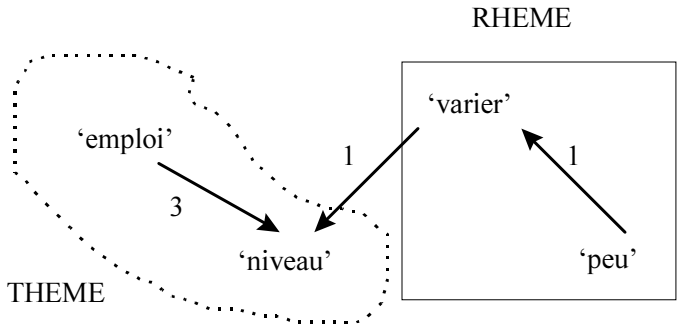

Figure 4. SemR for sentence (4)

It is worth noting that, within general English, one could find an acceptable paraphrase that has the semantic structure of the French sentence (4), i.e., (5), and likewise, sentence (6) would be an acceptable rendering in general French, if not in the LFS French sublanguage, of English (3).

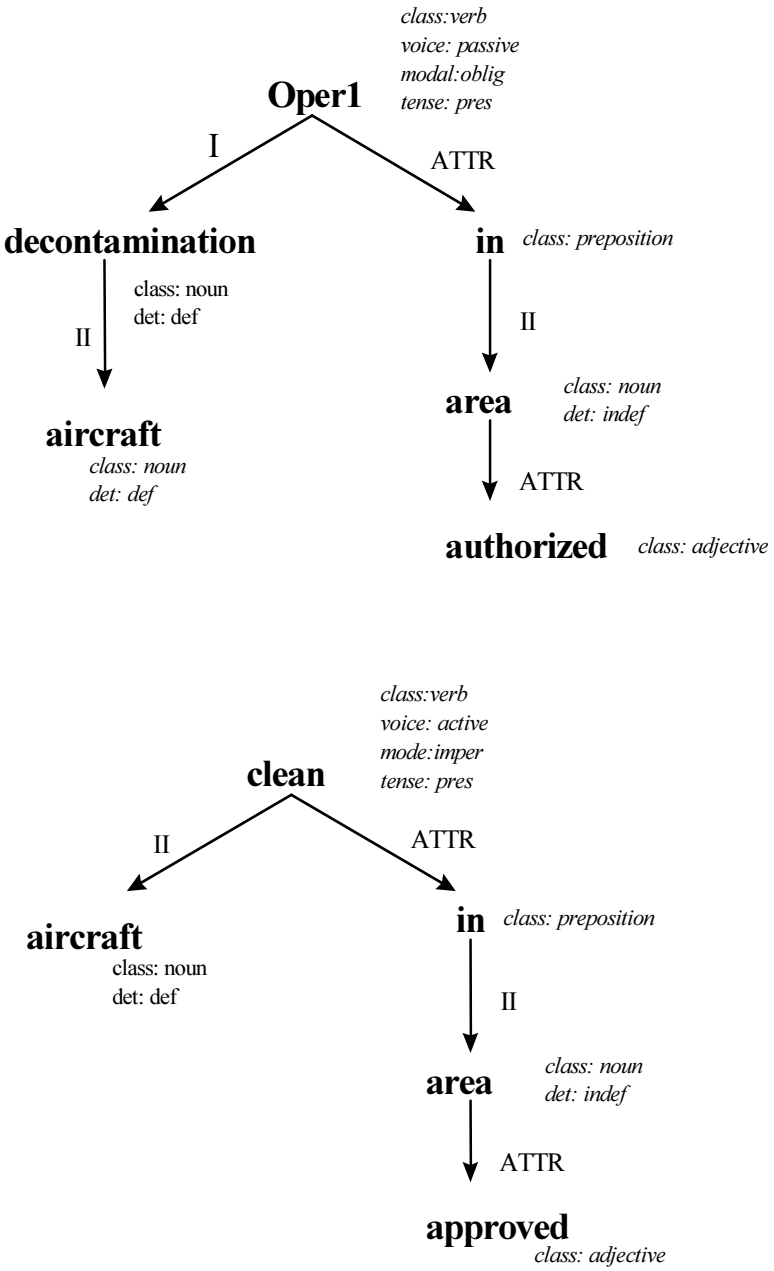

Figure 5. DsyntR trees for sentences (7) and (8), input and output to the CL revision mappings.

(5) *The employment level varied little.*
(6) *Le niveau d'emploi est resté pratiquement inchangé.*

Thus, it would have been possible to pick an interlingua resembling SemR for LFS, at the cost of occasionally generating "unprofessional" stylistic choices. The availability of a semantic level within our MTMs which reflects human paraphrase intuitions has made possible a choice of two different, but reasonable, interlingual solutions. (MTMs do not, however, offer much guidance in the matter of conceptual representations of domain knowledge.)

2.3 DSyntR used as transfer level in MT

Although the focus of this summary is on interlingual structures for generation, the related matter of interlingual machine translation (MT) is relevant. Indeed, when generating text in two or more languages, the choice of an interlingua is somewhat arbitrary in cases where there are divergences between the languages. One can, for example, use a DsyntR or SemR that is isomorphic to English to minimize the creation of a deep linguistic representation for that language, and derive the deep linguistic representation for French with a transformation from an English-like structure. This was actually done in one case in the FoG system. Thus the situation can be close to that of transfer-based machine translation, where analysis of the source language into something like a DSyntR or even a SemR for each sentence is followed by a cross-language mapping to similar structures in the target language. Syntactic-based transfer systems for MT, including "lexico-structural" transfer systems, have been in existence for at least 30 years, including the Montreal TAUM work, based on operator-argument grammar and phrase structure grammar, and work in Moscow by Apresyan and colleagues, which used MTMs for translation between Russian and English or French.

Since 1997, MTM-based machine translation has been under development at CoGenTex between English and Korean, based on earlier work between English and French (Palmer et al., 1998). In this bidirectional lexico-structural transfer system, DSyntRs of source-language sentences are mapped to corresponding DsyntRs of the target language.

2.4 DsyntR used as transfer level in CL revision

One of the active application areas in computational linguistics at the moment involves automatically checking and revising texts to conform to some standard of controlled language (CL), so that the resultant output can be easier to read by non-native speakers, and if needed, be more amenable to machine translation. Software designed to revise non-controlled human-authored text into text which conforms to CL standards must necessarily subject the source text to

grammatical analysis, and the CL revision problem has often been likened to machine translation between a natural language and a controlled version of the same language. Recently Nasr (1996) used an MTM for the revision of French sentences into CL form. An experimental implementation for English (Nasr et al., 1998) applies a sequence of mappings on DsyntRs to determine a DsyntR from which the RealPro generator (Lavoie and Rambow, 1997) can produce a CL-compliant sentence. Sentence (7) represents a non-controlled English sentence from a maintenance manual, while (8) is the final result of CL revision. The DsyntR trees for (7) and (8) are given in Figure 5.

(7) *The decontamination of the aircraft should be done in an authorized area.*
(8) *Clean the aircraft in an approved area.*

3 CURRENT LIMITATIONS OF MTMS FOR APPLICATIONS

In the application systems described above, MTMs have been chosen for their descriptive power to represent linguistic phenomena in a relatively clear, intuitive way. However, some of the advantages of MTMs in principle are directly related to their limitations in practice. We now turn to these.

First of all, MTMs "according to the book" provide so much descriptive detail that they require considerable time to build. For each lexical entry in the ECD, one must include a mass of detail, including the values of up to 60 lexical functions. This requirement, however, applies to the general language, and the differences between general language and sublanguage become evident in many application contexts. Indeed, few of the application systems described above make much use of lexical functions (LFs). Those that do (e.g., the LFS statistical summarizer) have so far used only a small subset of the possible ones, partly because the sublanguages in question restrict LF usage to a fraction of what a native speaker might use to express the same content in the general language. Moreover, each new sublanguage requires many specialized lexical entries and a few domain-dependent grammatical rules (for specialized mappings between descriptive levels). These cannot be taken from descriptive work on the general language, but need to be built from scratch.

Another drawback of MTMs, shared by other grammatical models, is their incompleteness of coverage of the general language. Here, MTT is more ambitious than many theories of language modelling in aiming to include communicative structure (CS) as an integral part of descriptions. Indeed, many problems of word order in generation cannot be resolved without CS markings (recall that the nodes of both DSynt trees and SSynt trees have no intrinsic order). A significant treatise on CS is only now on the horizon (Mel'cuk, forthcoming). Of course, some shortcuts (defaults) can be and are taken in building applications.

One area of concern is that many applications deal with paragraphs or much longer stretches of text beyond the single sentence. Many features of SemRs seem to be extendible for the representation of certain sequences of sentences. This will be necessary if we want to represent domains in which a single sentence in one language may be realized as two or more sentences in another. Even in the relatively simple sublanguages of economic statistics, occasional divergences in sentence scoping and communicative structure appear (cf. Lavoie, 1995), which call for some broader mechanism to show these equivalences prior to the realization of individual DsyntRs for each output sentence.

As with any sophisticated grammatical theory, Meaning-Text presents a considerable learning barrier to newcomers, including the computer scientists called upon to implement rule engines, dictionary maintenance software, etc. This situation would be aided if more attention could be given to stating explicit test criteria for making representational decisions in a highly pedagogical form. This would also facilitate the treatment of new and unusual phenomena which arise in each application, where experienced researchers may disagree on certain solutions (something that occurred on both FoG and LFS projects). Only in the past decade has there been significant movement of researchers and MTM resources between projects so that research insights can be compared and consolidated across the variety of languages being investigated (especially Russian, French and English).

Other difficulties in using MTMs can also be cited. In North America, at least, the predominance of phrase-structure grammars and phrasal categories in language theory shaped the whole tradition of computational linguistics until recently. Textbooks describing data structures for dependency trees and semantic networks (of the SemR variety) and their computer processing have not been widely available to students. These barriers have been overcome within individual projects, but a common culture of language processing based on these structures is only gradually taking shape today (see the recent workshop on dependency grammars at COLING-98).

4 CONCLUSIONS

This paper has aimed to illustrate the versatility of Meaning-Text models in a variety of language processing applications, especially involving interlingua, while drawing attention to their limitations and to the work remaining to be done for their wider use. On the one hand, the stratification of MTMs give two possible levels (DsyntR and SemR) for modelling interlinguae, or for carrying out language-to-language transfer. The ECD, coupled with deep-syntactic transformation rules, also makes possible certain operations required to convert non-controlled sentences to controlled form within a given language. On the

other hand, the complexity of MTMs, their incompleteness (especially in lexical description), and their relative unfamiliarity have conspired to inhibit wider usage. Thus there seems to be a relatively narrow spectrum of application problems where the language variation is sufficiently large to warrant using MTMs, but where the lack of descriptions is not too daunting to overcome in the lifetime of an application project. LFS was one such project, and other reporting domains (for example in generating sports summaries) may provide new opportunties to test and extend the models.

We can only hope that the growing number of researchers who are describing and applying MTMs will soon improve the pathways for exchanging resources and tools that will tip the balance towards their use in new areas.

REFERENCES

Bourbeau, L., D. Carcagno, E. Goldberg, R. Kittredge and A. Polguère, 1990. *Bilingual Generation of Weather Forecasts in an Operations Environment*. Proc. of COLING-90.

Caldwell, D. and T. Korelsky, 1994. *Bilingual Generation of Job Descriptions from Quasi-Conceptual Forms*. Proc. of the 4th Conference on Applied NLP.

Coch, J., 1996. *Overview of AlethGen*. Proc. of the 8th International NLG Workshop.

Goldberg, E., N. Driedger and R. Kittredge, 1994. *FoG: A New Approach to the Synthesis of Weather Forecast Text.* IEEE Expert, Vol.9, No.2, pp.45-53.

Helmreich, Stephen and David Farwell, 1998. *Translation Differences and Pragmatics-Based MT*. Machine Translation, vol.13, no.1, pp. 17-39.

Iordanskaja, L., M. Kim, R. Kittredge, B. Lavoie and A. Polguère, 1992. *Generation of Extended Bilingual Statistical Reports*. Proc. of COLING-92.

Iordanskaja, L., R. Kittredge and A. Polguère, 1991. *Lexical Selection and Paraphrase in a Meaning-Text Generation Model*. In C. Paris, W. Swartout and W. Mann, editors, Natural Language Generation in Artificial Intelligence and Computational Linguistics, pp.293-312, Kluwer.

Kittredge, Richard, 1995. *Efficiency vs. Generality in Interlingual Design: Some Linguistic Considerations.* Notes of the IJCAI-95 Workshop on Multilingual Text Generation, Montreal, pp.64-74.

Kittredge, R. and A. Polguère, 1991. *Dependency Grammars for Bilingual Text Generation: Inside FoG's Stratificational Models*. Proc. ICCICL Conf., Penang, pp.318-330.

Kittredge, R., A. Polguère and E. Goldberg, 1986. *Synthesizing Weather Forecasts from Formatted Data*. Proc. COLING-86.

Kittredge, R., L. Iordanskaja and A. Polguère, 1988. *Multilingual Text Generation and the Meaning-Text Theory*. Proc. of the 2nd International Conf. on Theoretical and Methodological Issues in Machine Translation of Natural Languages, pp.1-12.

Kukich, Karen, 1983. *Design of a Knowledge-Based Report Generator*, Proc. 21st Annual Meeting of the ACL.

Lavoie, Benoit, 1995. *Interlingua for Bilingual Statistical Reports*. Notes of the IJCAI-95 Workshop on Multilingual Text Generation, Montreal, pp.84-94.

Lavoie, B. and O. Rambow, 1997. *A Fast and Portable Realizer for Text Generation Systems*. Proc. of the 5th Conference on Applied NLP, pp.265-268.

Mel'čuk, Igor, 1988. Dependency Syntax: Theory and Practice. State University of New York Press, Albany.

Mel'čuk, Igor, forthcoming. Communicative Organization of Natural Language.

Mel'čuk, Igor and A. Polguère, 1987. *A Formal Lexicon in the Meaning-Text Theory*. Computational Linguistics, Vol.13, Nos.3-4, pp.261-275.

Nasr, Alexis, 1996. Un modèle de reformulation automatique fondé sur la Théorie Sens-Texte: Application aux langues contrôlées. PhD thesis, Université Paris 7.

Nasr, A., O. Rambow and R. Kittredge, 1998. *A Linguistic Framework for Controlled Language Systems*, Proceeding of CLAW-98, pp.145-158.

Palmer, M., O. Rambow and A. Nasr, 1998. *Rapid Prototyping of Domain Specific MT Systems*. Machine Translation and the Information Soup. Springer, pp.95-102.

Reiter, Ehud, 1995. *Templates vs. NLG*. Proceedings of the 5th European Workshop on NLG, pp.95-105.

Vander Linden, Keith and Donia Scott, 1995. *Raising the Interlingual Ceiling with Multilingual Text Generation*. Working Notes of the IJCAI-95, pp.95-101.

Richard I. Kittredge is Professor of Linguistics at the University of Montreal. He is on the editorial boards of the journal Machine Translation, and the Benjamins book series in Natural Language Processing. In 1990 he founded CoGenTex, a research and development company which specializes in NLG and related technologies. He can be reached at kittredg@iro.umontreal (see also www.cogentex.com/people).