# SUBLANGUAGE ENGINEERING IN THE FOG SYSTEM

**Richard Kittredge†‡**
†Department of Linguistics and Translation
University of Montreal
Montreal, Quebec
CANADA H3C 3J7
kittredg@iro.umontreal.ca

**Eli Goldberg**
Environment Canada
4905 Dufferin
Downsview, Ontario
CANADA M3H 5T4
goldberge@aestor.dots.doe.ca

**Myunghee Kim**
‡CoGenTex, Inc.
5911 rue Dolbeau
Montreal, Quebec
CANADA H3S 2G1
myunghee@cogentex.qc.ca

**Alain Polguère**
Department of English Language and Literature
National University of Singapore
10 Kent Ridge Crescent
SINGAPORE 0511
ellalain@leonis.nus.sg

## Abstract

FoG currently produces bilingual marine and public weather forecasts at several Canadian weather offices. The system is engineered to reflect "good professional style" as found in human forecasts. However, some regularization and simplification of the output has been needed. Sublanguage engineering issues include trade-offs in coverage and style, handling variation and evolution of sublanguages, "legislating" lexical semantics and assuring a language model able to accomodate new text types and support spoken output.

## 1  Background and System Overview

FoG (for Forecast Generator) was developed during 1985-89 (Kittredge et al., 1986; Bourbeau et al., 1990). After tests at Environment Canada during 1989-91, FoG entered regular use during 1991-92, first for marine forecasts, and more recently for public forecasts. Forty percent of the operational marine forecasts (roughly half of all marine forecast text) in Canada is now produced using FoG.

Meteorologists have been very receptive to using FoG, which is now a "back-end" facility of the FPA graphics workstation. The FPA supports the graphical analysis of weather while providing the rule-based concept formation needed to drive both text generation and non-linguistic applications. Meteorologists now concentrate on weather analysis and give less thought to how forecasts should be verbalized. Still, it has taken much time and effort to fit text generation into their work environment, and respond to new requirements. Operational experience has shown that some linguistic refinements first proposed during design were of low priority to users, compared with other features which were not originally anticipated.

Early work on FoG set up a specialized sublanguage grammar for marine forecasts, based on analysis of more than 100,000 words of archived English text. Corpus analysis of each forecast type has been supplemented by interviews and other means to check validity of categorizations and rules. In contrast to earlier work on machine translation (TAUM-METEO), where sublanguage grammars had to be relatively complete to recognize each possible human input text, generation of forecasts from concepts provides an opportunity to "engineer out" infrequent words and sentence patterns as long as each intended text content coming into the text planner is sayable in good quality text. The simplification and regularization of sublanguage grammars raises questions for engineering design. We are no longer just engineering the system to fit the sublanguage, but also engineering the output sublanguage itself to achieve goals such as simplicity (without significant loss of expressiveness) and clarity.

Different types of forecasts have differences in word usage, grammatical patterns and text structure, but the similarities are strong enough for them to be treated withing the same grammatical framework. This means that they require different text planners (and lexicons), but can use similar grammatical realizers. Significantly, English and French forecasts issued in Canada use the same principles for determining sentence boundaries, ordering and combining clauses, and formatting the output text. This greatly simplifies the problem of bilingual forecast generation in FoG.

FoG uses three major stages to compose forecasts: (1) graphically mediated content determination, (2) text planning resulting in interlingual forms, and (3) realization of English and French texts from the interlingua. Details are given in (Kittredge and

Polguère, 1991) and (Goldberg et al., 1994).

## 2 Sublanguage Engineering Issues

Early corpus analysis of marine forecasts identified a few kinds of information which were being conveyed with high frequency, as well as a "mixed bag" of phenomena of much lower frequency (e.g., WINDS HIGHER IN FJORDS and FOG LIFTING AS WINDS GRADUALLY INCREASE in Arctic regions). It was decided not to generate those sentence types requiring deep meteorological reasoning, to avoid high implementation cost. Over time, however, there has been some pressure to convey low frequency information which has significant value for marine safety (e.g., unexpectedly high winds in Arctic fjords). Use of a corpus has facilitated bringing low-frequency problems to the attention of system builders and users, so that deliberate design decisions can be made before the system is implemented.

An early goal in FoG was to generate text with stylistic variation by making use of paraphrase alternatives. Text generation typically provides an opportunity to introduce paraphrase variation, although the traditional problem has been finding ways of choosing from among the possible alternatives (Iordanskaja et al., 1991). However, many instances of apparent free variation turned out to have a tendency toward contextual determination, and it appeared easiest to build these tendencies into strict rules. In other cases individual forecasters voiced a clearcut preference for one variant form, which was subsequently implemented as the unique choice, at least for a given weather centre. The final result was the elimination of paraphrase from the generator. It is not clear that this is optimal, but it has simplified the design and implementation process during a phase when forecasters felt that there were more urgent problems.

One of the surprises in the development of FoG has been the constant evolution of language usage initiated by forecasters. New phenomena are being introduced (e.g., ultraviolet radiation warnings), other phenomena are de-emphasized, and better ways are found to say the same thing. The reasons for this have been quite varied and often specific to a given forecasting office and its client community. The constant "drift" of sublanguage usage at individual forecasting sites has led to maintenance of local variant systems. The flow of change requests has confirmed the need to keep the components of FoG in their most declarative and transparent form for easy maintenance.

Early work showed cases where roughly synonymous words turned out to have somewhat different fuzzy semantics. For example winds can both "strengthen" and "increase", but the former term tends to be used with high wind speeds. We incorporated a strict separation rule by "legislating" a point on the wind speed scale to separate the two word definitions. In other cases, apparently random variation in usage by forecasters led to an attempt to introduce a reasonable set of criteria for choosing one variant form over another. It appears that the very idea of free variation in forecast wording is difficult for forecasters to accept, and this natural tendency actually makes life easier (but less interesting) for system designers.

Future extensions to FoG are planned, including new forecast types (e.g., technical synopses) and an option for synthesized speech output, building on the existing linguistic model. We would also like to generate forecasts in languages such as Inuktitut, but this may require a deeper interlingual representation, such as the semantic net already used in other applications (Iordanskaja et al., 1991). However, languages like Inuktitut also use different conceptualizations of the weather than English and French, which might go beyond the capabilities of the FPA.

Recent attempts to produce spoken forecasts with concatenated speech techniques or commercial text-to-speech output devices suffer from a lack of good prosody. FoG's Meaning-Text language model provides for explicit prosodic structure, percolating from the interlingual representation to a new phonetic representation level. Contrastive stress will come from text planning, while most other features affecting pitch will come from surface syntactic specifications.

## References

L. Bourbeau, D. Carcagno, E. Goldberg, R. Kittredge and A. Polguère. 1990. Bilingual Generation of Weather Forecasts in an Operations Environment. In *Proc. of COLING-90*, v.3, pp.318–320, Helsinki.

E. Goldberg, R. Kittredge, and N. Driedger. 1994. FoG: A New Approach to the Synthesis of Weather Forecast Text. In *IEEE Expert (Special Track on NLP)*, April 1994.

L. Iordanskaja, R. Kittredge et A. Polguère. 1991. Lexical Selection and Paraphrase in a Meaning-Text Generation Model. In *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, (C. Paris, W. Swartout et W. Mann, eds.), Dordrecht: Kluwer Academic Publishers, pp.293–312.

Richard Kittredge and Alain Polguère. 1991. Dependency Grammars for Bilingual Text Generation. In *Proc. of the Int'l. Conf. on Current Issues in Comp. Linguistics*, pages 318–330, Penang.

R. Kittredge, A. Polguère and E. Goldberg 1986. Synthesis of Weather Forecasts from Formatted Data. In *Proc. of COLING-86*, pp.563–565, Bonn.