Multi-Lingual Text Generation and the Meaning-Text Theory

Richard Kittredge, Lidija Iordanskaja, Alain Polguère

Odyssée Recherches Appliquées, Inc.
1290 rue Van Home, suite 400
Montreal H2V1K6, Canada

ABSTRACT

We describe multi-lingual text generation as an alternative to automatic translation in specified technical sublanguages, illustrating the notion with the implemented RAREAS-2 system for synthesizing marine weather forecasts in English and French. We then review the Meaning-Text Theory (MTT) of Mel'cuk et al. as we have applied it to text generation in the GOSSIP system for producing English reports about computer operating systems. The experience gained from these two systems suggests possible approaches to multi-lingual generation using MTT.

Acknowledgments

Work on the GOSSIP system is supported by the U.S. Air Force, Rome Air Development Center, under contract F30602-86-C-0115. Work on RAREAS-2 was supported by the Atmospheric Environment Service, Environment Canada, under contract no. KM191-6-6533/01-SE.

1 Introduction

Until recently, the idea of approaching problems of machine translation (MT) through language generation may have seemed dubious. Throughout the history of MT, the accent has usually been on language understanding as the most difficult part of translation. Transfer problems have provided a secondary focus of research, particularly among practitioners of syntax-based indirect translation. But the generation (synthesis) component of translation systems has typically taken a back seat to the analysis and transfer components. And this has been quite natural, since many potential problems for language generation, such as the determination of clause boundaries and the ordering of sentences, clauses and phrases have scarcely arisen in systems where language transfer takes place at a superficial level.

But now that the research focus has shifted towards deeper semantic analysis and the use of semantic interlingua, a greater burden for creating correct output text has been shifted towards the generation component. Among some MT researchers, notably those at CMU [17], there has been an active integration of results and approaches from language generation research into the design of knowledge-based MT systems. Still, the overall goals of such projects typically require a prior concern with the general problem of language understanding.

The difficulty of the language understanding problem in its full generality has led most MT researchers to introduce some simplifications. Typically, a restriction on the domain allows the use of simplified or reduced lexica, grammars and knowledge bases associated with the sublanguage employed by domain experts. Additional simplification may be introduced by designing a system to accept only artificially "customized" language which consciously avoids ambiguous or otherwise problematic structures altogether. This second type of limitation, which presupposes significant control over the writing process, seems to be gaining in popularity, although not necessarily among those using artificial intelligence approaches to MT.

The temptation to simplify the language understanding problem is obviously leading us somewhere. It is leading us, ever so gradually, to explore the possibility of doing away entirely with automated translation of manually composed text in favor of automatic composition, starting from a single representation of content, of text in both the "source" and the "target" languages. We hasten to say that this is virtually impossible today in all but a small number of domains and situations. Not only must the domain and situation of language use be sufficiently constrained; there must also exist a machine-readable source of data which can serve as a basis for text

content. Nevertheless, the possible number of applications for multilingual text generation is bound to increase, and it may be worthwhile today to consider some of the factors working towards this growth of applications, and the tools needed to make multilingual generation a widespread reality.

In (§2) below, we look at some "easy" applications of bilingual text generation, including the synthesis of marine weather forecasts in English and French carried out by the RAREAS-2 system developed at Odyssée Recherches Appliquées in Montreal.

Then in (§3) we consider some of the problems posed by language generation in more complex domains, with less predictable content and text structure. We summarize our use of the Meaning-Text Theory (MTT) of Mel'čuk et al. [15,10] in the generation of reports on computer operating systems, and draw some conclusions (§4) for the use of MTT as a theoretical basis for multilingual generation.

2 Multi-Lingual Report Generation

The generation of texts in two or more languages from a common representation of content is not a recent notion. It was discussed as a possible alternative to machine translation during the initial design of the TAUM-METEO prototype in Montreal in 1974. Even at that time it was already a well-known (albeit futuristic) concept, which resurfaced from time to time during the following decade[1]. But to our knowledge it was not until the work by Kukich and Contant on generating stock market reports (based on prior linguistic work in Montreal) that the idea was given a realistic practical implementation. Kukich's ANA system of 1983 [7] used a rule base to select or compute significant facts from a database of half-hourly price quotations and related statistics for industrial stocks. After pre-linguistic operations on sentence content and ordering, ANA used a phrasal lexicon to encode messages (content representations) as parts of sentences. Soon after, in the FRANA system [2], Contant substituted a French linguistic component to encode the output of ANA's pre-linguistic modules. Although the ANA and FRANA systems could generate only one of the paragraphs typically found in stock market reports (concerning trends and volume of the industrial average), their output was so professionally idiomatic that it could not be distinguished from manually composed paragraphs. Bilingual report generation was a reality.

[1] cf. the work on on multilingual story summarization at the Yale AI Lab

2.1 RAREAS and RAREAS-2

In 1986 Kittredge and Polguère developed the RAREAS system [6] for generating Arctic marine weather forecasts from forecast data in professional forecaster's English. RAREAS was designed to accept the output of an expert system which computes the time and location of significant changes in the value of weather parameters including wind direction and speed, precipitation type, frequency and intensity, cloud cover and visibility. The bilingual version, RAREAS-2 [18], was completed in 1987 and delivered to Environment Canada for testing, extensions and implementation in regional weather offices.

Figure 1 gives the input formatted data for the generated English report fragment given as figure 2, and for the corresponding fragment in French, given as figure 3.

The formatted data identifies the Greenwich time of report validity, the data and area of interest, and then specifies the initial values for each important weather parameter (here using mnemonic labels). Subsequent changes in the value of a parameter are preceded by the number of hours until the forecast change. Additional weather parameters which are a function of the input parameters, such as dangerous wind and freezing spray conditions, are calculated by the pre-linguistic modules of the system.

Linguistic modules first calculate the values of significant semantic features of incipient lexical items, particularly concerning direction and degree of changes. For example, winds which shift in direction will be described lexically as *veering* or *backing* depending on whether the change in direction is clockwise or counterclockwise, respectively. Initial lexical instantiation uses the most precise term available in the lexicon. Subsequent segmentation into sentences may juxtapose clauses in such a way that lexical variation is desirable. Precise terms may then be replaced by synonymic variants or by more general (hyperonymic) lexemes.

```
2200 mon 83/09/22 end.
frob wind 220 30 &
            nt 5 300 35 & nt 18 speed 40
            wea rain cont heavy &
              nt 15 nl n 65 rain per moderate
            temp -3
            end.
```

Figure 1. Sample of RAREAS-2 formatted input

```
marine forecasts for arctic waters issued by environment
Canada
at 3.00 pm mdt monday 22 September 1983
for tonight and tuesday.

frobisher-bay
gale warning issued ...
freezing spray warning issued ...
winds southwesterly 30 veering and strengthening to
northwesterly
gales 35 late this evening then strengthening to northwesterly
gales 40 late tuesday afternoon.  cloudy with rain then
showers
developing north of 65 n latitude tuesday.  visibility fair in
precipitation.
```

Figure 2. English output for input of figure 1

```
previsions maritimes pour 1'arctique emises par environnement
Canada
a 15h00 har le lundi 22 septembre 1983
pour cette nuit et mardi.

frobisher-bay
avertissement de coup de vent en vigeur
avertissement d' embruns verglacants en vigeur  ...
vents du sud-ouest a 30 virant et se renforcant a coups de vents du
nord-ouest a 35 tard ce soir puis se renforcant a coups de vents du
nord-ouest a 40 tard mardi apres-midi.  nuageux avec pluie puis
averses commencant au nord de la latitude 65 n mardi.  visibilite
passable sons les precipitations.
```

Figure 3. French output for input of figure 1

2.2 Linguistic knowledge required for synthesis of forecasts

The RAREAS-2 architecture isolates different types of linguistic and non-linguistic knowledge within separate modules. Our grammatical, lexical, rhetorical and stylistic description was based on an examination of over 100,000 words of marine forecasts in English and French, for a broad selection of marine areas. Fairly detailed grammars were drawn up for the corresponding English and French sublanguages.

Linguistic knowledge for these forecasts consists of several types:

lexical semantics, including conditions for appropriate usage of words as a function of semantic configurations, particular data values, and word class co-occurrence restrictions;

frequency preferences among synonymous terms in the sublanguage of marine bulletins;

syntactic patterns, including the possible and preferred sentence patterns for expressing messages of given types; a second type of syntactic knowledge is embodied in the rules for deleting repeated sentence constituents when two or more propositions are fused into a single report sentence;

principles of text organization, specific to the variety of text being synthesized; clause ordering is a function of the relative saliency of different aspects of the content and of causal or temporal connections between meteorological events.

2.3 Restricted generality of the approach

Although the linguistic approaches used by the ANA/FRANA systems and the RAREAS systems are somewhat different, neither one employs a full-fledged linguistic model. Rather, these systems carry out fairly direct mappings from portions of content ("messages") to linguistically marked fragments of sentences. The relatively minor adjustments carried out on these fragments, as well as the mapping rules themselves, are quite domain-dependent. Such a direct approach to report generation is feasible (and efficient) only in sublanguages where the relationship between language structure and information structure is relatively transparent. The number of "natural" sublanguages, such as stock market summaries and weather forecasts, may be relatively small. But as the use of on-line databases increases,

there will be a growing opportunity to synthesize natural-sounding summaries in one or more languages. In the case of bilingual or multilingual communities which share the same databases, such as those in Canada, there is a particular impetus to reduce the costs and time delay entailed by human composition of text in the source language followed by human or mechanical translation into the target language(s). Even in the case of weather forecasts, which are translated in Canada by computer with a relatively high success rate, the occurrence of occasional untranslatable sentences or erroneous input requires that human revisers play a role in the information processing loop. In contrast, automatic generation of reports, by eliminating the unpredictability in human language production without compromising naturalness, facilitates the transfer of information with a potentially much smaller failure rate.

3 "Full-Fledged" Text Generation

In domains where the production of text does not follow stereotyped patterns with highly predictable content, the direct-mapping techniques of report generation do not suffice. For sublanguages significantly more difficult than those of weather forecasts and stock market reports one needs a linguistic model which captures the full human capacity for semantic paraphrase. As the informational and linguistic complexity and unpredictability of text increases, so do the constraints imposed on the final form of each individual text sentence. Only linguistic models which provide a full range of paraphrase options can work around conflicting constraints, or optimize the building of sentences over a set of preferences.

In this section we summarize some of the properties of the Meaning-Text Theory of Mel'čuk et al.[10,15](§3.1) and our use of an MTT model for the generation of texts in English dealing with the use of computer operating systems (§3.2). Such texts pose a wide variety of problems for generation. In the final section (§4) we touch on a few of the issues which our work has raised for the use of MTT and similar powerful linguistic frameworks when applied to multilingual text generation.

3.1 Some features of MTT

MTT describes the (bidirectional) mapping of linguistic meanings to texts through seven levels of representation, from semantic networks to surface phonetic representations, complete with prosodic markers. This rather extreme stratification of

linguistic phenomena allows each stage of the mapping to be stated simply while permitting the inclusion of a wider range of phenomena than is typically covered in linguistic models. For languages such as English the "interesting" representation levels are really three: semantic nets (SemR), deep syntactic trees (DSyntR), and surface syntactic trees (SSyntR). The two syntactic levels use dependency trees, with explicitly labelled grammatical labels on the arcs and lexemes (roughly, dictionary words) or grammatical words on the nodes.

MTT's semantic nets are used to represent decompositions of meanings which are justified on linguistic grounds. In particular, they reflect the intuitions of native speakers in paraphrasing complex lexical meanings with locutions involving simpler lexical meanings. The semantic labels which appear on network nodes are thus not presumed to be universal, since even the simplest lexical and grammatical meanings of languages tend to differ.

3.2 Text Generation using an MTT Model

3.2.1 Conceptual vs. semantic representations

Like most in the field, we view text generation as involving a non-linguistic planning stage in which the intended content of the text is derived from some external reality, perhaps using an expert system. But unlike many others, we assume that the representation of this content may be different from the meaning representation required for any given target language. Whereas the former is expressed in terms appropriate for the inferencing required in the underlying problem domain, the latter is ultimately dependent on the lexical and grammatical system underlying the paraphrase mechanism in the target language. In the case of text generation in technical sublanguages, the mapping from (conceptual) content representation to semantic nets may often be relatively straightforward. Furthermore, the semantic representations used in different languages to express the same conceptual content in parallel technical sublanguages may differ only slightly in comparison to the differences arising in more subjective domains.

3.2.2 Linguistic modules

When used generatively, MTT's linguistic modules map the networks of SemR to deep syntactic dependency tree structures (DSyntR) by a module of Semantic Rules. Deep dependency trees are mapped to surface dependency tree structures (SSyntR) by a module containing Deep Syntactic Rules. Our implementation maps SSyntR structures to a single morphological representation (MorphR) with a module of Surface Syntactic Rules, and completely dispenses with MTT's two phonetic levels, since these superficial levels are related by relatively trivial mappings for languages such as English.

We have made use of the fairly complete Surface Syntactic Rules for English which are given in[13]. Deep Syntactic and Semantic Rules for English exist in much more fragmentary form. Our implementation work has required developing some of these rules, as well as lexical entries, required for texts in our domain. Another important aspect of applying MTT to generation has been the design of algorithms for the efficient application of rules (which are left unordered in MTT), and for the resolution of conflicts between rules. For example, the conversion of semantic nets to dependency trees, in order to be computationally tractable, has required development of some (linguistic) principles to restrict search.

3.2.3 Implementation for generating operating system reports

The GOSSIP system (Generation of Operating System Summaries In Prolog) combines a direct implementation of MTT for sentence generation with a domain-oriented text planner. Since text planning "on first principles" appears inefficient and unreliable for our domain, we have opted for an approach using fixed plans. A database containing operating system audit information is monitored so that significant configurations (or regular time intervals) can trigger the choice of an appropriate text construction plan from a library. The plan specifies additional information which must be extracted from the audit data and how to convert the selected data into a "conceptual communicative representation" (CCR) [3]. The construction plan specifies a sequence of questions to be answered (goals to be satisfied) by the text.

The CCR for an entire text is a sequence of CCRs for individual messages (i.e., with answers to the questions in propositional form). An important aspect of each CCR is its "communicative structure", the specification of theme and rheme for the

given message. Our use of "theme" on the conceptual level is not unlike McKeown's "local focus" ([8]), but serves mainly to guide the subsequent linguistic generation process.

The communicative structure is preserved during the transition from CCRs to SemRs. It then is inherited by successive levels, restricting the choices allowed by mapping rules during the passage from SemR to MorphR, it helps dictate the choice of the root verbal lexeme in SyntR, the use of "topicalized" structures in SSyntR, and word order in MorphR. Figure 4 gives the communicative structure for a semantic net which encodes the meaning of sentences such as *System users ran compilers and editors during this time.*

GOSSIP is currently programmed in Quintus Prolog (version 2.0) running on a Sun 3 workstation. Figure 5 shows the Sun screen with a simple output text along with fragments of the deep and surface syntactic structures associated with the second sentence of the output text shown: *The users of the system ran compilers and editors during this time.*

4 Using MTT for Multi-Lingual Generation

The Meaning-Text Theory has been partially implemented in the French-Russian MT work of Apresyan [1], although analysis and transfer are limited to the surface syntactic (SSyntR) level of representation. Certain aspects of MTT have influenced the work of Nirenburg at CMU [17], particularly in the treatment of the lexicon. But our work at ORA on GOSSIP appears to be the first direct implementation of an MTT model for generation which uses the deeper levels of representation (SemR and DSyntR).

Our experience to date in implementing the MTT for English (by a multi-lingual implementation team) allows us to make observations about this model's adequacy mostly on the level of semantics, syntax and lexicon. MTT's semantic level has traditionally dealt with the amount of meaning that can be encoded in a single sentence. MTT therefore stops short of text structure. Any attempt to add text planning or rhetorical principles of text organization is really external to the theory. To the extent that languages differ in rhetorical organization, therefore, a separate mechanism must calculate the linguistic meaning representations which correspond, in each language, to the input (language-independent) content representation.

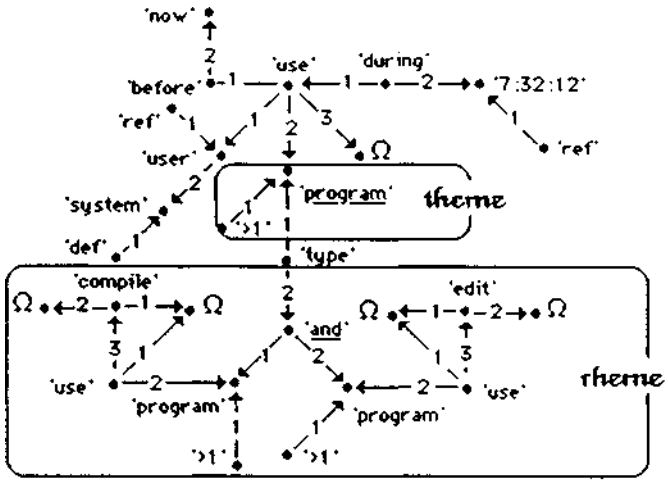

Figure 4.

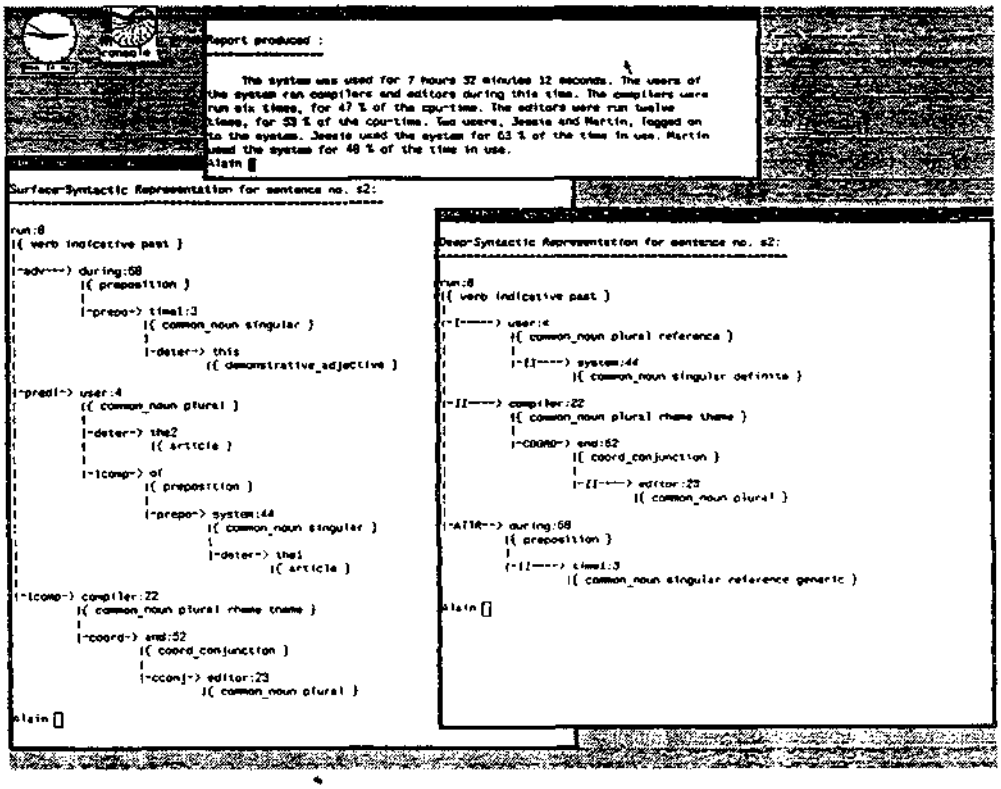

Figure 5.

MTT is particularly attractive in its power and flexibility to represent the way in which languages distribute elements of meaning differently in globally equivalent sentences. For example, the French sentence (2) is traditionally taken as a translation equivalent of English (1).

(1) John swam across the river.
(2) John a traversé la rivière à la nage.

A major redistribution of semantic material must take place during the standard translation from (1) to (2). moving the manner component of meaning from the predicate node to a sentence adverbial position, and incorporating the meaning of the English directional preposition *across* into the verbal meaning of the French *traverser = (Eng.) go across*. A detailed study of similar cases involving other verbs of motion provides support for the view that efficient language transfer must take place at a deep semantic level. Alternatively, bilingual generation of such sentences must begin to differentiate the grouping of semantic components at an early stage. MTT distinguishes a process of network reduction at the SemR level during generation which easily accommodates the differential grouping of "semes" in the two languages.

References

[1] Apresyan,Yu. (1985) "Linguistic Component of the French-Russian MT System ETAP-I" [in Russian] *Theory and models of knowledge (Theory and Practice of Creating AI Systems)* Scholarly Publications of Tartu University, v.714 (series on research in AI), pp.20-39.

[2] Contant, C. (1986) *Génération automatique de texte: application au sous-langage boursier,* M.A. thesis, Département de linguistique, Université de Montréal.

[3] Iordanskaja, L. and A. Polguère (1987) *Generation of Reports on the Activity of an Operating System Using Conceptual Communicative Representations,* technical memo, Odyssée Recherches Appliquées, Montréal.

[4] Iordanskaja, L., R. Kittredge and A. Polguère (1988) "Implementing a Meaning-Text Model for Language Generation" paper accepted for *Proc. COLING-88.*

[5] Kittredge, R. and I. Mel'čuk (1983) "Towards a Computable Model of Meaning-Text Relations within a Natural Sublanguage", *Proc. IJCAI-83,* pp.657-659.

[6] Kittredge, R., A. Polguère and E. Goldberg (1986) "Synthesizing Weather Forecasts from Formatted Data", *Proc. COLING-86,* Bonn.

[7] Kukich, K. (1983) "Design of a Knowledge-Based Report Generator", *Proc. 21st Annual Meeting of ACL,* Cambridge.

[8] McKeown, K. (1985) *Text Generation.* Cambridge University Press.

[9] Mel'čuk, I. (1974) *Towards a Theory of Linguistic Models of the Meaning-Text Type*.[in Russian] Nauka, Moscow.

[10] Mel'čuk, I. (1981) "Meaning-Text Models", *Annual Review of Anthropology,* vol.10, pp.27-62.

[11] Mel'čuk, I. (1982) "Lexical Functions in Lexicographic Description", *Proc. of the 8th Annual Meeting of Berkeley Linguistic Society.*

[12] Mel'čuk, I. (1984) *Dictionnaire explicatif et combinatoire du français contemporain.* Presses de 1'Université de Montréal.

[13] Mel'čuk, I. and N. Percov (1987) *Surface Syntax of English.* Benjamins.

[14] Mel'čuk, I. and A. Polguère (1988) "A Formal Lexicon", *Computational Linguistics* vol 13, nos.3-4. [Special Issue on the Lexicon].

[15] Mel'čuk, I. and A. Zholkovsky (1970) "Towards a functioning meaning-text model of language", *Linguistics,* vol.57, pp.10-47.

[16] Mel'čuk, I. and A. Zholkovsky (1984) *Explanatory Combinatory Dictionary of Modern Russian.* Wiener Slawistischer Almanach, Sonderband 14.

[17] Nirenburg, S. (1987) "A Natural Language Generation System that Emphasizes Lexical Selection", *Proc. Natural Language Planning Workshop,* Blue Mountain Lake.

[18] Polguère, A., L. Bourbeau and R. Kittredge (1987) *RAREAS-2: Bilingual Synthesis of Arctic Marine Forecasts*, technical report, Odyssée Recherches Appliquées, Inc., Montreal

[19] Zholkovski, A. and I. Mel'čuk (1967) "O Semantičeskom Sinteze" *Problemy Kibernetiki,* vol. 19. pp.177-238.