International Congress on
Terminology and Knowledge Engineering
29 Sept. - 1 Oct. 1987
University of Trier
Federal Republic of Germany

Victor Sadler

AI-DIRECTED INTERLINGUAL TERMINOGRAPHY IN TOMORROW'S MT SYSTEMS

Outline

0 Introduction: The DLT Project

This paper is concerned with terminological strategy in machine translation from the viewpoint of the DLT project. DLT (Distributed Language Translation) is a six-year MT project under development at BSO/Research (Utrecht, Netherlands) and funded on a fifty-fifty basis by the BSO software company and the Dutch government. The aim of the project is to produce a prototype English-to-French semi-automatic translation system for informative (including technical) texts by the year 1990. The system is designed for information distribution networks and personal desk-top equipment of the 1990's.

Ultimately, as its name suggests, DLT is intended to become a multilingual system, using an intermediate language (IL) and "distributing" the work load between SL-IL and IL-TL modules, which will typically be located in different places or countries. The main advantage, in terms of development costs, of an interlingual architecture in a multilingual system is well-known: the number of modules to be developed increases only linearly with the number of languages involved. One of the special characteristics of the system is its use of a human-language interface between the SL and TL modules. The language chosen for this purpose is a slightly modified form of Esperanto. In what follows the emphasis will be on the terminological consequences of this choice.

1 DLT and Terminography

By restricting its scope to 'informative' texts, the DLT system is excused in principle from coping with colloquialisms, puns and other features of creative writing. On the other hand, having the declared aim of translating material "ranging from technical instruction manuals to scientific literature abstracts and from business reports to nuclear waste disposal regulations" (8, p. x), DLT cannot step out of the way of one of the most labour-intensive aspects of MT systems: the building and maintenance of term banks. For this reason, aircraft maintenance handbooks were among the first test materials chosen for the development of the

prototype (Simplified) English to French system. This has involved some pioneering terminographic work in the IL. A trial demonstration of the DLT prototype is scheduled for December 1987, with technical texts from the chosen field as input.

2 Terminography in the IL

The general motivation of the choice of Esperanto as intermediate representation in the DLT system has been discussed elsewhere (5)(7)(8). In brief, Esperanto satisfies the following primary criteria:

a) The IL must have the same order of expressive power as the source and target languages;
b) It should be grammatically and semantically autonomous (and not defined in terms of one or other SL or TL);
c) It should facilitate automatic processing by a strong rule-based component;
d) It should also be user-friendly and readable, to facilitate development and maintenance.

However, it has been argued that the choice of Esperanto as intermediate language may have disadvantages as well as advantages. Hutchins (2, p. 289) points out the lack of technical vocabulary and terminological standardisation: "In effect, the project will be building an interlingual dictionary for international technical terms from scratch."

While it is true that Esperanto lacks the enormous corpus of scientific and technical writing available in the major languages, and English in particular, as well as their continually expanding vocabulary of terms, it should not be assumed that the language is completely undeveloped in this respect. On the contrary, a good deal of basic work has already been done, starting with pioneers such as Wüster, - particularly in such fields as informatics, mathematics, economics and chemistry - , and new glossaries are published regularly. In consequence, the position of Esperanto is similar to that of many other languages, particularly those of Third-World countries, in which there exists a basic scientific vocabulary in need of development and standardisation. DLT can thus avail itself of existing terminographical ground-work and of an international pool of experience in the scientific and technical application of its IL.

The initial experience in the field of aircraft maintenance may serve as an illustration. Preparatory to the trial translations scheduled for December 1987, a group of specialists competent both in Esperanto and in aircraft terminology were set the task of identifying or creating unambiguous Esperanto equivalents for 600 technical terms extracted from sample texts from aircraft maintenance handbooks. In the course of their work they decided to coin and define 3 new root morphemes. In 17 other cases they added new definitions to those of existing morphemes, thus introducing or increasing polysemy, although mainly of an interdisciplinary kind. In all the remaining 580 cases they were able either to identify existing Esperanto terms or to construct suitable terms from existing morphemes.

In the development of interlingual terminology in DLT, a choice has been made for the schematic, rather than the naturalistic, approach. Opting for schematic terms ensures that IL morphemes, which are invariable, are used as productively as possible in the interests of dictionary compaction. The choice reflects the importance, for an MT system, of maximizing the rule-base component and minimizing the lexical-entry component in the (virtual) dictionary. (Incidentally, it also echoes the historical development of Esperanto, in which naturalistic forms such as *administrator`o* have acquired an archaic flavour and tend to be replaced by regular agglutinative constructions such as *administr`ist`o.)*

The structure of Esperanto lends itself far more readily to term creation and systematisation, as well as to computer processing, than those of most other languages (1). There is no word class ambiguity. This, together with regular, exception-free derivation and valency inference and an exhaustive theory of word formation, makes the automatic synthesis and analysis of complex terms eminently feasible.

3 Knowledge-based (Term) Disambiguation

DLT is using Esperanto not only as the carrier of translated information between independent modules, but also to store Knowledge-of-the-World which, by definition, is language-independent. In the prototype system, the whole process of lexical disambiguation - a crucial stumbling-block in the history of MT (3) - is geared to a Lexical Knowledge Bank written in the IL. Alternative translations are matched against contextual models in the LKB and ranked in order of semantic appropriateness before a final selection is made for each sentence as a whole.

Needless to say, the process of building up a knowledge bank of this kind is expensive. This is why DLT policy has concentrated all the work of lexical disambiguation at the IL end of each module, making use of the same LKB in each SL or TL module. Here, full advantage can be taken of the special properties of Esperanto: its 100% agglutinative structure provides an explicit basis for the taxonomic organization of vocabulary and also allows surface variations to be automatically reduced to their lexical bones.

In this context, terminological disambiguation is merely a special case of the general problem, and the methods applied are the same. For example, the choice between the French *revêtement de l'avion* and *peau de l'avion* as possible translations of the English *aircraft skin* is ultimately determined by information stored in the (IL) Knowledge Bank, although triggered by pointers in the bilingual English-IL and IL-French dictionaries. The complex process of comparing contextual cues stored in the latter with the actual context of the sentence being translated is always performed on the IL representation of the sentence.

4 The Term Bank as Expert System

DLT is a *semi-automatic* translation system. An essential part of the design is the "disambiguation dialogue": a computer-initiated question-and-answer procedure to

eliminate source-language ambiguity vis-à-vis the IL representation, whenever attempts at automatic disambiguation remain inconclusive.

It has already been pointed out (8, p. III-104) that "DLT .. is an excellent platform for AI enhancements". Besides the special properties of the IL emphasized above, the interactive dialogue with an operator who in the future will more and more often be the author of the text offers an obvious incentive to treating the Lexical Knowledge Bank as an expert system (4).

The scientist or engineer who coins a new term while writing a report at some future DLT work station will be asked to specify its meaning. In the case of a monolithic word such as "quark", this will involve an active attempt at paraphrase or definition, which may in turn lead to further dialogue with the system. In the case of compounds or collocations, on the other hand, the system will attempt to analyze the combination and offer the most plausible results in descending order of probability, e.g. for an expression such as "quark drive":

1. Mechanism which drives a quark ?
2. Mechanism which drives by means of a quark ?
3. Driving mechanism powered by quarks ?
4. ...

- whereupon the writer can choose the intended meaning.

Obviously, the concept of the computer/writer dialogue is closely linked to that of machine learning. If no provision were made for self-improvement in the future DLT system, then questions of this kind would be repeated ad nauseam every time a new term was used. In fact, such paraphrases as the above will be generated automatically whenever the (knowledge-based) semantic analysis within the SL-IL module fails to establish a clear preference for one or other possible interpretation. They will be generated from alternative tree structures in the IL. This implies that the writer's choice will automatically link to the English term "quark drive" a specific IL representation which, stripped of brackets and labels, transforms to a linear string such as

per`kvark`a pel`il`o

where the implicit relation between the terms of the compound has been made explicit by means of the prepositional prefix "per". This IL representation (after possible compaction by the powerful word grammar module) is nothing less than an ad hoc technical term which will sooner or later be passed over a network and turn up in the input stream to the IL-TL module.

Whether the translation thus established will eventually be added to the SL-IL dictionary or will be treated as a one-off production and later discarded by some garbage collection routine is a matter for statistical and lexicographical heuristics to decide in the framework of general DLT software support policy (8, p. VI-27).

At the IL-TL module, the IL term may very well not be echoed by any specific term in the target language. In this case a paraphrase will be generated from the IL representation, using general metataxis rules for the language pair

concerned and once more calling on the word grammar module for the IL to break down into analytic dependency structures any complex synthetic forms (such as "per`kvark`a") not represented in the IL-TL dictionary.

On the other hand it is perfectly possible that a specific term already does exist in the target language, corresponding to the term which has only now been coined in English. In this latter case there must also already exist an IL entry in the IL-TL dictionary, similar, but not necessarily identical, to the input IL term. In the event of the IL entry not being identical, its equivalence can be tested with the aid of the IL word grammar and a set of transformation rules capable of identifying structurally different but semantically equivalent expressions. For example, an expression such as

per`komput`il`a tekst`o`pri`labor`o

("word processing") will be automatically matched with an alternative form such as

tekst`pri`labor`ad`o per komputor`o

if no exact equivalent is found in the IL-TL dictionary. It should be emphasized that the process described here for technical coinages is in no essential way distinct from the process to be applied to non-technical ambiguities. For example, an expression such as "police protection" may well require interactive disambiguation in a given context (protection of or by the police?), and the resulting (semantically more explicit) IL interpretation can equally be regarded as a "term" from the theoretical viewpoint. This principle ties in with the observation that there is no sharp boundary between technical terms and common-language words and that they must therefore be treated as parts of one single language system (6).

5 Machine-aided Terminography

While it is true that a great deal of terminographic effort will be required in order to equip DLT's intermediate language with the mass of terms it will need in a working system, the AI-directed design of the project offers considerable benefits for the terminographer. While we wait for a self-improving terminographic automaton, the interactive creation of terms can be greatly aided by the Lexical Knowledge Bank being built for DLT and by the kind of total-access search techniques being developed to interface with it. A few example applications:

- The system can support the generation of new terms by analogy. Just as the unaided human terminographer may look to analogous concepts for a suitable new term (e.g. E "space shuttle" --> D "Raumfähre" rather than "Raumweberschiffchen"), so the computer-aided DLT terminographer may request LKB search facilities to suggest logically or ontologically analogous concepts, asking for example "What are the sister concepts to *shuttle* in the conceptual taxonomy?" or "What words have a similar contextual pattern to *shuttle* ?" ("goes back and forth" etc.).

- Where multilingual terminology is already available, the DLT system can generate sets of possible IL translations from each NL in turn, calculate the best matches between the sets and offer the resulting IL terms for confirmation or editing. For example, starting with the English "outer wing tank test" and the French "essai des réservoirs externes voilure", and assuming that the individual words are already known to the system, we can require it to match all the possible IL translations of the highly ambiguous English string with those of the French term and rank the resulting match scores.
- Software tools can be provided which make use of the existing word grammar module to check the structural regularity of the proposed complex IL terms and display any potential ambiguities by means of paraphrases.
- Where the creation of new morphemes is preferred, the system can provide a check on accidental homonymy, definition consistency etc.

6 Conclusions

The kind of encyclopaedic Lexical Knowledge Bank being built up for the DLT system, written as it is in the IL, lies at the heart of the multilingual DLT architecture. Thus it will be equally accessible via a DLT work station from any source language for which the software has been developed. This fact, coupled with the AI-based inference and disambiguation processes inherent in the DLT design and the use of a computer-friendly and highly structured intermediate language opens new perspectives for international standardisation and automation in the field of terminography. The economic advantages of the interlingual architecture and of a central, monolingual knowledge bank are expected to far outweigh the cost of developing extensive terminology in the intermediate language.

Literature

1. EICHHOLZ, R. The creation of technical terms in Esperanto. (At this conference).

2. HUTCHINS, W.J. Machine Translation: Past, Present, Future. Chichester: Horwood, 1986.

3. MELBY, A.K. Lexical transfer: A missing element in linguistics theories. In: 11th INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS. Proceedings of Coling '86. Bonn: Institut für angewandte Kommunikations- und Sprachforschung, 1986, p. 104-106.

4. PAPEGAAIJ, B.C. Word expert semantics: an interlingual knowledge-based approach. V. Sadler / A.P.M. Witkam (eds.). Dordrecht / Riverton: Foris, 1986, 254 p.

5. SCHUBERT, K. Wo die Syntax im Wörterbuch steht. In: BURKHARDT, A., KÖRNER, K-H. (ed.). Pragmantax (Akten des 20. Linguistischen Kolloquiums, Wolfenbüttel/Braunschweig 1985). Tübingen: Niemeyer, 1986, p. 449-458.

6. SCHUBERT, K. Interlingual terminologies and compounds in the DLT project. In: Proceedings of the International Conference on Machine and Machine-Aided Translation, Birmingham, April 1986.

7. SCHUBERT, K. Linguistic and extra-linguistic knowledge. Computers and Translation 1 (1986), no. 3, p. 125-152.

8. WITKAM, A.P.M. Distributed Language Translation: Feasibility study of a multilingual facility for videotex information networks. Utrecht: BSO, 1983.