ELECTRONIC DICTIONARIES IN CANADA

Pierre Isabelle

Communications Canada
Canadian Workplace Automation Research Center
1575 Chomedey Blvd.
Laval, Quebec
CANADA H7V 2X2

November 1988

1. Introduction

In current usage, the term "electronic dictionary" (ED) is applied to a wide variety of tools, from the electronic bilingual pocket dictionary to the lexical components of the most sophisticated NLP systems. A broad distinction is usually drawn between human-oriented and machine-oriented ED's, according to whether the ED is meant to be used by a human or by an NLP system.

It seems, however, that the distinction between human- and machine-oriented dictionaries now tends to be somewhat deemphasized: many computational linguists believe that techniques can be developed that will make standard dictionaries play a useful role in NLP systems. Given that trend, our survey includes not only NLP-oriented ED's, but also some terminology banks (TERMIUM and BTQ), one computerized standard dictionary (NOED), and Igor Mel'cuk's theoretically-oriented dictionaries.

2. The TERMIUM terminology bank

The TERMIUM terminology data bank has been under ongoing development at the Canadian Department of Secretary of State since 1974. TERMIUM is made available to the translators and terminologists of the Canadian Bureau of translations, as well as to many other public and private translation services nationwide.

Each record in the bank is in principle associated with a unique concept, and searches proceed from a particular concept to the terms (mostly English and French) which designate it.. Records also typically include definitions, synonyms, abbreviations, sources, etc. Currently, TERMIUM comprises approximately 750,000 records associated with some 1.7 million terms.

The Department of Secretary of State is currently testing a CD-ROM PC-based prototype version of TERMIUM. The success of the prototype has led the Department to plan the development of an industrial version to be integrated into a projected translator's workstation. Over the years, TER-

MIUM has been an important source of data for the dictionary builders of several MT experiments in Canada. But so far, there has been no attempt to establish a more direct link.

3. The BTQ terminology bank.

The development of the Banque de terminologie du Québec (BTQ) started in 1972 at the Office de la langue française du Québec. To date, the BTQ contains approximately 900,000 records pointing to nearly 3 million terms. As with TERMIUM, both French and English terms are provided. In fact, the two banks are rather similar in their overall structure.

However, there are important differences. While the primary goal of TERMIUM was to provide terminology for the translators of the Bureau of Translations, the BTQ was rather aimed at providing French terminology for the workplace in Quebec. For this reason, the coverage of the two banks is largely complementary.

To my knowledge, no attempt has been made to use the BTQ in NLP systems. However, one of the current projects at the BTQ is to apply NLP techniques to the development of a *terminologists' workstation*.

4. The New Oxford English Dictionary project

A team of researchers at the University of Waterloo is cooperating with Oxford University Press on the production of an electronic version of The New Oxford English Dictionary. The NOED is to be turned into a full-fledged *textual database* (of some 600 million characters), structured around the concept of *p-string* ("parsed string").

The resulting product will facilitate ongoing updates. Moreover, it will be possible to have the machine extract from the NOED an incredible amount of information that is buried there, but has so far been out of reach given the lack of retrieval apparatus. For example, one will be able to ask the computer to exhibit the list of words whose first attested occurrence goes back to Shakespeare's works.

The first release of the electronic NOED, on CD-ROM, is scheduled for 1990. Of course, this product will in no sense be directly usable in current NLP systems. However, the electronic NOED project represents an important contribution to the goal of bridging the gap between "ordinary" and "machine-oriented" dictionaries.

5. The Explanatory Combinatorial Dictionary of French

Igor Mel'cuk, Alexander Zholkovsky and their colleagues have been working for some 20 years on a linguistic theory called the Meaning-Text model. The theory is aimed at formalizing paraphrase relationships within a language. A key element of the Meaning-Text model is the "explanatory combinatorial dictionary" (ECD). An ECD provides a particularly comprehensive description of the meaning and use of the words of a natural language.

An ECD entry includes a definition given in a form close to the natural language statements of standard dictionaries, but making use of *disambiguated lexemes* in place of words, and of explicit variables to designates the *actants* of the situation described. The French verb *aider* would be defined as follows:

> aider $^{\mathbf{1a}}$: W *aide* Y *à* Z-*er* = W *est un facteur* $^{\mathbf{2a}}$ *contribuant à ce que le processus ou l'action* Z *de* Y *ait lieu.*

ECD entries also incorporate a *valency description* which stipulates how the actants are to be syntactically realized: noun phrase, prepositional phrase, infinitival complement, etc. Finally ECD descriptions provide an extremely detailed specification of lexical co-occurrence restrictions by means of *lexical functions*, a formal device used to capture sense relationships (such as synonyms and conversives), derivational relationships, as well as co-occurrence patterns such as *support* *verbs*. For example, the semantic relationship between *to offer* and *to make an offer* is captured by the **oper1** function:

Oper1(offer) = make [an offer]

Oper1(crime) = commit [a crime]

Mel'cuk and his colleagues of l'Université de Montréal have been working for a few years on an ECD for French, a long term enterprise given the complexity of ECD descriptions. So far a few hundred entries have been collected and published. At the COLING-88 conference, Richard Kittredge and Alain Polguère from the Montreal-based Odyssey Research Inc. presented their ongoing work on implementing a system based on the Meaning-Text model and an ECD-type dictionary.

6. Spelling checkers and conjugation programs

French spelling checkers compatible with popular word processors for personal computers have recently made their appearance in Canada. For example, Edit Inc. (Montreal) sells a French spelling checker for Word Perfect files. Since French morphology is relatively rich, the better packages are based on morphological analysis. Some packages (e.g. HUGO from Logidisque) are able to detect some types of grammatical mistakes.

French writers can also buy morphological synthesis software (e.g. SM-Verbes, Le Conjugueur) to let their personal computers help them with the many quirks of French verb conjugation.

7. Generic morphological analysis/synthesis

Some companies offer morphological analysis/synthesis modules that can be incorporated right into your NLP system. A good example is the Dictionnaire micro-informatisé du français (DMF) sold by a Montreal-based firm (Progiciels Bourbeau et Pinard Inc.). The DMF lists more than 50,000 French words together with a full description of their morphological properties and a partial description of their syntactic properties. It comes in a highly compressed format, together with programs for morphological analysis and synthesis.

8. Text-analysis packages

The Centre d'analyse de textes par ordinateur at the University of Quebec in Montreal has developed a LISP-based package called DEREDEC, that provides a complete environment to develop text analyzers. High-level facilities are offered to develop dictionaries, grammars and semantic modules. Some DEREDEC modules are available for French texts: a 70,000-word morphological analyzer and a large-scale "surface grammar" of French.

DEREDEC has been used so far mostly as a tool that helps researchers in the humanities explore large bodies of textual material.

9. Computer-aided instruction

The ENGLISH I package (Educational Software Products, Toronto) provides assistance for English language acquisition, by detecting the common mistakes made by students. The system uses a 60,000 morphological dictionary. Moreover it incorporates an electronic version of the Gage dictionary of English.

10.The CRITTER dictionaries

One of our projects at the Canadian Workplace Automation Research Center is aimed at the development of an improved conceptual model for MT. As a tested for the model, we are building a system named CRITTER, which translates agricultural market reports between English and French (in a reversible manner).

The CRITTER monolingual French and English dictionaries are rather small (currently 800 entries each) and oriented towards the sublanguage of market reports. However, they embody some interesting features. Since CRITTER is based on a "lexicalist" model, the lexical component is responsible for the description of *polyvalence* phenomena (passivization, raising, control, etc.).

In spite of this additional burden placed on the lexical component, our dictionary entries remain relatively simple and perspicuous, thanks to the use of an elaborate hierarchical system of lexical classes. For example, the verb *give* is assigned to a class called *dative_alternation*. In class definition module, *the dative_alternation* is defined in terms of the union of the class *bi_transitive* (cf. Jack gave Jill a hog) and the class *transitive_oblique* (cf. Jack gave a hog to Jill). These two classes are in turn defined as specializations of the class *transitive*, which features both active and passive forms (see Figure 1).

ENTRIES	dict(give, T) :- T.citation_form = give, T.infl_class = vb12, T.sem_form = give(A, B, C), dative_alternation(T, {A, B, C}).
CLASS DEFINITIONS	dative_alternation(T, Sem_Vars) :- bi_transitive(T, Sem_Vars) ; transitive_oblique(T, Sem_Vars). bi_transitive(T, {A, B, C}) :- T.subcat = [_, _, COMP2], COMP2.cat = np, COMP2.sem_form = C, transitive(T, {A, B}). transitive(T, Sem_Vars), :- active_form(T, Sem_Vars) ; passive_form(T, Sem_Vars).

Jack gave Jill a hog.	Jill was given a hog by Jack.
Jack gave a hog to Jill.	A hog was given to Jill by Jack.

Figure 1.
Monolingual dictionary descriptions in the CRITTER system