ELECTRONIC DICTIONARIES
AND MACHINE TRANSLATION SYSTEMS

Pierre Isabelle

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

November 1988

1. Introduction

In recent discussions on electronic dictionaries (ED's), some authors have contrasted two approaches to dictionary building (cf. for example the paper by Wilks and associates in the Proceedings of COLING-88). In the "demo" approach, one encodes by hand a small but rich lexicon geared to the needs of a specific application. In the "book" approach, one attempts to develop procedures for transforming the content of standard dictionaries and encyclopedias into representations suitable for NLP tasks.

This binary opposition tends to conflate several issues that should probably be kept separate:

1) In building NLP systems, should we attempt to use general-purpose dictionaries, or is it best to develop application-specific or sublanguage-specific dictionaries?
2) What should we use as the primary source of data in the construction of ED's for NLP purposes: text corpora, existing human-oriented reference tools (dictionaries, encyclopedias, books, etc.) or both?
3) Should lexical information be hand-coded or automatically extracted (whether the sources are standard dictionaries, corpora, etc.)?
4) Can ED's be formulated in a theory-neutral way, so that they will be useful regardless of one's theoretical preferences?

There is, to be sure, some level of independence between these four issues. General-purpose and sublanguage-specific dictionaries can both be developed either on the basis of corpora (given some more or less precise linguistic theory) or on the basis of standard (possibly specialized) dictionaries. Whichever you chose, you still have to decide between hand-coding your dictionary or trying to devise a procedure that will derive it from the data. And once you have made that choice, you can either aim for a theory-neutral end-product or a highly biased information structure. I will therefore try to address each one of the four questions separately.

2. Sublanguage-specific and general-purpose dictionaries

During my years at the TAUM group, I learned the importance of focusing on restricted sublanguages. Richard Kittredge and others have repeatedly pointed out that sublanguages typically exhibit patterns that seem foreign to the "general" language. This is particularly true at the level of lexical patterns (e.g. subcategorization and semantic relationships).

The following sentence exemplifies some patterns that are typical of agricultural market reports:

Cows closed steady to easier.

This sentence is "anomalous" in many ways. The noun *cows* is taken to denote the price of cows. The verb *close* has a peculiar subcategorization pattern: it acts more or less as a copula, relating its subject to an adjectival complement (with the stipulation that the property denoted by the complement holds at the time of closing). The adjectives *steady* and *easier* are taken to be co-members of a certain contrast set, while in the language at large, they are normally thought of as belonging to disjoint sets. These two run-of-the-mill adjectives have become **terms** of the domain (market reports), and acquired a special meaning (which can be formalized as precise functions from commodities to numerical values). General vocabulary and domain-specific terminology do not, in general, constitute disjoint sets.

Thus, in particular sublanguages, some of the structural relationships that hold in the general vocabulary are remapped to suit a restricted conceptual universe. This remapping operation is not in general totally predictable.

These facts have very important consequences on the outlook for the development of large-scale ED’s. An MT system aimed at producing high-quality translations in some given sublanguage needs a detailed description of the lexical structures that are specific to that sublanguage. General-purpose dictionaries are usually much too imprecise.

At the same time, there is very little evidence that we are in a position to automatically produce translations of acceptable quality outside the realm of very narrow sublanguages. I believe that concentrating on very narrow sublanguages is still the best strategy for MT. Building sublanguage-oriented ED's will hopefully help us better understand sublanguage-specific lexical structures, and their relationship with the general vocabulary.

Still, it is clear that the development of large-scale general-purpose ED's would be extremely valuable from the point of view of translators, for such dictionaries could provide the basis required to develop various translator's aids (as opposed to pure MT).

3. Sources for data: corpora and standard reference tools

There is no question that dictionaries, encyclopedias and reference books, whether domain-specific or general-purpose, are useful to those who develop NLP systems. In view of what we said above, it is equally clear that they seldom if ever provide all the information that is needed.

When good dictionaries and term banks are available, one does not hesitate to pump some or all of their content into his system. These tools are based on very large scale corpora, and there is no point in duplicating efforts. However, in many cases there doesn't exist any adequate reference tool. Take again the case of agricultural market reports. As far as I know, no existing dictionary will provide answers to questions such as the following. What is the precise meaning of *cows traded easier* as opposed to *cows traded steady*? Do *ample supply* and *good supply* refer to the same or to different points on the relevant scale?

To find the answer to such questions, one must either scrutinize a corpus, or ask the competent "speakers" of the sublanguage. NLP practitioners cannot in general rely on existing resources to rid them completely of the pains of lexicography, and this situation is not likely to change in the near future.

4. Hand-coding and automatic extraction

Ideally, an NLP system would be able to learn both from standard reference tools and from text corpora (not to mention competent speakers). Learning issues are obviously critical to the future of NLP. However, in my opinion, we currently know so little about the content of what is being learned, that it is difficult to even raise the right questions about learning techniques and strategies.

Standard dictionaries are certainly a good starting point for research projects on the acquisition of lexical knowledge. But since a learning tool is always relative to a learning device, we also have to rethink our NLP systems.

Take for example the problem of using standard bilingual dictionaries and term banks in MT systems. Most existing systems fail to draw a neat separation between SL, TL and contrastive lexical knowledge. In particular, lexical transfer rules are often little more than a misplaced, directional and highly procedural restatement of the properties of TL lexical items (cf. the "rule" to translate *know* as either *connaître* or *savoir*).

If we want to tackle the learning problem, our MT systems have to evolve toward more realistic models. We have to make them capable of dealing with declarative and non-directional descriptions of lexical phenomena. This is a problem we have addressed to some extent in our CRITTER system at the CWARC. In the longer term, we would like our systems to be capable on interpreting dictionary entries written as quasi-natural metalanguage statements (cf. Isabelle & Macklovitch, Transfer and MT Modularity, COLING-86). Much research will be required before we can make our MT systems capable of learning from the examples that are typically found in standard dictionaries.

5. Theoretical neutrality

Dictionaries are certainly not atheoretical entities. Try to imagine for example, how far apart an *LFG* dictionary would be from a *japhetist* dictionary written by Marr, had japhetist lexicography flourished long enough. Still, there is no reason why we couldn't develop lexical descriptions based on a core of assumptions shared by a group of closely related linguistic theories. The result will fall short of providing everything that anyone needs, but it might still be useful to everyone. The contents of the entries will of necessity be oriented towards a rather "surfacy" and "observational" level, where linguistic theories are less likely to have profound disagreements.

Lexicalist models tend to include more information in the dictionary, thereby increasing the risks of theoretical disagreements over the precise contents of the entries. One possible way of alleviating this problem is to use "layered" descriptions, in which the upper layers are mostly concerned with observations that are not highly theory-specific. For example, in the CRITTER system, the syntactic behavior of verbs is described in terms of a hierarchy of classes such as *simple_transitive*, *transitive_oblique*, *bi_transitive*, *dative_alternation*, etc. These classes are defined in a separate component, and are ultimately reduced to a small number of primitive notions. For example, the class *dative_alternation* is defined as a class whose elements belong at the same time to the class *bi_transitive* and to a subclass of the class *transitive_oblique*. The classes *simple_transitive*, *transitive_oblique* and *bi_transitive* are all defined in terms of the class *transitive* plus some additional stipulation. It is only at the bottom of the hierarchy that the definitions acquire a strong theoretical bias. For example, one of several possible definitions for the class *transitive* involves two alternative instantiations of an HPSG-like *SUBCAT* feature: one for the active pattern , the other for the passive pattern.

As far as I can tell, our dictionaries could in a large measure be made to fit different theoretical commitments, by simply redefining the low-level lexical classes.