Programming of Future Generation Computers II
K. Fuchi, L. Kott (Editors)
Elsevier Science Publishers B.V. (North-Holland), 1988

BUILDING ELECTRONIC DICTIONARIES FOR NATURAL LANGUAGE PROCESSING

Laurence DANLOS and Maurice GROSS

Laboratoire d'Automatique Documentaire et Linguistique (LADL)
Université de Paris 7
2 Place Jussieu, 75251 Paris Cedex 05, France*

The complexity of building electronic dictionaries for use in natural language processing systems is underestimated most of the time. In particular, it is generally but wrongly believed that commercial dictionaries in a machine readable form will do nicely. The main differences between electronic and commercial dictionaries will be presented along with some of the linguistic problems that arise when building electronic dictionaries.

1. INTRODUCTION

Building a dictionary involves the preliminary step of entering a list of words of the corresponding language. Such a step has always seemed trivial to most computational linguists who have concerned themselves with the complex procedures of syntactic and semantic analysis since the early days of mechanical translation. They generally considered that appending a dictionary to their programs was just a matter of "keypunching" the entries of commercial available dictionaries. However, we found that the problem was by no means simple when we decided to build an electronic dictionary of French at the LADL. Lots of questions, both linguistic and computational, had to be solved in order to get a meaningful representation of what is commonly thought to exist under the term "the set of words of a given language". Furthermore, inflecting the entries of a commercial dictionary to get the full set of words found in texts (e.g. nouns in plural, conjugated verbs) has never been a deep concern, especially to those dealing with English which is a language with a rather poor system of inflection. Again, it is generally believed that the information contained in available dictionaries and grammars is sufficient to build a program of inflection, which is far from being the case.

We will discuss these subjects and present the results obtained so far for French. These results should also be relevant for other languages.

* This work was supported by the Ministère de la Recherche et de l'Enseignement Supérieur within the framework of the Programme de Recherches Coordonnées en Informatique et Linguistique.

2. ELECTRONIC VERSUS ORDINARY DICTIONARIES

By electronic dictionary, we mean a computerized dictionary intended for use in rather sophisticated computer operations such as recognizing a complex technical term in a text or parsing a text to locate certain data. These operations are preliminary steps which are necessary in complex systems of translation or information retrieval. Electonic dictionaries and ordinary commercial dictionaries (which have been computerized) differ on several crucial issues, among which coverage and information.

2.1. Coverage of a dictionary

According to the uses electronic dictionaries are supposed to offer, they should be complete. For example, a spelling error corrector is reliable only if it is based on a complete dictionary, otherwise it considers as mispelled all the well-spelled words that are not included in the dictionary. Moreover, a parsing process is likely to fail in the analysis of a sentence which includes an "unknown" word.

Coverages provided by commercial dictionaries are determined by a compromise between several non linguistic parameters. It even appears that the rational approach to coverage is done in terms of marketing : the public and selling price of a dictionary are determined first, then its linguistic coverage is determined from these data. Suppose, for example, that some publisher feels there is a need for a language rather than encyclopedic dictionary, and that the price is placed at 200 French Francs. Since the amount of information attached to each entry (explanations and examples) is constrained by traditions and competition among publishers, and since the size of the dictionary in number of characters is determined according to its price, the number of entries of a 200 FF language dictionary is going to be 60.000[1]. This way of determining the number of entries makes the use of ordinary dictionaries for natural language processing impossible.

Ordinary general purpose dictionaries include some of the every day language words plus some of the very common technical terms. There exist some specialized commercial dictionaries. They present the same deficiencies as ordinary commercial dictionaries. For natural language processing, one may want several electronic dictionaries:
- a general dictionary which includes all the non technical words and which is used for any application,
- for each domain (e.g. computer science, medecine), a specialized dictionary which includes all the technical terms pertinent to the domain and which is used only for certain applications.

The plurality of dictionaries raises the question of the limits of a given domain, which is not a simple matter. However, it seems desirable since a unique electronic dictionary with all the technical terms of any domain would raise tremendous problems of memory size and search programs.

2.2. Information attached to a word

Another difference between electronic and commercial dictionaries lies in the nature of the information attached to lexical items. There is no limit to the amount of information that one may want to attach to a word: syntactic, semantic, stylistic, historical, encyclopedic data can be introduced, depending on applications. Let us concentrate here on the grammatical information that is needed to inflect a word or to keep it invariable. For European languages,

---

[1] This information comes from Jean Dubois, author of several dictionaries published by Larousse.

this information is restricted to gender, number, case, tense, mood and person, and is often linked with endings. Some formal requirements are lacking in ordinary dictionaries. For example, it is easy to verify that even a mark such as the plural of nouns has been poorly represented in ordinary dictionaries (and in grammars as well) which generally use a default rule such as: " if unspecified, the plural is formed by adjunction of an *-s* ". Such a rule applies to all subentries of a given entry. For example, it applies to all subentries of the noun *ocean* (the plural of which is *oceans*), although *ocean* in *Pacific ocean* has no plural. Many other words which do not have any plural are not indicated as such. For example, it is not indicated that the French noun *cours* , which corresponds to the two English nouns *course* and *courses* , cannot be added an *-s* in the plural. Similarly, it is not indicated that the English noun *police* does not take an *-s* in the plural and this noun is not marked as not being in the singular.

3. LINGUISTIC PROBLEMS WITH THE COVERAGE ISSSUE

Ordinary dictionaries are designed for human beings who already have a good knowledge of language. Therefore, many words the meaning of which is obvious are omitted from dictionaries. The reason is to avoid commercial dictionaries to be redundant to a point where they would become unwieldy. Yet, the implicit knowledge must be made explicit in electronic dictionaries, which requires linguistic problems to be solved. As an illustration of this claim, let us consider adjectives derived from verbs by suffixation of *-able* . For example, *swallowable* is a well-formed adjective derived from the verb *swallow* . The interpretation of this adjective is computable since its derivation follows a regular transformational pattern, namely :

One can swallow this thing
= *This thing can be swallowed*
= *This thing is swallowable*

However, beside pairs such as *swallow-swallowable* which present a regular morpho-syntactic and semantic relation between the verb and the adjective, the following cases are observed:
- pairs such as *remain-*remainable* for which there is no adjective in *-able* derived from the verb,
- pairs such as **charit-charitable* for which the adjective in *-able* is not derived from a verb,
- pairs suchs as *drink-drinkable* for which the morpho-syntactic (and semantic) relation between the verb and the adjective is not clearly regular.

As a consequence, either the full list of adjectives in *-able* regularly derived from a verb must be compiled and integrated in an electronic dictionary, or explicit rules must be stated that predict the productive forms.

Let us now examine the steps which are necessary to reach one of these two solutions. First, the search for *-able* adjectives is not a simple yes-no answer to the question: "Does the verb V have a derived adjective in *-able* ? Consider the pair *break-breakable* . The adjective *breakable* is associated with the verb *break* in the sentence

(1) *The bomb explosion broke the TV set*
= *The TV set can be broken*
= *The TV set is breakable*

but it is not in the sentence:

(2) *John broke with Mary*
**Mary is breakable*

One could be satisfied with an "existential" answer such as "there exists one meaning (use or entry) of the verb *break* which has an adjectivable form *breakable* " . This existential answer is clearly unsatisfactory. It does not take into account the various uses of *break* which must be recorded as many subentries of this verb in any dictionary, especially in an electronic dictionary since a full lexico-syntactic description of verbs is needed for sentence parsing or generation. Hence, a good knowledge of the formal conditions under which the adjectivization in *-able* occurs can be obtained only by examining a lexicon in which the structures (and meanings) of verbs have been clearly separated. Commercial dictionaries do separate sentences such as (1) and (2), however they do not include a systematic separation of subentries of a given word. In particular, attempts to separate metaphoric (figurative) meanings from proper ones have considerably blurred the notion of subentries of a given word. Consider for example:

(3) *This letter broke Mary*

It is customary to call (3) a metaphor of (1). Again, the adjectivation

**Mary is breakable*

is forbidden, but for unclear reasons since the transformational source of this *-able* adjective is accepted:

Mary can be easily broken

Moreover, the sentence

(3a) *This letter shattered Mary*

is perceived as a metaphor of

(1a) *The bomb explosion shattered the TV set*

and here the adjectivization in *-able* is accepted both for the proper and figurative uses of *shatter* :

The TV set is shatterable
Mary is shatterable

Therefore, the search for *-able* adjectives has to go through the following steps:
1) for each reading of a verb, separate its "metaphoric" use(s) from its "proper" one(s),
2) for each "metaphoric" and "proper" use of the reading of a verb, determine if the *-able* adjectivization is accepted.
Step 1) (i.e. the metaphoric issue) is far from being solved. The term "metaphor" suggests a relation between (1) and (3) or (1a) and (3a). However only one aspect of the relation can be made clear : it has a diachronic (etymological) origin. Too many questions remain, in particular: what is the degree of generality between (1) and (3)? In fact, there exist sentences which are intuitively similar to (1) or (1a)

(1b) *The bomb explosion (missed + reached + imploded) the TV set*

for which there is no corresponding metaphor:

(3b) **This letter (missed + reached + imploded) Mary*

Therefore, the lexicon has to be investigated before any systematic relation between (1) and (3) or (1a) and (3a) can be laid down. Furthermore, the nature of such a relation is unclear.

For example, why can a change of proper to metaphoric use of a verb block a morpho-syntactic process such as the adjectivization in *-able* ? As a matter of fact, since step 1) has never been thoroughly studied, step 2) cannot be carried out.

In conclusion, in the present state of knowledge, nobody is in a position to compile the full list of *-able* adjectives or to lay down rules that predict the productive forms. Still, rules would be better than a full list. Consider for example the two sentences:

Lebanon cannot be de-unifilized
Lebanon is not de-unifilizable

They are immediately interpretable, yet verbs such as *unifilize* cannot be included as entries in a dictionary since they are coined in unique textual or extra-linguistic conditions. However, these verbs (and their associated *-able* adjectives) can be subjected to rules that establish syntactic and semantic links with the proper noun UNIFIL. Rules that predict *-able* adjectives would state that *abbreviatable* and *abjurable* are well-formed adjectives regularly derived from transitive verbs. These two adjectives, which could possibly appear in proper contexts, cannot be found in the Oxford English Dictionary.

4. COMPOUND TERMS

The next step in complexity lies in the construction of dictionaries of compound terms. We have adopted the classification in parts of speech used for simple terms:

- compound verbs: *look upon, kick the bucket, take into account,* etc.

- compound adjectives: *free of charge, well done, well-to-do,* etc.

- compound adverbs: *from time to time, time and again, in fact,* etc.

- compound nouns: *sulfuric acid, border town, deed of gift,* etc.

- other varied compounds such as determiners *(as many as, a handful of)* or conjunctions *(as soon as, to the extent that, in order (that + to))*, etc.

Within each of these major categories, we have defined subclasses in terms of the categories that make the compounds.

From the point of view of the recognition of complex utterances in a text, at least two main types of entries have to be distinguished depending on the variability of the terms:

- totally frozen compounds such as many adverbs, for example *as a matter of fact* or *by chance* where the nouns can neither be put in plural nor modified by any adjective. The compound *in order to* does not belong to this category because modifying insertions are possible, e.g. *in order presummably to*.

- variable compounds: they range from nouns with a plural form, the simplest form change, to discontinuous verbs such as *take X into account* in which the verb *take* undergoes morphological changes and the sequence *X* that separates the two fixed components may include a variety of objects and/or adverbs.

In current dictionaries, compounds raise an obvious lexicographic problem: since they are made of at least two parts, how should they be entered? Namely, should the adjective *ill fated* be entered under *ill* or under *fated* ? Entering it twice is cumbersome. Devising rules

that allow to choose between one or the other part is a complex matter. On the one hand, entering *ill fated* under *ill* would make that the number of items under the entry *ill* might become too large (hence difficulties of consulting) since there are numerous similar forms, e.g. *ill advised, ill humoured, ill tempered,* etc. This is a case when an argument of frequency in the lexicon may play a role. On the other hand, entering the term under *fated* would hide the syntactic pattern of prefixation by *ill* . Here a linguistic argument runs against a statistical one. Choosing systematically the left most component would coincide with the idea of displaying the syntactic pattern of prefixation by *ill* . However the suffixation rule would then be hidden for other compounds such as *context free, delivery free,* etc. Entering a term under the part which is first in alphabetic order has drawbacks too. All in all, decisions about entering terms are often arbitrary. An awkward consequence of this difficulty is that the number of compounds in dictionaries is quite restricted.

The main possible reason why many compound terms cannot be found in dictionaries is that they are difficult to define in a simple way. Their only general feature across languages is their lack of compositionality. For example, a term such as *tear gas* can be given neither a syntactic structure nor a transformational source: even a form like *gas for tears* could not be a basis for a semantic interpretation. Notice that *gas which causes tears* is not a sufficient definition either, since onion vapours have the same effect and are not *tear gases* . Part of the meaning is linked to the use of *tear gas* as deterrent. This type of comment is the essence of non compositionality. If one wants to state more precisely the lack of compositionality of a given expression, one has to show that the expression cannot be analyzed with respect to a full grammar. In principle, a full grammar includes:

- a set of syntactic rules,
- a set of semantic interpretation rules that map syntactic structures into semantic structures.

Since no full grammar of any language has ever been built so far, determining lack of compositionality is a process that has to be performed in a specific way for each expression and which thus varies from one expression to another. Let us consider a few examples of compound adverbs:

- *time and again* is a unique (ill formed) coordination of a noun and an adverb. No variations are allowed:

**instant and again*
**time and more*

Moreover, no interpretation can be based on the meanings of *time* (and of *again* , although perhaps less) since the coordination is ill formed;

- *by and large* is even worse with respect to the preceding argumentation;

- *roughly speaking* is not a stylistic permutation of *speaking roughly*. However, there could be evidence to include in the grammar the following derivation which applies to a productive set of adverbs:

Technologically speaking, [John is right]
= *From a technological point of view, [John is right]*
= *When one speaks from a technological point of view, [John is right]*

Now for our example, the question comes to whether the adjective *rough* enters into the derivation or not:

?From a rough point of view, [John is right]

If the answer is positive, the compound adverb is compositional, otherwise, it has to be considered as an idiom.

This situation affects compound nouns as well:

- the meaning of *couple charged device* is very precise, whereas the meaning of each part is loose and ambiguous;

- even *magnetic tape* is not compositional, although the meanings of *magnetic* and *tape* seem precise. Whatever the transformational source one may give to this compound, this transformational source cannot be mapped into a semantic representation that has the technological significance of this term. Most technical terms work in the same way: the words they consist of ring a bell which is relevant to the whole meaning, however in no case can they account for the extra-meaning.

Besides the lack of compositionality of compounds, there may be features that can help to detect them:

- in French, the term "mot composé" is reserved to terms that are hyphened. However, there are neither rules nor guidelines to specify the use of hyphens. The study by Mathieu-Colas [1] made through the main dictionaries of French shows that they disagree in a large number of cases. Hyphens, which can be a stylistic device, are used in various ways by different authors and/or at different times (at the beginning of the 19th century, hyphens were more common, e.g. *pomme de terre* could be found with hyphens);

- in German, compound nouns are not graphically distinguishable from simple nouns. In such cases, a linguistic analysis is the only way to separate compositional items from non compositional ones. As a matter of fact, most languages function like German. Even in French, many compounds are written as one word, for example the following technical terms:

électrocardiogramme, [acide] phénylacrylique, etc.

- in Italian and Spanish, many diminutive and augmentative forms of nouns and adjectives are to be found. It would be quite redundant to list them all. To analyse them as compounds seems a better solution.

It should be noted that bilingual dictionaries often contain more compound terms than monolingual dictionaries. In technical dictionaries, it could not be any other way since most technical terms are compounds. For general purpose dictionaries, translation is often a test to detect non compositionality: compound terms can be translated into simple ones (e.g. *pomme de terre* (French) <--> *potato* (English)), or else they are not translated word-by-word (e.g. *pièce de collection* (French) <--> *collector's item* (English)).

5. FRENCH ELECTRONIC DICTIONARIES

Today, the dictionaries built at the LADL have reached the following stages:

- DELAS is an electronic dictionary of simple basic (non inflected) words develloped by Courtois [2]. A word is defined as a sequence of characters occuring between consecutive separators (e.g. blanks). The DELAS contains 60.000 words. Each word has a code that determines its part of speech and its grammatical changes for inflected forms. A word such as *voile* is recorded in the DELAS as:

- a masculine noun *(veil)*
- a feminine noun *(sail)*

- DELAF is the dictionary of inflected words obtained by applying an inflection program to the entries of the DELAS. This program expands the DELAS to a set of over 500.000 inflected forms and it associates the grammatical category(ies) of each form. A word such as *voile* is recorded in the DELAF as:

- a masculine singular noun *(veil)*
- a feminine singular noun *(sail)*
- a verb *(veil)* either in the indicative mood (1st and 3rd person of the singular in the present tense), or in the subjunctive mood (1st and 3rd person singular in the present tense), or in the imperative mood (2nd person of singular).

- DELAC is the dictionary of compound nouns. The description of compound nouns is a large entreprise as shown in 4. So far, 40.000 forms *N Adj* (e.g. *tube cathodique, cordon bleu, vin blanc*) have been described with their possible variations for feminine or plural by Gross [3]. When achieved, the DELAC is expected to contain 400.000 compound nouns of the every day language.

- As they are less variable than other compounds, compound adverbs have been described as complex words. These 5.000 adverbs have been classified according to the grammatical nature of their components by Gross [4].

The dictionaries we have mentioned so far give only grammatical information for inflection. The syntactic and semantic information is stored in "lexicon-grammars". The concept of lexicon-grammar originated in the observation that the syntactic description of a given verb is actually the description of the elementary sentence in which it enters (i.e. subject-verb-objects). This observation is also relevant for adjectives that enter in a *be* construction as well. A full description of lexicon-grammars will not be given here. However, let us at least say that:

- A lexicon-grammar of about 12.000 French simple verbs has been built by Gross [5], J.P. Boons, A. Guillet and Ch. Leclère [6] and [7]. Verbs have been classified in about 50 classes according to the shape of the elementary sentences into which they enter, that is roughly the number of objects they take (0, 1 or 2) and the nature of their prepositions (namely "zero", *à* or *de*). For each verb, an average of 40 syntactic properties (out of a set of 500) has been recorded: passivization, reflexivization, shape of clitic pronouns, etc. The format adopted for this lexicon-grammar (i.e. table) is neutral with respect to computer applications. Therefore, mechanical procedures have been devised by Salkoff [8] and Danlos [9] to translate these lexico-syntactic tables into the format required by parsing or generation systems.

- A lexicon-grammar of about 18.000 compound verbs has been built by Gross [10] on the same principles as for the previous one.

Lexicon-grammars for other sentence types, mainly nominal, have also been built by Giry-Schneider [12] , [13] and Danlos [14].

REFERENCES

[1] Mathieu-Colas, M., 1987, Variations orthographiques, Programme de Recherches Coordonnées Informatique et Linguistique, rapport no 5.

[2] Courtois, B., 1985, Dictionnaires électroniques du français, Programme de Recherches Coordonnées Informatique et Linguistique, rapport no 1.

[3] Gross, M., 1986, Lexicon-Grammar, The Representation of Compound Words, *in Proceedings of ColingS86,11th International Conference on Computational Linguistics* , Bonn.

[4] Grosss,M, 1988, *Syntaxe transformationnel du français : syntaxe de l' adverbe*, Cantilène, Paris.

[5] Gross, M., 1975, *Méthodes en syntaxe* , Hermann, Paris.

[6] Boons, J.P., Guillet, A., Leclère, Ch., 1976a, *La structure des phrases simples en français : les constructions intransitives,* Droz, Genève.

[7] Boons, J.P., Guillet, A., Leclère, Ch., 1976b, *La structure des phrases simples en français : classes de constructions transitives,* Rapport de recherche de l'Université de Paris 8, numéro 8.

[8] Salkoff, M., 1973, *Une grammmaire en chaîne du français* , Dunod, Paris.

[9] Danlos, L. , 1987, *The Linguistic Basis of Text Generation,* Cambridge University Press, Cambridge.

[10] Gross, M., 1982, "Une classification des phrases "figées" du français", in P. Attal et XCCl. Luller, eds., *Actes du Colloque de Rennes 1980,* Amsterdam, Benjamin.

[11] Giry-Schneider, J., 1978, *Les nominalisations en français : l'opérateur "faire" dans le lexique* , Droz, Genève.

[12] Giry-Schneider, J., 1987, *Les prédicats nominaux en français : les phrases simples à verbe support*, Droz, Genève.

[13] Danlos, L. , 1988, "Les phrases à verbe support *être Prép* ", *Langages 90*, Larousse, Paris.