ADAPTING AN ENGLISH MORPHOLOGICAL ANALYZER FOR FRENCH

Roy J. Byrd and Evelyne Tzoukermann

IBM Research
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

A word-based morphological analyzer and a dictionary for recognizing inflected forms of French words have been built by adapting the UDICT system. We describe the adaptations, emphasizing mechanisms developed to handle French verbs. This work lays the groundwork for doing French derivational morphology and morphology for other languages.

1. Introduction.

UDICT is a dictionary system intended to support the lexical needs of computer programs that do natural language processing (NLP). Its first version was built for English and has been used in several systems needing a variety of information about English words (Heidorn, et al.(1982), Sowa(1984), McCord(1986), and Neff and Byrd(1987)). As described in Byrd(1986), UDICT provides a framework for supplying syntactic, semantic, phonological, and morphological information about the words it contains.

Part of UDICT's apparatus is a morphological analysis subsystem capable of recognizing morphological variants of the words whose lemma forms are stored in UDICT's dictionary. The English version of this analyzer has been described in Byrd(1983) and Byrd, et al. (1986) and allows UDICT to recognize inflectionally and derivationally affixed words, compounds, and collocations. The present paper describes an effort to build a French version of UDICT. It briefly discusses the creation of the dictionary data itself and then focuses on issues raised in handling French inflectional morphology.

2. The Dictionary.

The primary role of the dictionary in an NLP system is to store and retrieve information about words. In order for NLP systems to be effective, their dictionaries must contain a lot of information about a lot of words. Chodorow, et al.(1985) and Byrd, et al.(1987) discuss techniques for building dictionaries with the required scope by extracting lexical information from machine-readable versions of published dictionaries. Besides serving the NLP application, some of the lexical information supports that part of the dictionary's access mechanism which permits recognition of morphological variants of the stored words. We have build a UDICT dictionary containing such morphological information for French by starting with an existing spelling correction and synonym aid dictionary[1] and by adding words and information from the French-English dictionary in Collins(1978).

French UDICT contains a data base of over 40,000 lemmata which are stored in a direct access file managed by the Dictionary Access Method (Byrd, et al. (1986)). Each entry in this file has one of the lemmata as its key and contains lexical information about that lemma. Other than the word's part-of-speech, this information is represented as binary features and multi-valued attributes. The feature information relevant for inflectional analysis includes the following:

```
(1) features:

    part-of-speech
    singular
    plural
    masculine
    feminine
```

[1] We are grateful to the Advanced Language Development group of IBM's Application Systems Division in Bethesda, Maryland, for access to their French lexical materials. Those materials include initial categorizations of French words into parts-of-speech and paradigm classes.

```
invariant
first (second, third) person
infinitive
participle
past
present
future
imperfect
simple past
subjunctive
indicative
conditional
imperative
```

Some of these features are explicitly stored in UDICT's data base. Other features — including many of the stored ones — control morphological processing by being tested and set by rules in ways that will be described in the next section. Stored features and attributes which are not affected by (and do not affect) morphological processing are called "morphologically neutral." Morphologically neutral information appears in UDICT's output with its stored values unaltered. Such information could include translations from a transfer dictionary in a machine translation system or selectional restrictions used by an NLP system. For French, no such information is stored now, but in other work (Byrd, et al. (1987)) we have demonstrated the feasibility of transferring some additional lexical information (for example, semantic features such as [+human]) from English UDICT via bilingual dictionaries.

It may be useful to point out that, given the ability to store such information about words, one way of building a lexical subsystem would be to exhaustively list and store all inflected words in the language with their associated lexical information. There are at least three good reasons for not doing so. First, even with the availability of efficient storage and retrieval mechanisms, the number of inflected forms is prohibitively large. We estimate that the ratio of the number of French inflected forms to lemmata is around 5 (a little more for verbs, a little less for adjectives and nouns). This ratio would require our 40,000 lemmata to be stored as 200,000 entries, more than we would like. The second reason is that inflected forms sharing the same lemma also share a great deal of other lexical information: namely the morphologically neutral information mentioned earlier. Redundant storage of that information in many related inflected forms does not make sense linguistically or computationally. Furthermore, as new words are added to the dictionary, it would be an unnecessary complication to generate the inflected forms and duplicate the morphologically neutral information. Storing the information only once with the lemma and allowing it to be inherited by derived forms is a more reasonable approach. Third, it is clear that there are many regular processes at work in the formation of inflected forms from their lemmata. Discovering generalizations to capture those regularities and building computational mechanisms to handle them is an interesting task in its own right. We now turn to some of the details of that task.

3. Morphological Processing.

3.1. The mechanism. The UDICT morphological analyzer assumes that words are derived from other words by affixation, following Aronoff(1976) and others. Consequently, UDICT's word grammar contains **affix rules** which express conditions on the base word and makes assertions about the affixed word. These conditions and assertions are stated in terms of the kinds of lexical information listed in (1).

An example of an affix rule is the rule for forming French plural nouns shown in Figure 1. This rule — which, for example, derives *chevaux* from *cheval* — consists of five parts. First, a **boundary** marker indicates whether the affix is a prefix or a suffix and whether it is inflectional or derivational. (Byrd(1983) describes further possible distinctions which have so far not been exploited in the French system.) Second, the **affix name** is an identifier which will be used to describe the morphological structure of the input word. Third, the **pattern** expresses string tests and modifications to be performed on the input word. In this case, the string is tested for *aux* at its right end (since this is a suffix rule), two characters are removed, and the letter *l* is appended, yielding a potential base word. This base word is looked up via a recursive invocation of the rule application mechanism which includes an attempt to retrieve the form from the dictionary of stored lemmata. The fourth part of the rule, the **condition**, expresses constraints which must be met by the base word. In this case, it must be a masculine singular (and not plural) noun. The fifth part of the rule, the **assertion**, expresses modifications to be made to the features of the base in order to

```
-pn: aux21* (noun +masc +sing -plur) (noun +plur -sing)
| |  |      |                        |
| |  |      |                        assertion
| |  |      condition
| |  pattern ("check for 'aux', remove 'ux', add '1', lookup")
| affix name ("plural noun")
affix boundary ("inflectional suffix")
```

Figure 1. The structure of a UDICT morphological rule.

describe the derived word. For this rule, the singular feature is turned off and the plural feature is turned on. Features not mentioned in the assertion retain their original values; in effect, the derived word contains inherited morphologically neutral lexical information from the base combined with information asserted by the rule.

For the input *chevaux* ("horses"), the rule shown in Figure 1 will produce the following analysis:

```
(2) chevaux: cheval(noun plur masc
        (structure <<*>N -pn>N))
```

In other words, *chevaux* is derived from *cheval.* It is a plural noun by assertion. It is masculine by inheritance. Its structure consists of the base noun *cheval* (represented by "<*>N") together with the inflectional suffix "-pn".

In order for rules such as these to operate, there is a critical dependance on having reliable and extensive lexical information about words hypothesized as bases. This information comes from three sources: the stored dictionary, redundancy rules, and other recursively applied affix rules.

While the assumption that affixes derive words from other words seems entirely appropriate for English, it at first seemed less so for French. An initial temptation was to write affix rules which derived inflected words by adding affixes to non-word stems. This was especially true for verbs where the inflected forms are often shorter than the infinitives used as lemmata, and where some of the verbs — particularly in the third group — have very complex paradigms. However, our rules' requirement for testable lexical information on base forms cannot be met by a system in which bases are not words. The machine-readable sources from which we build UDICT dictionaries do not contain information about non-word stems. It is furthermore difficult to design procedures for eliciting such information from native speakers, since people don't have intuitions about forms that are not words. Consequently, we have maintained the English model in which only words are stored in UDICT's dictionary.

UDICT's word grammar includes **redundancy rules** which allow the expression of further generalizations about the properties of words. In a sense, they represent an extension of the analysis techniques used to populate the dictionary and their output could well be stored in the dictionary. The following example shows two redundancy rules in the French word grammar:

```
(3) : ∅ (adj -masc -fem)(adj +masc)
    : e∅ (adj +masc) (adj +fem)
```

The first rule has no boundary or affix name and its pattern does nothing to the input word. It expresses the notion that if an adjective is not explicitly marked as either masculine or feminine (the condition), then it should at least be considered masculine (the assertion). The second rule says that any masculine adjective which ends in *e* is also feminine. Examples are the adjectives *absurde*, *évitable*, and *vaste* which are both masculine and feminine. Such rules reduce the burden on dictionary analysis techniques whose job is to determine the genders of adjectives from machine-readable resources.

For inflectional affixation, we normally derive the inflected form directly from the lemma. However, recursive rule application plays a role in the derivation of feminine and plural forms of nouns, adjectives, and participles — which will be discussed under "noun and adjective morphology" — and in our method for handling stem morphology of the French verbs belonging to the third group, which will be discussed under "verb morphology".

3.2. Noun and adjective morphology. For nouns and adjectives, where inflectional changes to a word's spelling occur only at its rightmost end, the word-based model was simple to maintain.

```
a. -vpres: ent$ (v +inf) (v -inf +ind +pres +plur +pers3)
b. -vsubj: es$ (v +inf) (v -inf +subj +pres +sing +pers2)
c. -vimpf: ions$ (v +inf) (v -inf +ind +impf +plur +pers1)
d. -vpres: e$ (v +inf) (v -inf +ind +imp +pres +plur +pers1 +pers3)
e. -vpres: ons$ (v +inf) (v -inf +ind +imp +pres +plur +pers1)
```

Figure 2. Morphological rules which invoke the spelling rules.

As shown in Figure 1, the pattern mechanism supports the needed tests and modifications. For recognition of feminine plurals, we treat the feminine-forming affixes as derivational ones (using an appropriate boundary), so that recursive rule application assures that they always occur "inside of" the plural inflectional affix. For example *heureuses* is analyzed as the plural of *heureuse* which itself is the feminine of *heureux* ("happy"). Similarly, *élues* ("chosen or elected") is the plural of *élue* which, in turn, is the feminine of *élu* itself analyzed as the past participle of the verb *élire* ("to vote"). The final section of the paper mentions another justification for treating feminine-forming affixes as derivational.

3.3. Verb morphology. Many French verbs belonging to the first group (i.e., those whose infinitives end in *-er*, except for *aller*) show internal spelling changes when certain inflections are applied. Examples are given in (4) where the inflected forms on the right contain spelling alterations of the infinitive forms on the left.

(4)a. peser - (ils) pèsent
b. céder - (que tu) cèdes
c. essuyer - (tu) essuies
d. jeter - (je, il) jette
e. placer - (nous) plaçons

These spelling changes are predictable and are not directly dependent on the particular affix that is being applied. Rather, they depend on phonological properties of the affix such as whether it is silent, which vowel it begins with, etc. There are seven such **spelling rules** whose job is to relate the spelling of the word part "inside of" the inflectional affix to its infinitive form. These rules are given informally in (5). (The sample patterns should be interpreted as in Figure 1 and are intended to suggest the strategy used to construct infinitive forms from the inflected form. "C" represents an arbitrary consonant, "D" represents *t* or *l*, and "=" represents a repeated letter.)

(5) spelling rules:

`i1yer*` - change *i* to *y* and add *er*, as in *essuies/essuyer*

`ç1cer*` - change *ç* to *c* and add *er*, as in *plaçons/placer*

`ge0r*` - add *r*, as in *mangeons/manger*

`èC2eCer*` - remove grave accent from stem vowel and add *er*, as in *pèsent/peser*

`èC2éCer*` - change grave accent to acute on stem vowel and add *er*, as in *cèdes/céder*

`èCC3éCCer*` - like the preceding but with a consonant cluster, as in *sèchent/sécher*

`D=1er*` - remove the repeated consonant and add *er*, as in *jette/jeter*

It would be inappropriate and uneconomical to treat these spelling rules within the affix rules themselves. If we did so, the same "fact" would be repeated as many times as there were rules to which it applied. Rather, we handle these seven spelling rules with special logic which not only encodes the rules but also captures sequential constraints on their application: if one of them applies for a given affix, then none of the others will apply. The spelling rules are invoked from the affix rules by placing a "$" rather than a "*" in the pattern to denote a recursive lookup. In effect, the base form is looked up modulo the set of possible spelling changes. Example affix rules largely responsible for (and corresponding to) the forms shown in (4) are given in Figure 2.

Verbs of the third group are highly irregular. Traditional French grammar books usually assign each verb anywhere from one to six stem forms. Some examples are given in (6).

(6) stems for third group verbs:

a. *partir* has stems *par-, part-*

```
a. -vcond: rions5* (v +stem -inf) (v +cond +pres +plur +pers1)
b. +vstem: saulvoir* (v +inf -stem) (v +stem -inf)
c. saurions: savoir(verb cond pres plur pers1 (structure <<*>V -vcond>V))
```

Figure 3. An example of stem morphology.

b. *savoir* has stems *sai-, sav-, sau-, sach-, s-*

c. *apercevoir, concevoir, décevoir, percevoir, recevoir* have stems in *-çoi-, -cev-, -çoiv-*

d. *contredire, dédire, dire, interdire, médire, maudire, prédire, redire* have stems in *-dis-, -di-, -d-*

Since our derivations are to be based on lemmata, we need a way to associate infinitives with appropriate stem forms. The mechanism we have chosen is to let a special set of **verb stem rules** perform that association. Recognition of the inflected form of a third group verb thus becomes a two-step process. In the first step, the outermost affix is recognized, and its inner part is tested for being a valid stem. In the second step, a verb stem rule attempts to relate the stem proposed by the inflectional affix rule to an infinitive in the dictionary. If it succeeds, it marks the proposed stem as a valid one and the entire derivation succeeds.

Consider, as an example, the rules and system output shown in Figure 3. During the analysis of the input *saurions* ("(we) would know"), the rule in Figure 3(a) will first recognize and remove the ending *-rions*, and then ask whether the resulting *sau* meets the condition `"(v +stem -inf)"`. Application of the verb stem rule in Figure 3(b) will successfully relate *sau* to *savoir* and assert its description to include `"(v +stem -inf)"`, thus meeting the condition of rule (a). The result will be the successful recognition of *saurions* with the analysis given in Figure 3(c). Note that the structure given does not mention the occurrence of the `"+vstem"` affix; this is intentional and reflects our belief that the two-level structural analysis — inflectional affix plus infinitive lemma — is the appropriate output for all verbs. The intermediate stem level, while important for our processing, is not shown in the output for verbs of the third group.

The French word grammar contains 165 verb stem rules and another 110 affix rules for third group verbs. Given the extent of the idiosyncrasy of these verbs and their finite number (there are only about 350 of them), it is natural to wonder whether we might not do just as well by storing the inflected forms. In addition to the arguments given above (about redundant storage of morphologically neutral lexical information, etc.), we can observe that there *are* generalizations to be made for which treatment by rule is appropriate. The lists of verbs shown in (6c,d) have common stem patternings. Lexicalization of the derived forms of these words would not allow us to capture these generalizations or to handle the admittedly rare coinage of new words which fit these patterns.

4. Summary and further work

A recognizer for French inflected words has been built using a modified version of UDICT, which is programmed in PL/I and runs on IBM mainframe computers. Approximately 400 affix and verb stem rules were required, of which over half are devoted to the analysis of French verbs belonging to the third group. 15 redundancy rules and 7 spelling rules were also written. In addition to many minor changes not mentioned in this paper, the major effort in adapting the formerly English-only UDICT system to French involved handling stem morphology. French UDICT contains a dictionary of over 40,000 lemmata, providing fairly complete initial coverage of most French texts, and forming a setting in which to add further, morphologically neutral, lexical information as required by various applications.

We are testing French UDICT with a corpus of Canadian French containing well over 100,000 word types. (The corpus size is close to 100,000,000 tokens.) Initial results show that the recognizer successfully analyzes over 99% of the most frequent 2,000 types in the corpus, after we discard those which are proper names or not French. For a small number of words (fewer

than 25), spurious information was added to the correct analysis. Work continues toward eliminating those errors.

We believe that the resulting machinery will be adequate for building dictionaries for other European languages in which we are interested (Spanish, Italian, and German). In particular, we believe that the spelling rule mechanism will help in recognizing German umlauted forms and that the stem mechanism will serve to handle highly irregular paradigms in all of these languages.

Expressing spelling rules in a more symbolic notation (rather than as logic in a subroutine invoked from affix rules) would facilitate the task of the grammar writer when creating morphological analyzers for new languages. For French, the bulk of the work done by spelling rules is on behalf of verbs of the first group. However, some of the spelling changes are also observed in other verbs and in nouns and adjectives. Currently those effects are handled by affix rules. We look forward to generalizing the coverage of our spelling rules and thereby further simplifying the affix rules.

We also plan to expand our word grammar to handle the more productive parts of French derivational morphology. The attachment of derivational affixes to words is constrained by conditions on a much more extensive set of lexical features than the attachment of inflectional affixes. For example, we have observed that feminine-forming suffixes apply only to nouns which denote humans or domestic animals. The idiosyncrasy of this constraint is typical of derivational affixes and provides further justification for our earlier decision to treat feminine-forming suffixes as derivational. By discovering and exploiting such regularities within our framework, we expect to cover a large set of derivational affixes.

References.

Aronoff, M. (1976) *Word Formation in Generative Grammar*, Linguistic Inquiry Monograph 1, MIT Press, Cambridge, Massachusetts.

Byrd, R. J. (1983) "Word formation in natural language processing systems," *Proceedings of IJCAI-VIII*, 704-706.

Byrd, R. J. (1986) "Dictionary Systems for Office Practice," IBM Research Report RC 11872, T.J. Watson Research Center, Yorktown Heights, New York.

Byrd, R. J., G. Neumann, and K. S. B. Andersson (1986a) "DAM - A Dictionary Access Method," IBM Research Report, IBM T.J. Watson Research Center, Yorktown Heights, New York.

Byrd, R. J., J. L. Klavans, M. Aronoff, and F. Anshen. (1986b) "Computer Methods for Morphological Analysis," *Proceedings of the Association for Computational Linguistics*, 120-127.

Byrd, R. J., N. Calzolari, M. S. Chodorow, J. L. Klavans, M. S. Neff, and O. A. Rizk (1987) "Tools and Methods for Computational Lexicology,". IBM Research Report RC 12642, IBM T.J. Watson Research Center, Yorktown Heights, New York. (to be published in *Computational Linguistics* 1987)

Chodorow, M. S., R. J. Byrd, and G. E. Heidorn (1985) "Extracting semantic hierarchies from a large on-line dictionary," *Proceedings of the Association for Computational Linguistics*, 299-304.

Collins (1978) *Collins Robert French Dictionary: French-English. English-French.* Collins Publishers, Glasgow.

Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow (1982) "The EPISTLE Text-Critiquing System," *IBM Systems Journal* 21, 305-326.

Klavans, J., Nartey, J., Pickover, C. Reich, D., Rosson, M., and Thomas, J. (1984) "WALRUS: High-quality text-to-speech research system," *Proceedings of IEEE Speech Synthesis and Recognition*, pp. 19-28.

McCord, Michael C. (1986) "Design of a Prolog-Based Machine Translation System", *Proc. Third International Conference on Logic Programming*, Springer-Verlag, 350-374.

Neff, M. S. and R. J. Byrd (1987) "WordSmith Users Guide: Version 2.0," IBM Research Report RC 13411, IBM T.J. Watson Research Center, Yorktown Heights, New York.

Sowa, J. F. (1984) "Interactive Language Implementation System," *IBM J. of Research and Development*, vol. 28, no. 1, January 1984, pp. 28-38.