The Application of a Morphological Analyzer to On-Line French Dictionaries

Evelyne Tzoukermann and Roy J. Byrd

Abstract

A morphological analyzer for French has been developed which provides improved access to on-line French and bilingual dictionaries. Since users are able to look up inflected forms as well as citation forms, the dictionary becomes easier to use. Furthermore, coverage is increased by four to five times with a morphological analyzer. The system itself consists of its own internal dictionary containing about 40,000 lemmata, each followed by its part(s)-of-speech, grammatical features and attribute-value pairs. The morphological analyzer comprises a set of inflectional and derivational rules which are applied to the user's input. The lemma which results from the morphological analysis is used as the citation form with which to probe the dictionary.

1. Introduction

Published dictionaries in machine readable form are now in widespread use by researchers in natural language processing (NLP) (Boguraev and Briscoe 1986, Michiels 1982). The Lexical Systems project at IBM Research has developed a system intended for use with the COLLINS ROBERT FRENCH-ENGLISH ENGLISH-FRENCH DICTIONARY (1978). In order to facilitate access to this dictionary, we have developed a morphological analyzer for French. The analyzer determines the lemma of a word that has been input by the user. That lemma is used as the citation form with which to probe the dictionary. As a result, consultation of the dictionary becomes easier and more flexible. The French morphological analyses are produced by a newly constructed version of UDICT, the first version of which was built for English (Byrd 1985, 1986). French UDICT consists of its own internal dictionary and a morphological analyzer. In this paper, we discuss the details of each of these two parts of the system and examine how they are related.

We emphasize one type of application for French UDICT: the enhanced use of French dictionaries within the WordSmith on-line dictionary reference system (Neff and Byrd 1987). The system as it stands already provides a tool for the office worker, the translator and the foreign student. The addition of French morphology will produce an even more valuable tool for these users. Moreover, French UDICT will also support other research work in NLP. With the aid of a recognition system, one can do different kinds of text analysis—lexical, syntactic, and even semantic. Currently, the analyzer is used as a support for automatically labelling the words within a French text. Among other types of application—machine translation, spelling checkers and so on—which we will not go into here, there are other uses for the kind of dictionary system that we describe here.

2. Organization of the Dictionary

The French UDICT dictionary is a data base consisting of about 40,000 entries. Entries in this data base have a very simple form. Since they are currently only accessed through the morphological analyzer, only grammatical information is present. (The English UDICT dictionary contains additional information. See Klavans and Wacholder (1988) for a description of the features and attributes in the Udict system.) The information is coded in a binary representation. Every entry consists of the part-of-speech, grammatical features and attribute-value pairs.

The **parts-of-speech** are as follows:

ADJ (adjective), ADV (adverb), DET (determiner), CONJ (conjunction), INTERJ (interjection), NOUN (noun), PREP (preposition), PRON (pronoun), VERB (verb).

The **grammatical features** are as follows:

MASC (masculine), FEM (feminine), SING (singular), PLUR (plural), IND (indicative), SUBJ (subjunctive), COND (conditional), IMP (imperative), INF (infinitive), PART (participle), PRES (present), PAST (past), IMPFT (imperfect), FUT (future), PSIMP (simple past), PERS1 (first person), PERS2 (second person), PERS3 (third person), STORED (the word is stored in the dictionary), IRREG (irregular), CLASS3 (third class), INV (invariant), ATOMIC (atomic), MONO (monophonemic).

The last four categories deserve further explanation.

The CLASS3 feature: Recall that French verbs are traditionally divided into three groups according to their infinitives: the verbs ending in -er form the first group as in "chanter" (to sing), in -ir the second as in "blanchir" (to whiten) and in -ir and -re the third as in "partir" (to go) or "vendre" (to sell). This last group is a closed class of verbs (no new verbs of this class are coined). Because verbs of the second and third group have common endings in the infinitive and sometimes within the conjugation, we owe to distinguish the different paradigms. A traditional criterion is that second group verbs, but not third group verbs, take "-iss-" as an infix within the present participial ending. Thus, we get "fin-issant" (ending) but not "*dorm-issant". Using the CLASS3 feature for the verbs of the third group, we can easily make the distinction between the two classes (i.e. the verbs in -ir which are not marked CLASS3 belong to the second group). For "dorm-isse" (subjunctive imperfect of "dormir" (to sleep)) the following rules in (1) and (2) will be successively applied:[1]

(1) -vsubj: isse4* (v +stem -inf) (v -inf +subj +impf +sg +pers1)
(2) +vstem: dorm0ir* (v +inf +class3 -stem) (v +stem -inf +subj +psimp)

and will provide the final analysis in (3):

(3) dormisse
(dormir VERB SUBJ IMPFT SING PERS1 STORED CLASS3
(LEMMA dormir)
(STRUCTURE ⟨⟨*⟩V -vsubj⟩V))

whereas for "fin-isse" (subjunctive present and imperfect of "finir" (to finish)), two different rules will be applied:

(4) -vsubj: isse3r* (v +inf -class3) (v -inf +subj +pres +sg +persl +pers3)
(5) -vsubj: isse3r* (v +inf -class3) (v -inf +subj +impf +sg +persl)

(6) finisse
(finir(VERB SUBJ PRES SING PERS1 PERS3 STORED (LEMMA finir)
(STRUCTURE ⟨⟨*⟩V -vsubj⟩V)) (finir(VERB SUBJ IMPFT SING PERS1 STORED (LEMMA finir) (STRUCTURE ⟨⟨8⟩V -vsubj⟩V))

The INV feature: The invariant feature prevents some masculine adjectives and nouns from having a feminine form. For example, "loupe" (magnifying glass) cannot be the feminine form of "loup" (wolf) whose feminine "louve" is irregular. Therefore, "loup" will be stored in the dictionary as:

(7) loup : NOUN SING MASC INV

The ATOMIC feature: This feature applies to words whose stored information is either complete or idiosyncratic and which, therefore, do not require further morphological analysis.

(8) fut : être (verb irreg ind psimp pers3 sg atomic)
vais : aller (verb irreg ind pres persl sg atomic)

The forms "fut" (was or were) and "vais" ((I) go) are irregular and need to be stored without any kind of analysis.

The MONO feature: Te monophonemic feature identifies third class verb stems consisting of one or two phonemes -such as "p-" for "pouvoir" which forms "pus" ((I) could) or "pl-" for "pleuvoir" which forms "plut" (rained). The feature is required to distinguish occurrences of these stems from ordinary words which end with the same letters. Therefore, it allows the program to have control of suffix length. Thus, "sent" ((he) feels, smells) will be analyzed as "sen-t" from "sentir" (to feel, to smell) and not as *s-ent from "savoir" (to know).

The **attributes** are:

ADDENDA—the addenda file from which word information was obtained.
BASE—morphological base, normal spelling, citation form, etc.
LEMMA—citation form for inflected words.

3. Description of the Morphological Analyzer

The UDICT morphological analyzer applies a set of morphological rules to the user's input in order to arrive at a morphological analysis. For inflectional analysis the rules deal with, on the one hand, adjectives and substantives (recognition of gender and number) and, on the other hand, the verbs (recognition of mood, tense and person). As we have mentioned above, the organization of the system follows the same logic as the one used for English. The underlying approach is word based morphology introduced by Mark Aronoff (Aronoff 1976). The basic idea is that

words are derived from words, every word being a separate entry in the dictionary. However, in order to handle peculiar characteristics of Romance languages, some major changes had to be implemented. This lead us to a stem based morphology approach (Corbin 1987), mostly motivated by the verb inflection in French but also by derivation.

3.1. Inflection

In performing inflectional morphology on French, we had to face the complexity of the verbal system, particularly for processing the verbs of the third group (Tzoukermann 1986). These verbs show a variable number of stems that is difficult to systematize. In this group, one can find verbs with one stem such as "rend-" in "rendre" (to give back), with two stems such as "meur-" and "mour-" in "mourir" (to die), with three stems such as "li-", "lis-" and "l-" in "lire" (to read), with four stems (ex: "doi-", "doiv-", "dev-" and "d" in "devoir" (must), with five stems such as "sai-", "sav-", "sau-", "sach-" et "s-" in "savoir" (to know) with six stems such as "peu-", "pouv-", "peuv-", "pou-", "puiss-", "p-", in "pouvoir" (can). On the other hand, the third group's paradigm (i.e., the set of endings added onto the stems) is almost always regular. In order to deal with this multiplicity of stems—which characterizes French and other Romance languages—we have tested a new technique based on a non-word stem. For every inflected verb, a two-step procedure has to be applied, using different strategies according to the verb group.

For the verbs of the first group, although the paradigm of the conjugation is very regular, a certain number of spelling and morpho-phonological rules has to be applied on the stem. Seven rules have been written in a separate routine which is applied after the application of a possible suffix rule. Thus, having recognized a verbal suffix, like the simple past "*-ai*" of "plaç-ai" ((I) put) in (9),

(9) -psimp: ai$ (v +inf) (v -inf +ind +psimp +sg +pers1)

the dollar sign causes all the spelling rules to be applied. The first step provides the information about the mood, the tense, the person and the number of the verb. Afterwards, the spelling rules are executed and in the case of "plaç-ai", the change from "-ç-" to "-c-" is done. The second and last step adds the infinitive "-er" to the rewritten base "plac-".

For the verbs of the third group, the first step is also to apply inflectional rules which recognize the verbal paradigm:

(10) -vimpf: ions4* (v +stem -inf) (v -inf +ind +pres +pl +pers1)
-vsubj: ions4* (v +stem -inf) (v -inf +subj +impf +pl +pers1)

The second step is to apply **verb stem rules**. Verb stems deal with collections of verbs (from 2 to 30) that conjugate in the same way. For example, "venir" (to come) and "tenir" (to hold), which have 27 derived forms, are handled by rules for "-enir" verbs. For this verb category, there are six stems as shown in (11):

(11) a +vstem: ien3enir* (v +inf -stem) (v +stem -inf) v-ien-
b +vstem: en0ir* (v +inf -stem) (v +stem -inf) v-en-
c +vstem: ienn4enir* (v +inf -stem) (v +stem -inf +subj) v-ienn-

d +vstem: in2enir* (v +inf -stem) (v +stem -inf) v-in-
e +vstem: în2enir* (v +inf -stem) (v +stem -inf +psimp) v-în-
f +vstem: iend4enir* (v +inf -stem) (v +stem -inf) v-iend-

Thus, in the analysis of "venions" ((we) came), first, the system identifies the verbal suffix common to all the verbs of the third group (10) giving the information on mood-tense-person(s); second, it recognizes the infinitive from the stem (11)b "ven-". The result is:

(12) venions
venir(VERB SUBJ PRES PERS1 PLUR STORED CLASS3 (ADDENDA 1) (LEMMA venir) (STRUCTURE ⟨⟨*⟩V -vsubj⟩V))
venir(VERB IND IMPF PERS1 PLUR STORED CLASS3 (ADDENDA 1) (LEMMA venir)
(STRUCTURE ⟨⟨*⟩V -vimpf⟩V)))

3.2. Derivation

Work in derivational morphology is currently in progress. Derivation raises some interesting theoretical and practical questions that we will discuss above. Like inflection, derivation allows, on the one hand, to recognize derived forms that we would not find in a dictionary and, on the other hand, to handle coinages. Actually, the coinage question is too complex to handle because of the varied nature of productivity. We will show a few examples below.

Consequently, two different sources of information have been used to write the derivational rules: some grammatical knowledge (Grévisse 1986), and a corpus of Canadian French containing well over 140,000 word types. The prefix rules are written as follows:

(13)a
a #: al* (a) (a) amoral
b co #: co2* (v) (v) coexister
c pre #: pré3* (v) (v) preetablir
d re #: re2* (v) (v) redonner
e multi #: multi5* (a) (a) multinational
f pluri #: pluri5* (a) (a) pluridisciplinaire
g post #: post4* (a) (a) postsecondaire
h poly #: poly4* (a) (a) polytechnique
i anti #: anti4* (n) (n) antiboycottage
j super #: super5* (n) (n) superpetrolier

The rules from (13)a to (13)d refer to existing words in French whereas (13)e through (13)j refer to new formations.

Prefix rules do not change their part of speech: "amoral" (amoral) is an adjective like "moral" (moral) is; "redonner" (to give again) is a verb like "donner" (to give). However, the semantics of prefixes is an open question: if we define a rule "dé-" for words like "charger" (to load) and "décharger" (to unload), can we accept that "détailler" (to detail) is derived from "tailler" (to cut)?

On the other hand, suffix rules change their part of speech as given in (14):

(14) a #itude: itude5* (a +masc +sg) (n +sg -masc +fem) platitude
b #onner: onner3* (n +sg +masc -fem) (v +inf -masc -sg-inv) échelonner
c #al: al2* (n +masc +sg) (a +sg +masc -inv) gouvernemental
d #al: a12e* (n +fem +sg) (a +sg +masc -fem) pyramidal

The substantive "platitude" (flatness) is derived from the adjective "plat" (flat) in (14)a. As opposed to prefix rules, the semantics of suffixes is simple. However, the attachment to the stem is sometimes complex: to derive "bifurc-ation" (bifurcation) from "bifurqu-er" (to bifurcate) we need additional spelling rules; to derive "bouvier" (herdsman) from (ox) "boeuf" (ox) some additional stem rules are neccessary.

An experiment is currently in process in the Speech Recognition group for automatically labelling French texts. This experiement utilizes the forward—backward algorithm based on statistical and probabilistic approaches. We rewrote the morphological output into 95 tags and built a dictionary of over 112,000 entries from the Canadian French corpus. Although the work is not finished, the results are close to 98% of correct tags.

4. Future Plans

We have tested the inflectional module with the Canadian French corpus. (The corpus size is close to about 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. The derivational module which is a more endless task still needs some improvements.

Complementary work has to be started in generation. Words can be generated from the form recognized by the morphological analyzer. This facility will prove useful when working with dictionaries or texts. The whole system will have a lot to offer the translator, lexicographer, student and researcher in NLP.

5. Conclusion

Because of the rich productivity of French morphology, the analyzer can improve the access to the dictionary by an expansion factor of five. We calculate a factor of 4 for most of the adjectives, a factor of 2 or more for the nouns and from 9 for verbs like "pleuvoir" (to rain) to 40 for regular verbs. The morphological analyzer for English is presently implemented in the WordSmith system of dictionary facilities at IBM Research (Neff and Byrd 1987). We plan to add the French system described here and to link it to the Collins French-English dictionary. The numerous users of French and French-English dictionaries will have important new facilities available to them. Moreover, because the system lemmatizes the forms, using an on-line dictionary will be made easier.

Notes

[1] The morphological rules consist of six parts: a) a boundary specification ("+" or "—"), b) an affix name, c) a pattern, d) a condition, e) an assertion and f) an example. Consider the rule (1) above. It is a inflectional rule (-) with boundary specification "vsbuj". It applies to words ending in "-isse", for example, "dormisse, finisse, etc". The characters "-isse" must be removed and the word matched against the condition (v + stem -inf). That is, in the case of "dormisse", the non-word stem "dorm" that is formed; the program continues and checks for a + vstem rule such as (2). If the condition matches the rule asserts that the input word "dormisse" is the first person singular of the imperfect subjunctive of the verb "dormir".

References

Cited Dictionary

COLLINS ROBERT FRENCH-ENGLISH ENGLISH-FRENCH DICTIONARY. 1978. B.T Atkins, A, Duval et al. (eds.). London and Glasgow:Collins Publishers.

Other Literature

Anshen, F. M. Aronoff, R. J. Byrd, and J. L. Klavans. 1986. 'The Role of Etymology and Word Length in English Word Formation' in *Proceedings of the Conference "Advances in Lexicology."* Center for the New OED, University of Waterloo, Canada.

Aronoff, Mark. 1976. *Word Formation in Generative Grammar.* Linguistic Inquiry Monograph 1. Cambridge, Massachusetts: MIT Press.

Boguraev, Branimir and Ted Briscoe. 1986. *Large Lexicons for Natural Language Processing: Utilising the Grammar Coding System of LDOCE.*

Byrd, R. J. 1983. 'Word formation in natural language processing systems' in *Proceedings of IJCAI* VIII, 704—706.

Byrd, R. J. 1985. *UDICIT Users Guide.* IBM Research Report. Unpublished Lexical Systems project report.

Byrd, R. J. 1986. *Dictionary Systems for Office Practice.* IBM Research Report RC 11872.

Byrd, R. J., J. L. Klavans, M. Aronoff, and F. Anshen. 1986a. 'Computer methods for morphological analysis' in *Proceedings of the Association for Computational Linguistics.* 120—127.

Byrd, R. J. and E. Tzoukermann. 1988. 'Adapting an English morphological analyzer for French' in *Proceedings of the Association for Computational Linguistics.* 1—6.

Corbin, Danielle. 1987. *Morphologie dérivationnelle et structuration du lexique.* Tübingen: Max Niemeyer Verlag.

Grévisse, M. 1986. *Le bon Usage.* Gembloux: Duculot.

Klavans, Judith L. and Nina Wacholder. (forthcoming) *Features and Attributes in the UDICT Lexicon.* IBM Internal Research Report.

Michiels, Archibal. 1982. *Exploiting a Large Dictionary Data Base.* Unpublished PhD Dissertation. Liege, Holland: University of Liege.

Neff, M. S. and R. J. Byrd. 1987. *WordSmith Users Guide.* IBM Research Report. Yorktown Heights, New York: T.J. Watson Research Center.

Tzoukermann, Evelyne. 1986. *Morphologie et génération des verbes français.* Unpublished PhD Dissertation. Paris, France: Institut National des Langues Orientales.