THE USE OF FINITE AUTOMATA IN THE LEXICAL REPRESENTATION OF NATURAL LANGUAGE

Maurice Gross

Université Paris 7

Laboratoire d'Automatique Documentaire et Linguistique [1]
Centre d'Etudes et de Recherches en Informatique Linguistique[2]

Finite automata are tools which are well adapted to the representation of phenomena observed at various levels of the description of natural languages.

There are numerous cases where an utterance (word, phrase or sentence) is subject to formal changes that leave invariant its essence, in general its meaning. These changes generate new utterances that all share features, they must then be grouped into families or equivalence classes. Some of these cases correspond to natural linguistic phenomena subject to rules, others, such as spelling, are more artificial. Handling all of them is essential for computer processing of texts.

Much emphasis has been put in the literature (both linguistic and computational) on the recursive linguistic phenomena that are not finite-state, hence promoting context-free or recursively enumerable grammars as the major models of language (N. Chomsky 1959). At this point, linguists seem to have forgotten that the bulk of known linguistic phenomena is indeed finite-state, in syntax as well as in phonology.

This statement can be made on the basis of the description of written French elaborated at the LADL-CERIL and not from isolated examples, the representativity of which is always a question. We study here typical cases of commonly encountered phenomena that have immediate implications for natural language recognition.

O. An elementary step of linguistic analysis

Computerized texts are processed in order to locate some information, for example, to determine a set of documents pertaining to this information or else in view of translating them. The first processing step then consists in recognizing words. The simple words of a text are rather well defined: strings of certain characters between consecutive separators[3]. But words undergo grammatical variations, it is thus necessary to reconstruct from the inflected forms of words found in texts the normal forms of words which constitute the dictionary entries (which can be roots or words). This activity, often called lemmatization, consists essentially in cutting off grammatical suffixes from occurrences of inflected forms and in checking compatibility between roots and suffixes. Since suffixes depend on word classes and often on individual words, a complete dictionary of the language has to be built in order to represent all and only the possible combinations of roots and suffixes.

We call the electronic dictionary of simple words, the dictionary of word entries (or roots) containing the grammatical codes that determine all and only the inflected forms together with their grammatical value(s). The dictionary should meet the following condition:

---

1. Unit 819 of the CNRS.
2. FIRTECH Industries de la langue française.
3. There are however difficulties in drawing a clear line between the set of computer characters that constitutes the alphabet of the text and the set of separators (cf. M. Silberztein 1988, present volume).

(T) Given a text and a recognition procedure of words that uses the electronic dictionary, all the simple words of the text should be recognized, that is, there should not be any failure of the dictionary look up process.

This ideal requirement should be corrected to take into account two important facts about texts:

- they contain mispelled words,
- they contain proper names, more generally arbitrary forms (symbols, numerical forms) that cannot be listed a priori in a dictionary.

Such a dictionary system is being built for French, it is called DELA (B. Courtois 1987), and presently contains about 65.000 entries which generate more than 500.000 inflected forms. The numerous gaps in ordinary (paper) dictionaries, that is, their fundamental inaptitude to meet the test (**T**), are being bridged progressively, and more information than the minimal grammatical features we mentioned is being introduced (M. Gross 1988). In particular, codes that systematize derivational processes are added. For example, from a verb such as *to abbreviate* one can derive two other verbs: *to reabbreviate* and *to disabbreviate*; from these verbs one can form the three adjectives: *abbreviatable, reabbreviatable, disabbreviatable*. These derivatives are missing in current dictionaries, but they must be entered in an electronic dictionary.

Listing such words alphabetically in a dictionary provides satisfactory computational solutions for retrieving them. But when one deals with such families of words, the regular syntactic and semantic relations that define them have to be recorded in view of processing operations that go beyond simple retrieval of the words. In our example, the semantic effect of affixation by *re-*, *dis-*, *-able* being general, the meaning relations between words (i.e. **repetition, suppression, potential**) that mirror the morphological relations could be used in formal procedures, hence should appear in the dictionary.

This example illustrates only one of the many reasons that lead us to entirely revise the notion of dictionary for the automatic treatment of texts and that force us to define precise representations for their entries.

1. Spelling

1.1 Spelling of simple words

In principle, the spelling of *simple* words is well standardized in most of the languages that use a Roman alphabet. It is a necessary condition to the existence of convenient dictionaries. However, many variations are allowed:

- for example the use of capital letters as variants of small case letters in certain contexts: inside English titles, for all words at the beginning of a sentence, etc.;
- in French, accents may be omitted from capital letters.

Thus a word such as *été* which means both "summer" and "been", is spelled either *Été* or *Eté* when it occurs at the beginning of a sentence. Hence the three possible spellings represented by the graph of figure 1.

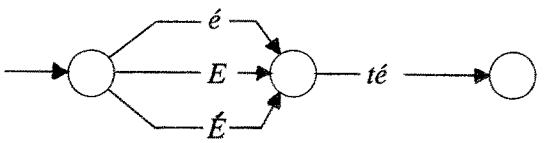


Figure 1

Notice that for other stylistic reasons (title, emphasis, etc.) words may have to be spelled entirely in capitals letters. Our example then has the two other spellings:

ETE, ÉTÉ

which can be incorporated into the preceding graph, yielding:

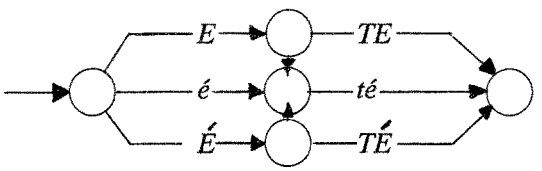


Figure 2

A recognition system for words must take into account the fact that there are 4 variants for the standard word form *été*[4]. Ignoring altogether differences in cases and accents renders equivalent all the forms of the graph of figure 2. As a matter of fact, early computer and communication systems have used this approximation. But from a linguistic and computational point of view, this solution is inadequate. In the particular example of *été*, using *ETE* as the canonical form for the equivalence class of spellings has no special implication for the recognition process of this word, but for the form *étés*, which only means "summers", using the canonical form *ETES* would introduce an artificial ambiguity with the verbal form of *être* (to be): *êtes* (are).

By transforming the electronic dictionary of French DELA and comparing the different versions, one can enumerate the artificial ambiguities generated by such approximations (S. Woznika 1987): The result is that more than 10 % of the inflected words, that is 50.000 words, become ambiguous[4].

A similar experiment has been performed by transforming the four "accents" of French (i.e. acute, grave, circumflex and umlaut) into one single diacritic sign: the "flat" accent. In this case, only a few dozen inflected words become ambiguous; for example, *prés* (meadows) and *près* (near), *pécher* (to sin) and *pêcher* (to fish); *dés* (dice) and *dès* (as soon) take on the same form *dès,* but no ambiguity with *des* (some) is generated as with the suppression of all accents. No ambiguity results for the words *êtes* and *étés*.

Whatever the practical solution adopted for recognizing words in texts, representations such as those of figure 2 will have to be used, at least in the ambiguous cases.

Another type of spelling variants occurs with words which are not yet well established in the vocabulary of the language, this is the case for foreign imports or with slang words which begin acquiring respectability when exhibited in print. For example the word *kosher* can be found in French spelled at least in any of the following forms (M. Mathieu-Colas 1987):

[4]. In principle, other mixtures of cases are not found in the non technical vocabulary (cf. M. Silberztein 1987).

kasher, casher, kashère, cashère, kacher, kachère

They are all pronounced in the same way and it is clear that the variations are due to the various transcription solutions allowed by French. As a consequence, the forms can be organized in a graph in a natural way:

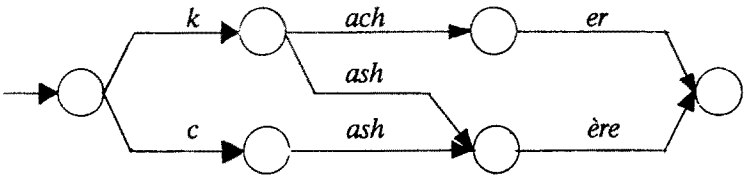


Figure 3

The shape of the graph avoids the combination of *ach* with the initial *c*, perhaps because of a potential ambiguity: *cacher* is a verb (to hide). In the same way one could easily add 6 variant forms with an *o* or *aw* instead of the *a*. With the same restriction on *c*, the modified graph of figure 3 would not introduce the two forms *cocher* (coachman) and *cochère* (about coach) which already exist, thus avoiding two ambiguities.

Remark.

The representations could be extended to spelling mistakes which are recognized as frequent or that can be expected sooner or later. For example, the spelling of *kasher* with *kh* instead of *c* or *k* is not found and would probably be considered as a mistake. One can expect to find it, given the already wide range of variations more or less accepted. This new possibility can be added to the graph of figure 3.

1.2 Spelling of compound words

While recording and spelling simple words is a well normalized activity, it is not the case for compound words in French. One reason could be the fact that compounds are not systematically entered into dictionaries. Since they are composed of at least two simple words, entering them together with one or the other simple component is an art that lexicographers practice with wide variations.

All the spelling problems mentioned for simple words also affect compounds. But new factors intervene. For example, in French, the use of hyphens in compound terms is not subject to rules. Consider the ways the compound noun *Middle Ages* is spelled in current French dictionnaries:

- it can be found with a hyphen or a space between the two simple words,
- it can be spelled with initial capitals or not.

Hence we find: *moyen(-)âge, Moyen(-)Age, Moyen(-)Âge, Moyen(-)âge*. These 8 forms are generated by the graph of figure 4:

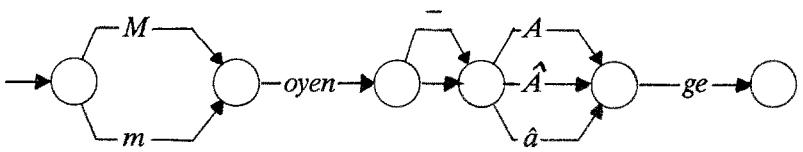


Figure 4

However, the 4 shapes:

moyen(-) Age, moyen(-) Âge

are not observed, either they should be substracted from the set generated by the graph, or else a different graph should be used as in figure 5:

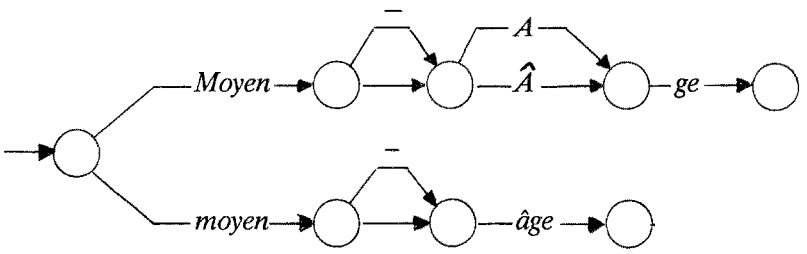


Figure 5

Remark

Notice that the "missing" forms cannot be excluded by a general rule constraining the use of capital letters in compound nouns. In fact, the normal spelling of Pacific Ocean is with one capital on the second word: *océan Pacifique*.

Another source of variation for French compound words is the plural mark, in general an *-s* which is not pronounced. Hence, detecting its presence is a problem of syntactic analysis of the context, but often, this analysis fails to provide an answer. Consider for example the compound *un simulateur de vol* (a flight simulator), there is no phonetic, syntactic or semantic reason not to spell it *un simulateur de vols*. As a matter of fact, there are semantic rules (taught in school!) which determine the number of some compounds: for example *un essuie-mains* (a hand towel) is spelled with *mains* in the plural "because one usually dries both hands when one uses it".

In the same way, a *presse citron* (a lemon crusher) is spelled with *citron* in the singular, "because one crushes only one lemon at a time"[5]. This line of reasoning can lead to the spelling *vols*. However, the singular form is also observed and we thus have to represent the compound with an optional *-s* on *vol*[6].

[5]. Such "rules" never reflect reality, for example it is never the case that whole lemons are crushed, they are always cut in halves each of which is crushed.

[6]. Notice that the problem almost arises in English: *flight simulator* or *flights simulator* ? Since the second word begins with an *s*, the interdiction of the plural of *flight* may be hard to perceive. In the general case however, say with *baby boom*, it is clear that *babies boom* is not a possible variant.

2. Derivational morphology

As already mentioned, families of words related by prefixation and suffixation processes can be organized by means of a finite-state graph. Such a representation is particularly interesting in certain productive cases. We now study an example that presents a double productivity:

- nouns of countries or ethnical names range in the thousands. For example, the word *France* belongs to a set of proper nouns which can grow without any limit,
- such nouns give rise to a wide variety of derivatives. For example, we will see that at least 30 simple words can be formed on *France.* The two processes combine: namely each of the thousands of proper names enters derivational processes similar to those of *France* which we now examine.

Several subfamilies of such words are represented in figure 6:

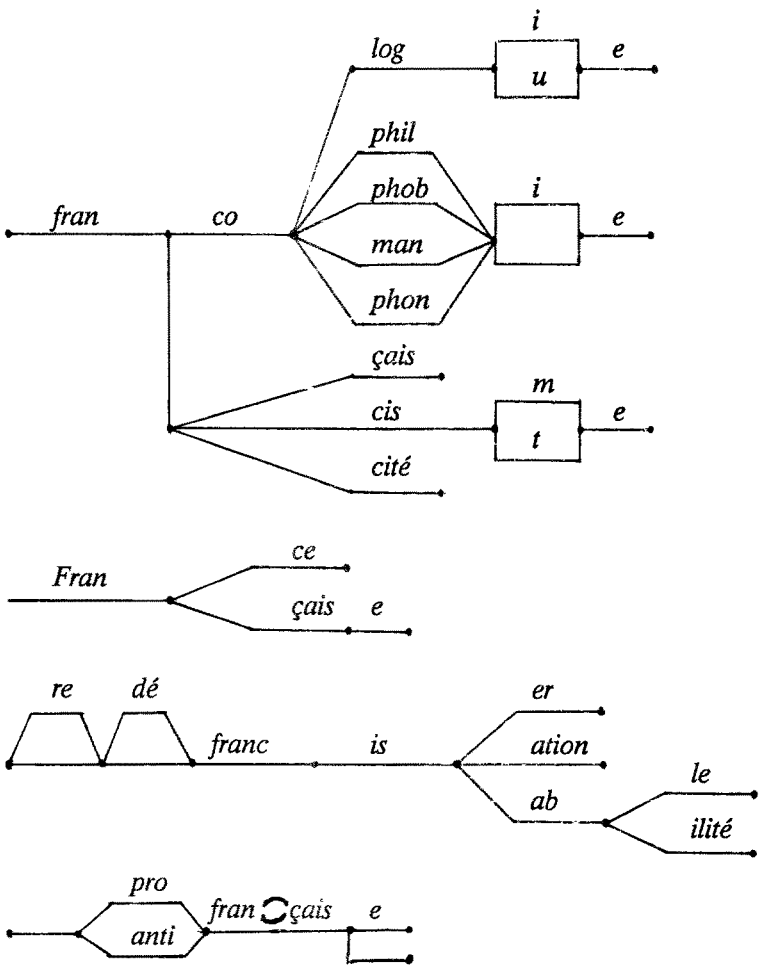


Figure 6

- a group of words formed on the form *franco-* with suffixes carrying obvious meanings,
- a group of words built only with suffixes on the root *fran* (in fact *franc*, but the two spellings *ç* and *c* of the sound *s* are separated);
- two words with capital *F*: *France* and *Français* (a Frenchman);
- a group of prefixed and suffixed words built on *-franc-is-*, that is a verbal derivative of the adjective *français*;
- a group of adjectives with prefixes *pro-* and *anti-* which can also be spelled as compounds, with a blank or a hyphen.

These graphs do not provide a complete representation of the forms of the family. There are in fact more simple words associated to the subgroup of the verb *franciser* (to make something French) which must be conjugated, yielding about 40 new forms; the noun *Français* (Frenchman or Frenchmen) and the adjective *français* both exist in the feminine and feminine plural. However, the noun *français* (French language) has no feminine form; also, the unmarked plural of this noun is semantically different from the plural of the noun *Français*. Moreover, the representation makes it difficult to introduce the prefixe *in-* which applies only to a subset of the words derived from the verb. For example, we have *inrefrançisable* (which cannot be made French again), whereas *inrefranciser* is not accepted[7]. Also, *Francophonie* is spelled with capital *F* (the meaning is roughly "the community of French speaking countries".

All the relations represented in figure 6 and the restrictions mentioned tend to mirror, but not always, syntactic relations that preserve meaning (J. Dubois 1962, D. Corbin 1988). More precisely, each meaning of the words should be described by means of an elementary sentence and related to other forms by formal transformations. Relations between words will then appear as particular restrictions to words of the syntactic transformations. We now list the various elementary sentences and the transformations involved.

There is a series of sentences with human subject:

Max est un franco (mane + phile + phobe)
[NA] = *Max est franco (mane + phile + phobe)*
[AN] = *Max a une certaine franco (man + phil + phob)ie*

[NA], the first relation transforms the noun into an adjective; [AN], the second relation is a nominalization of the adjective (A. Meunier 1977). The following sentence is also obtained by nominalization:

Max est un francologue (?un francologiste)
[NN] = *Max fait de la francologie*[8]

We have as above:

Max est un francophone
[NA] = *Max est francophone*

As already mentioned, the noun *Francophonie* has acquired an autonomous meaning (the set of French speaking countries), it is no longer related syntactically to

7. For a detailed discussion, cf. M. Gross 1988.
8. These forms may sound odd to French speakers, but they are the equivalent of well accepted pairs such as:

Max est sinologue
= *Max fait de la sinologie*

the noun or to the adjective *francophone*, the relation is etymological. The verb *francophoniser* is used in Quebec, derivatives such as *infrancophonisable* or *francophonisation* follow mechanically.

The human noun *Français* is related to the human adjective by means of the general rule [NA]:

Max est un Français (Max is a Frenchman)
[NA] = *Max est français* (Max is French)

(notice the use of capitals).

The non human adjective *français* (French) is to be distinguished from the human adjective of citizenship which has an identical form. It is this non human adjective that enters in relation with the causative verb *franciser*. We have then the following complex derivation (M. Gross 1981):

(Ce mot + Ce produit) n'est pas français
[Caus. *rendre*] = *Max rend français (ce mot + ce produit)*
[Caus. *-is*] = *Max francise (ce mot + ce produit)*[9]

Then the verb *franciser* has variant forms generated by morpho-syntactic processes. The suffixes that appear in figure 5 correspond to syntactic relations such as:

On francise ce mot
[*pouvoir* i.] = *On peut franciser ce mot*
[Passive] = *Ce mot peut être francisé*
[*Adj-able*] = *Ce mot est francisable*
[*N-abilité*] = *Ce mot a une certaine francisabilité*

On a francisé ce mot
[*N-ation*] = *On a fait la francisation de ce mot*
[Passive] = *Ce mot a été francisé*
[*N-ation*] = *Ce mot a (eu + subi) une francisation*

[*être N-ation*] = *Coquetel est une francisation de cocktail*

One could also propose the relation:

Ce terme est un francisme (This term is French-like)
[AN] = *Ce terme a une certaine francité*
(This term belongs to the French domain)

Although syntactically motivated, this relation does not account for the meanings (sometimes unclear) of these word forms. It should be noted that we have dealt here with two different meanings of *francisation*, that is, two different words that cannot be clearly distinguished by morphological processes.

The prefixes *re-* and *dé-* apply here in a fairly regular way: *re-* can mean "again" and *dé-* indicates a "reverse" process. We mentioned the "negative" suffixe *in-* which only applies to the three adjectival forms in *-able*, and thus cause the representation to be more complex. The relations involved are of the type:

9. Actually, the analysis should be more complex in order to account for the syntactic and semantic similarity between *franciser* and the operator verb *transformer* as observed in the pair:

Max a transformé ce mot étranger en un mot français
Max a francisé cocktail en coquetel

[*re-V*]	*Ce mot peut être refrancisé*
[*Nég* i.] =	*Ce mot ne peut pas être refrancisé*
[*in-Adj-able*] =	*Ce mot est inrefrancisable*
[AN] =	*Ce mot n'a aucune refrancisabilité*

The proper human noun Français (*Frenchman*) is related to *France* and to the human adjective *français* (French). One of the reasons for isolating this noun-adjective pair is its property of accepting the prefixes *pro* and *anti* which are not accepted with the other meaning (i.e. with non human subjects). Underlying this prefixation are the syntactic relations:

(Bob + ce livre) est (contre + pour) (la France + les Français)
= *(Bob + ce livre) est (anti + pro)français*
(Bob + this book) is (against + for) (France + the French)

It should be noted that the rules used so far in our relations involve support or operator verbs (e.g. *être, avoir, pouvoir, rendre*), that is, grammatical verbs with limited semantic function (e.g. modality, causative). Hence, the syntactic relations we presented do not make meaning explicit. In certain cases, we may need to make the relations more specific by introducing elementary sentences which contain verbs that carry complete meanings. Z.S. Harris 1976 has proposed such a device as an explanation of the meaning, and at any rate, as an etymological possibility. For example, one could introduce pairs such as the following, in order to complement the syntactic derivations given above:

Bob étudie la France (Bob studies France)
= *Bob est un francologue* (=Bob is a francologist)

Bob (aime + hait) la France (Bob (loves + hates) France)
= *Bob est franco(phile + phobe)* (=Bob is franco(phile + phobe))

La France obsède Bob (France obsesses Bob)
= *Bob est francomane* (=Bob is francomaniac)

Bob parle le français (Bob speaks French)
= *Bob est francophone* [10] (=Bob is francophone)

Bob étudie le français (Bob studies French)
= *Bob est un franciste* (=Bob is a specialist of French)

These semantic relations involve explicit sentences which could be used as semantic representations for the equivalence classes of sentences generated by the syntactic transformations. These representations are summed up in the table of Figure 7.

[10]. This nominal sentence is more specific than the verbal one, from an aspectual point of view: there is a permanent component in the fact that Bob speaks French.

Derived words	Semantic predicates	Basic words
francologue	**étudier**	*la France*
francomane	**obsédé par**	*la France*
francophile	**aimer**	*la France*
francophobe	**détester**	*la France*
francophone	**parler**	*le français*
franciste	**étudier**	*le français*
franciser	**rendre comme**	*français*

Two words of the third column are linked by a relation such as: *Le français se parle en France* (French is spoken in France).

Figure 7

One of the reasons for structuring this set of words is its productivity. Hundreds of other names of place can enter the same type of graph. However every such graph will have to be constructed "manually", mechanical substitution of roots would run into difficulties due to numerous idiosyncrasies. For example, the correspondance between country names and language names or between countries names and citizen names is not one-to-one. There are social or political groups other than countries, their names give rise to similar families, with gaps (e.g. no language name). Morphology can be fairly regular with differences such as in the pairs: *English-anglo, American-americano*. But they can also be highly irregular. Already, pairs such as *Fran(ce + çais)-franco*, where the reason for the sound /k/ in *co* is not clear, are not uncommon: *Belgique-belgo* raises a similar question for the sound /g/ in *go*. And often, some forms are unrelated in shape to the intended root: there are series such as:

Allemagne (N),	*allemand (N, Adj), germanique (Adj),*	*germano*
Espagne (N),	*espagnol (N, Adj), hispanique (Adj),*	*hispano*
Suisse (N),	*suisse (N, Adj), helvétique (Adj),*	*hélvéto*

Such forms are often complementary, in the sense that if we wanted to construct for *Germany* the dictionary article we elaborated for *France*, we would have to replace the words constructed on *fran* by words constructed sometimes on *alleman*, sometimes on *german*, sometimes on both (e.g. *germanophile*, **allemanophile*, whereas *études allemandes* and *études germaniques* are both accepted and mean both German studies).

There are other families of compounds that raise similar problems, for example, scientific terms in the fields of physics, chemistry, biology, medicine, etc.

An interesting situation arises with all these families of words when they are composed in the following way: forms ending with *-o* can be considered as adverbs which modify adjectives derived from place, ethnic or field names. For example, we obtain compound adjectives as the product of the set of adverbial prefixes by the set of adjectives of nationality:

(P)	*anglo*	*anglais*
	franco	*français*
	germano	*allemand, germanique*
	hispano	*espagnol, hispanique*
	luso	*portugais, lusitanien*

The interpretation of the compounds is that of conjunctions, with variations that depend on the noun to which the combination is attached:

- *un mélange franco-anglais* is a mixture of French and English stuff,
- *un accord franco-anglais* is an agreement between France and England (Great Britain),
- *la frontière franco-anglaise* would be the boundary between France and England.

Products such as (P) can in principle be generalized[11] to any length, as in:

un accord anglo-americano-allemand

the general shape would then be:

(Q) **(anglo)^n-(anglais)**

representable by a finite automaton. The number of elements that can be substituted in the parentheses may not be finite.

In the same way, a productive family of adverbs is of the form:

à la (française + anglaise + italienne + etc.*)*

They can be analyzed as a reduction of the adverbs:

à la manière (française + anglaise + italienne + etc.*)*

where the *N* =: *manière* (manner, way) makes their meaning more explicit (i.e. *in the French way*).

The representation we outlined is far from complete, for there are many combinations frozen to various degrees involving the simple words *France*, *français*, etc. which have lost their original syntactic relations to these words. This is the case for example in *Revolution française* which has acquired a meaning of its own, *Banque de France*, the province *Ile de France* (and the derivative *ilofrancien*: inhabitant of this province), the city *Fort de France*, etc. Examples such as *études françaises* (French studies), *marine française,* (French Navy) seem different: on the one hand, the feeling is that one is dealing with compositional phrases, since productive syntactic patterns appear to hold:

études sur la France = études françaises
marine de la France = marine française, etc.

On the other hand, these relations do not preserve meaning exactly (i.e. the "institutional" character of the compound *NAdj* is not perceived in the source term.

3. Syntax

Similar problems of representation are raised by many combinations of words into sentences, solutions similar to those we applied to characters and parts of words can be used. The description of sentences then provides clear situations about the computational function of automata representations. We now present several cases of such formal descriptions.

11. Also, similar products exist, with different types of interpretation and combinatorial restrictions for other large parts of the vocabulary (e.g. socio-economic, cardio-vascular).

3.1 The grammar of French pre-verbal particles.

This example is a case of a strictly grammatical automaton, that is, a device enumerating a family of elementary sentences which have no meaning relations to each other. The automaton aims at generating all and only the sentences that contain non empty sequences of distinguished words, called pre-verbal particles (*Ppv*), also known as clitics, conjoined pronouns, etc. It should be noted that the justification for building a local grammar is linguistic in a specific way: the distinction between *Ppv*s and other pronouns is being formalized. When one studies sentences such as:

Elle ne le lui donne pas
(*She does not give it to him*)

one observes that no insertion is allowed in the sequences $(Ppv)^n$ V (V is the verb; the symbol * signals ungrammaticality):

**Elle souvent ne lui en donne pas*
**Elle ne souvent ne lui en donne pas*
**Elle ne lui en souvent donne pas*

whereas in English for example, similar insertions are accepted:

She often gives some to him

French thus presents a phenomenon of pronominalization of complements that has drawn the attention of linguists because of its complexity. For example, the sentence form with two complements. (N_1 and *à* N_2):

N_0 V N_1 *à* N_2 =: *Max montre ce lit à Luc*
(Max is showing this bed to Luke).

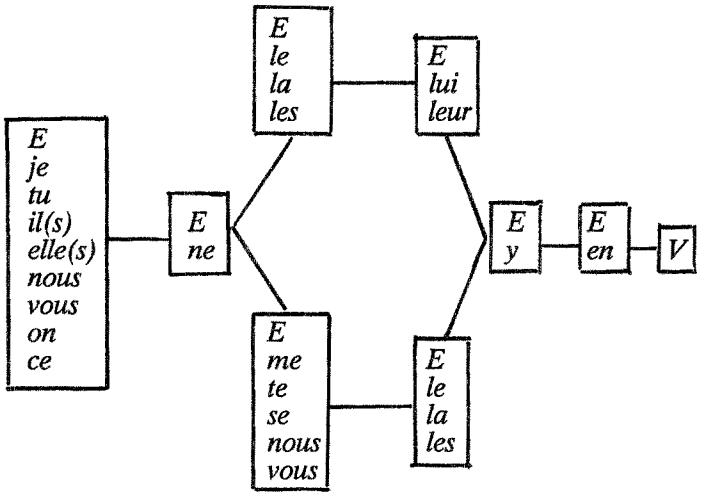
leads to pronominal forms such as:

Max le montre à Luc
Max lui montre ce lit
Max le lui montre

where, in appropriate contexts, complements have been replaced by pre-verbal pronouns. The arrangement of pronouns is complex, for example, if the direct object is indefinite, its pronoun (*en*) occupies a different position, as in:

Max lui en montre un
(Max is showing one to him)

Subject pronouns have to be included in the description. They introduce other restrictions linked to the grammatical features of the subject and of the complements. The number of complements that are sources of *Ppv*s is limited and verbs take at most three or four complements that are sources of *Ppv*s. Even if we add subject pronouns and the negative particles *ne* to the set of *Ppv*s, the combinatorial problem remains finite and even relatively small. It is quite natural to represent all and only the sequences of *Ppv*s by means of a finite automaton. The combinations of *Ppv*s have been represented on the condensed graph of figure 8. For reasons of convenience, the notations differ slightly from those of the preceding finite automata, but their formal equivalence should be obvious. D. Perrin 1989 has discussed for such graphs various processes of representation that are better suited to computer treatment.



The symbol *E* represents the empty word. This graph does not take into account elision nor agreement phenomena of reflexive verbs.

Figure 8

The graph of figure 8 is incomplete because of elisions that may occur when some of the *Ppv*s come into contact. For example, the particle *ne* changes its form to *n'* when followed by a vowel. The vowel can belong to another *Ppv* as in *n'en*, *n'y* or to the verb, as in *n'arrive pas*. Such constraints are strictly finite and could be incorporated in the finite automaton of figure 8, making it more complex. However, this elision phenomenon occurs in more general contexts than combinations of *Ppv*s. It is observed with articles, (*le*, *la*), prepositions (*de*), conjunctions (*que*) in combination with practically any word beginning with a vowel **v**. The rule accounting for this contraction can be written:

$$\begin{array}{ll} & \#\,(n + m + t + s + l + d + qu)e\,\#\,\mathbf{vx} \\ = & \#\,(n + m + t + s + l + d + qu)'\,\mathbf{vx} \end{array}$$

where the symbol # represents a blank. In other words, this formula can be seen as a transduction of sequences of characters. This transducer applies to the output of the automata of figure 8 providing all and only the correct sequences, and it may apply to many other syntactic forms.

Many independent linguistic phenomena interfere in general ways, they impose solutions by composition of automata. For example, so-called reflexive verbs (e.g. *se souvenir*) impose identity of person and number between the subject and an intrinsic *Ppv* which behaves exactly as a *Ppv* with an overt complement source:

je me souviens de cela,	*je m'en souviens* (I remember it)
tu te souviens de cela,	*tu t'en souviens*
elle se souvient de cela,	*elle s'en souvient*

**je (te + se + nous + vous) en souviens*
**elle (me + te + nous+ vous) en souvient*, etc.

Reflexivization depends on particular verbs such as *se souvenir* (*Vrflx*) which may have other *Ppv*s not involved in the process (e.g. *en* in the above examples). With other verbs, the process involves two complements:

*Elle se montre à Bob, *Elle se lui montre*

It is possible to represent all possibilities by graphs of the type of figure 8, however, a rule (i.e. a transduction) is sometimes more appropriate. For example, the agreement rule could apply to the string schema:

Ppv (*E* + *ne)* **Ppv** (*Ppv*)^{k} *Vrflx*

when the first *Ppv* is the subject, optionally followed by the negative particle *ne*, followed by the *Ppv* which must agree in gender and number with the subject, followed by *k* (= *0, 1, 2*) other *Ppv*s.

Let us now mention an application of this description of *Ppv*s to the automatic resolution of an ambiguity mentioned in 1, one of the systematic ambiguities generated by the suppression of accents. This ambiguity is seen in MANGES for example, which corresponds to both forms: 2nd person singular of the present tense: *manges* (eat) and masculine past participle: *mangés* (eaten). All transitive verbs of the first conjugation group, that is several thousands of verbs including most productive classes, have this ambiguity. However, the conjugated form *manges* necessarily implies the presence of the *Ppv* subject *tu* in its immediate context, namely:

- the interrogative form *manges-tu*, or
- the subgraph of figure 8 restricted to the subject *Ppv tu*.

Hence, mechanical disambiguation can be performed for shapes such as *MANGES* whose indeterminacy is detected by a dictionary look-up procedure: if the form has *tu* in its context, it is the finite verb, if *tu* is not found, it must be the participle form with accent.[12]

3.2 Frozen sentences

Another situation where finite automata find a general application is the description of elementary sentences which have frozen (idiomatic) parts such as:

Bob a (*nourri* / *réchauffé*) (*un serpent* / *une vipère*) (*sur* / *dans* / *en*) *son sein*

(1) *Bob* (*nursed* / *nourished* / *warmed* / *cherished*) (*a serpent* / *a snake* / *viper*) *in his bosom*

The subject, a human noun, is free, in other words it has a wide range of variations, but the other terms of the sentences are entirely constrained in English as in French. These sets of sentences (12 for French, 12 for English) are naturally organized by a finite automaton which then represents the meaning common to all variant shapes.

Other variations can be either included in the basic automata or introduced by finite transductions of the basic forms. For example, the dependency (identity of person-number) between the possessive adjective, and the subject is finite, in the sense that it holds between the two positions N_0 and N_2 of a finite structure of the form $N_0\ V\ N_1\ Prep\ N_2$. A transducer which would apply in many other similar situations could represent this agreement constraint:

You warmed a serpent in your bosom

12. Except, when additional ambiguities occur accidentally.

**You warmed a serpent in his bosom*

It is the limitation on the number of basic complements (never more than three) that makes sentence patterns strictly finite.

More varied forms of sentences are obtained when the syntactic positions where the frozen parts occur are varied. For example, the two sentences:

(2) *(God, The Lord) recalled Bob to Him*

both mean *Bob died*. They also have the form $N_0\ V\ N_1\ Prep\ N_2$, with N_0 and N_2 frozen[13], N_0 was free and N_1, N_2 frozen in example (1). The application of the [Passive] rule to (2) can be seen as a finite transduction of the structure $N_0\ V\ N_1\ Prep\ N_2$, which yields:

(3) *Bob was recalled to (God, the Lord)*

Example (1) may also undergo [Passive], as in:

(4) *Max is a serpent which was nursed in Bob's bosom*

(3) and (4) share a special feature of [Passive]: the pronominal elements (*his*, *him*) must be replaced by their full forms. Hence, two similar but not entirely identical transductions will have to apply to (1) and (2). Such differences are typical, they constitute one source of the enormous variety of syntactic forms allowable for a given meaning unit. Other formal changes come from the various adjunction processes (adverbs, relative clauses, noun complements, etc.). But in most cases, these adjunctions and the changes of word order can be described by composing finite transducers on strings of words or of grammatical categories.

3.3 Other forms

Many local grammars of phenomena that are stictly finite as in the case of French *Ppv*s have to be written. Some examples such as grammars of numerals (integers, fractions, decimal numbers) are well known. But often, they cannot be directly included in more complex grammars that describe numerical data: prices, lengths, weights, time, etc. The reason is that the particular units of price, length, weight, time, etc. constrain the numerals attached to them (e.g. hours in a day are only twenty four). Hence, specific grammars must be written in each situation[14], more precisely, specific composition operations of automata will have to be defined in each situation.

4. Conclusion

We have attempted to show that finite automata can adequately represent certain formal variations of linguistic units (from words to sentences). A number of questions arise immediately about any actual programme of description of a given language.

There are **linguistic questions**, such as:

- separating the meanings of each word, that is establishing a catalogue of ambiguities to be solved;
- describing the formal (i.e. morphosyntactic) variations possible for each word meaning;
- establishing reasonably complete lists of simple words and of processes of word formation;

[13]. A person-number agreement constraint may also hold between N_0 and N_2, but need not be represented since these two components cannot vary.

[14]. D. Maurel 1987, has written for french a detailed grammar of dates.

- establishing reasonably complete lists of compound words, including lists of elementary sentences whose variations are the object of a lexicon-grammar (M. Gross 1975; J.-P. Boons, A. Guillet, Ch. Leclère 1976).

All these questions are not independent. Thus a complete research programme must be defined to construct the linguistic system which will bear on the general vocabulary and on specialized vocabularies which may have their own linguistic properties. Without this complex device no significant picture of the language and no application can emerge.

There are also **computational questions** to be solved, such as:

- the choice of representations for families of strings. These representations will have to be implemented in computer programs;
- implementation of dictionaries of automata. These automata will be the basis of powerful look up procedures that attach to a particular word form general information invariant with respect to the family of strings defined by the automata;
- implementation of computer tools for constructing the dictionaries of automata, which includes maintenance tools for editing automata (adjunctions and suppressions of paths and word forms.

The experience we gained from the description of French indicates that many of these questions are still difficult to solve, even though the methods and the technology are available. Some of the questions we listed may look like problems already dealt with by specialists, they are in fact entirely new in most cases:

- traditional dictionaries are all phototypeset, that is available in computer form, nonetheless they are of limited use when an electronic dictionary has to be built,
- in the same way, traditional studies in morphology, namely in word formation, have never been pushed to the point where they could lead to automata representation, much less to dictionaries of automata,
- interesting software sometimes similar to that needed here already exists. But relational data bases, "tree editors" and automata systems such as the one H.Johnson and R. Kazman 1986 have developed for the Oxford English Dictionary do not seem specific enough to meet the demands an electronic dictionary puts on computer systems.

Many linguistic questions are language dependent, but formal and computer systems could be made general enough to be common to at least the main European languages. A coordinated effort of construction appears both necessary and feasible.

References

Boons, Jean-Paul, Alain Guillet, Christian Leclère 1976. *La structure des phrases simples en français, I. Constructions intransitives,* Genève : Droz, 377 p.

Chomsky, Noam 1956. Three Models for the Description of Language, *IRE Transactions on Information Theory*, IT2, pp. 113-114.

Corbin, Danièle 1988. *Morphologie dérivationnelle et structuration du lexique*, Thèse de Doctorat: Université Paris 7.

Courtois, Blandine 1987. *Le système DELA de dictionnaire électronique du français*, Rapport de recherches du LADL, Université Paris 7.

Dubois, Jean 1962. *Etudes sur la dérivation suffixale en français moderne et contemporain*, Paris : Larousse, 118 p.

Gross, Maurice 1975. *Méthodes en syntaxe*, Paris: Hermann, 414 p.

Gross, Maurice 1981. Les bases empiriques de la notion de prédicat sémantique, in *Langages*, N° 63, A. Guillet et C. Leclère éds. : *Formes syntaxiques et prédicats sémantiques*, pp. 7-52.

Gross, Maurice 1988. Sur la structure des articles d'un lexique-grammaire, *Linguistica Computazionale*, Pise.

Gross, Maurice 1989. La construction de dictionnaires électroniques du français, *Annales des Télécommunications*, Paris: CNET.

Harris, Zellig 1976. *Notes du cours de syntaxe*, Paris : Le Seuil, 237 p.

Kazman, Robert 1986. *Structuring the Text of the Oxford English Dictionary through Finite State Transductions,* University of Waterloo, Waterloo, Ontario: Internal Report CS-86.20.

Mathieu-Colas, Michel 1987. Variations graphiques de mots composés, *Rapport N° 4 du Programme de Recherches Coordonnées Informatique Linguistique*, Université Paris 7: LADL.

Maurel, Denis 1987. Grammaire des dates, *Mémoires du CERIL,* Vol. 1, CNAM et Université Paris 7, Paris : CERIL, pp.218-240.

Meunier, Annie 1977. Sur les bases syntaxiques de la morphologie dérivationnelle, *Lingvisticae Investigationes*, J. Benjamins B.V., Philadelphia-Amsterdam, pp. 287-332.

Perrin, Dominique 1989. Automates et algorithmes sur les mots, *Annales des Télécommunications*, Paris: CNET.

Silberztein, Max 1987. L'inversion de textes, *Rapport de recherche du Programme de recherches coordonnées Informatique Linguistique*, Paris: Université Paris 7: LADL.

Silberztein, Max 1988. The Lexical Analysis of French, present volume.

Woznika, Stanislas 1987. *Dictionnaire des homographes du français*, Rapport de recherches du LADL, Université Paris 7.