

## A bootstrap method for constructing local grammars

Maurice Gross  
LADL, Université Paris 7  
75251 PARIS CEDEX 05

Local grammars are finite-state grammars or finite-state automata<sup>1</sup> that represent sets of utterances of a natural language.

Local grammars have been used to describe a wide variety of sets of strings (M. Gross 1989 ; E. Roche, Y. Schabes 1997 ; W.A. Woods 1970), ranging from finite sets of words related by prefixation and suffixation (figure 1) to sets of sentences syntactically and semantically related (figures 2a and 2b). Formal representations of finite-state grammars are quite varied, although equivalent. As can be seen on figures 1 and 2, we have chosen a particular representation by graphs which is well-adapted to the syntax of natural languages. Using a particular example, we will present in detail a method for constructing local grammars using large corpora.

-----  
**Figure 1**  
-----

**Figure 2a**

The sentences represented in this graph, called **IssueTop**, describe the numbers of shares whose price rises or falls on the Stock market. The shaded box called **AdvUnchanged** contain adverbial forms that modify the main sentences. These adverbials are represented by a graph given in figure 2b.

-----  
**Figure 2b**  
-----

The graph is a subgraph of the graph of figure 2. When the graph **IssueTop** is applied to a text in order to locate the corresponding sentences, the graph **AdvUnchanged** is automatically called by the parsing program.

### 1. Dictionary entries

The example we construct is centered on the word *health*. Current dictionaries provide the following information about the notions expressed by the word entry *health*:

- 1) the well-being, mental or physical, of a person,
- 2) a metaphorical use of the primary meaning, as in the *health of the economy*, *financial health*, etc.
- 3) the use of the word in certain phrases or sentences:
  - a) *John is in (good + poor) health*
  - b) *Bob wished Eva a good health*
  - c) when proposing a toast: *To your health !*
- 4) examples of compound nouns where the word *health* is used more or less in connection with its meaning 1): *health club*, *health food*, *health insurance*, etc.

Such informal descriptions are far from sufficient for the automatic analysis of texts. Specialized dictionaries add to this description a certain number of technical terms that

<sup>1</sup> Equivalent to regular or rational expressions, or kleene languages.

may help a human user, but a computer program must have access to exhaustive lists of terms including their variants and abbreviations, in order to match the content of texts. Hence, an amount of information much larger than the content of published dictionaries is needed for applications such as information retrieval and mechanical translation.

The construction we are undertaking takes place in an environment where basic linguistic resources are already available for computer processing. We briefly present this environment:

The first tool is the dictionary of simple words. The simple words are those commonly encountered in texts, separated by delimiters, mainly spaces and punctuation marks. Words are inflected in many languages: verbs are conjugated, adjectives and nouns carry gender, number and case endings. Such forms can be extremely numerous, to the point that the processes that enumerate them may be difficult to formalize in a complete way. Dictionaries and morphological rules have the function of reducing sets of inflected words to one canonical form: the entry of a dictionary. For European languages, the number of inflected words ranges in the  $10^6$  (e.g. in French, the number is  $10^6$  to  $10^7$ , depending on the set of prefixes and suffixes taken into account in the morphological analysis, see figure 1), the number of the entries of current dictionaries of French and English ranges in the 100 000.

Isolating a simple word from its context, as parsing programs do<sup>2</sup> in a first step, can be quite misleading. Already the example of *health* shows the difficulty: *health* often translates into *santé* in French as *in good health* = *en bonne santé*, however *health food* is not translated by *nourriture de santé* but by *nourriture bio*, *health spa* is translated into *ville d'eau* or *ville de cure thermale* and the French terms do not make use of the translations of the English simple words. This phenomenon is often referred to as idiomaticity, but it ranges far beyond what is commonly understood by idioms.

As a consequence, dictionaries of compound words (i.e. generalized idioms) must be constructed for computer analysis, (and for language teaching as well), they are classified according to their main syntactic properties:

- adverbs: it is clear that *from time to time*, *as a matter of fact*, etc. must be represented as single units ; for French, about 15,000 compounds of this type have been described, a number to be compared with the 2 or 3000 simple word entries found in published dictionaries ;
- - adjectives can also be complex, as is *matter of fact* in the sentence *Bob is a very matter of fact person* ;
- verbs: many verbs are frozen with one or more of their arguments (subject, complement): *Bob took your point of view into account*, *Eva took the bull by the horns*, etc. In French, we described more than 25,000 such units, whereas the number of simple verbs is about 15,000;
- compound nouns are common items, idiomatic ones being the most conspicuous: *red herring*, *red tape*, etc. But in fact, all technical terms are of the same type: it is not possible to construct their meanings using the meanings of the simple words and syntactic rules that compose them. The exact meaning of a term (or description of a device) such as *electron tunnelling microscope* cannot be deduced from the three component words, even if the word *microscope* suggests the general line of functioning of the apparatus. Technical terms are numerous, in every field, not only in science and technology, there are millions of them;
- proper names. Texts contain numerous proper names, of place, of individuals, of institutions. Electronic catalogs of such terms exist, but they will have to be further elaborated to be usable in computer parsers of texts;
- numerical expressions. Utterances such as *ten million dollars*, *220 V* are common in texts, they involve grammars specific of each unit: *dollar*, *volt*, etc.

<sup>2</sup> Learners of foreign languages often operate in the same way : a word not understood by a reader is looked up in a dictionary, and its context can be difficult to match with the relevant information found in the dictionary.

In order to retrieve information from texts stored in electronic libraries, one must use full terms and not key-words such as *electron*, *tunneling* which are too ambiguous. At present, tools that search information, say on the WEB, are limited to keywords, hence, rather inefficient. The procedure we outline here is designed to cope with the problem of collecting, storing and using complex terms, most are nouns or nouns phrases. The goal is the analysis of large amounts of texts, in any field of knowledge.

## 2. Resources

### 2.1. Linguistic resources

The lexicographic data that are available in computer form as of today are the following:

- *published dictionaries*, they can be monolingual or bilingual. They are available in paper form. Publishers sometimes provide a computer version, hence they become readable on computer screens, that is, their printed form can be read from a disk or a CD-ROM on a screen. The screen can be printed, providing the original paper form. In some cases, it is authorized to download there dictionaries on a hard disk, and the files can be read by a word processor and processed. In this case, some data can be borrowed in computer form, but they still must be encoded before being added to electronic dictionaries ;
- *electronic dictionaries*. These dictionaries are built for use by programs, their content is made of alphanumerical codes which represent the grammatical data that can be reasonably formalized today. Figure 3 is a sample of such a dictionary;
- *corpora* have become increasingly available. Some texts (literature, newspapers) are being published on CD-ROM and electronic libraries are being built for them. But in the last few years, the INTERNET offers through almost any site linguistic materials, that is, large amounts of texts which can be easily downloaded.

A sample of terms extracted by **grep** from the electronic dictionaries available at the time of construction of the local grammar of *health*

academic:health:center(N1)/Bldg;Med/an  
 adverse:health:effect(N1)/Sick/an  
 American:health:security:act(N1)/Pol/the  
 comprehensive:health:reform(N1)/Med/a  
 economic:health(N1)/Eco/an  
 federal:health:official(N1)/Hum;Med;Pol/a  
 financial:health(N1)/Eco/a  
 fragile:health(N1)/Sick/a  
 health:care(N1)/Med/a  
 health:concerns(N1P)/Med;Psy/E  
 health:clinic(N1)/Bldg;Med/a  
 health:club(N1)/Bldg;Med/a  
 .....  
 health:care:coverage(N1)/Eco;Med/a  
 health:care:industry(N5)/Eco;Med/a  
 health:care:plan(N1)/a  
 health:care:professional(N1)/Hum;Med/a  
 health:care:provider(N1)/HumColl;Med/a  
 health:care:system(N1)/a  
 health:care:worker(N1)/Hum;Med/a  
 health:food:fair(N1)/Eco;Med/a  
 health:food:regime(N1)/Eco;Med/a  
 health:maintenance:organization(N1)/HumColl;Med/a  
 national:health:insurance:program(N1)/Pol/a  
 national:health:plan(N1)/Pol;Med/a  
 neighborhood:health:clinic/Conc;Med/a  
 preventive:health:care(N1)/Med/a  
 primary:health:care(N1)/Med/E  
 private:health:care:program(N1)/Pol/a  
 private:health:system(N1)/Pol;Med/a  
 standard:health:plan(N1)/Pol;Med/a  
 World:Health:Organisation(N1)/Bio/a

Each entry contains:

- a morphological code that describes the plural of the variable noun, for example *N1* stands for nouns that simply take an *-s*,
- semantic codes, such as **Med** for **medicine**, **Pol** for **political**, etc.
- a determiner, which indicates a phonological information (*an* is when the word begins with a vowel).

**Figure 3**

## 2.2. Standards

Published dictionaries and grammars are mostly the work of one author. It is impossible to merge two dictionaries or two grammars in order to improve or enlarge them. Each author has a personal point of view, difficult to spell out, which makes it impossible to reproduce the method(s) he used in the descriptions. Electronic dictionaries and grammars must be of a size and a precision that cannot be reached by the methods used for published works. Vocabularies and grammars have to be divided so that specialists can work independently on parts which will have to be merged coherently into a single system. Achieving this goal involves methodological, theoretical and practical constraints which result in a set of standards. We will not go here into the details of these requirements and will just mention two major principles:

- empirical evaluation of linguistic facts must be REPRODUCIBLE, a condition met in all hard sciences, but practically unknown in linguistics,
- the use of the formalism of finite automata, which has proven to be well adapted both to description work and to computation.

### 2.3. Programs

Computational tools and fixed file formats are necessary to accumulate data. A system called INTEX has been constructed at LADL (M. Silberztein 1993) which integrates the various file formats for dictionaries and finite state graphs. These formats are compatible with most European languages, including Serbo-Croatian whose dual alphabet requires special adjustments (D. Vitas & S. Krstev 1996).

### 3. The bootstrap approach

We present the way electronic dictionaries and grammars are constructed in practice, using the resources we described.

The initial steps are :

1) extraction, say, by means of the **grep** function, of the entries already available in the existing electronic dictionaries. We provide a sample of this search in Figure 3.

A sample of terms extracted by **grep** from the electronic dictionaries available at the time of construction of the local grammar of *health*

academic:health:center(N1)/Bldg;Med/an  
 adverse:health:effect(N1)/Sick/an  
 American:health:security:Act(N1)/Txt ;Law ;Pol/the  
 comprehensive:health:reform(N1)/Med/a  
 economic:health(N1)/Eco/an  
 federal:health:official(N1)/Hum;Med;Pol/a  
 financial:health(N1)/Eco/a  
 fragile:health(N1)/Sick/a  
 health:care(N1)/Med/a  
 health:concerns(N1P)/Med;Psy/E  
 health:clinic(N1)/Bldg;Med/a  
 health:club(N1)/Bldg;Med/a  
 .....  
 health:care:coverage(N1)/Eco;Med/a  
 health:care:industry(N5)/Eco;Med/a  
 health:care:plan(N1)/Med/a  
 health:care:professional(N1)/Hum;Med/a  
 health:care:provider(N1)/HumColl;Med/a  
 health:care:system(N1)/Med/a  
 health:care:worker(N1)/Hum;Med/a  
 health:food:fair(N1)/Eco;Med/a  
 health:food:regime(N1)/Psy;Med/a  
 health:maintenance:organization(N1)/HumColl;Med/a  
 national:health:insurance:program(N1)/Pol/a  
 national:health:plan(N1)/Pol;Med/a  
 neighborhood:health:clinic/Bldg;Med/a  
 preventive:health:care(N1)/Med/a  
 primary:health:care(N1)/Med/E  
 private:health:care:program(N1)/Pol ;Med/a  
 private:health:system(N1)/Pol;Med/a  
 standard:health:plan(N1)/Pol;Med/a  
 World:Health:Organisation(N1)/HumColl ;Med/a

Each entry contains:

- a morphological code that describes the plural of the variable noun, for example *N1* stands for nouns that simply take an *-s*,
- semantic codes, such as **Med** for **medicine**, **Pol** for **political**, **Hum** and **HumColl** for human individual and collective humans, that is, organizations, **Bldg** for building, etc.  
etc.
- a determiner, which carries phonological information (*an* is when the word begins with a vowel).

Figure 3

2) starting from a text, we search all occurrences of the word *health*. By means of the INTEX program, we build a concordance of this word (Figure 4 is a sample of the concordance).<sup>3</sup>

At this point, a linguist has to sort 'manually' all new occurrences according to their lexical status:

- some uses of *health* are free, namely, they are used in contexts which provide the information by means of the rules that combine *health* with the terms of its context, this the case for phrases such as : *Mr. Li's health, the health of that single child*, etc. Such forms are not retained, they should be analyzed by means of the general rules;

- other uses are frozen or institutionalized, as such, they should be stored in the lexicon in order to upgrade the coverage of processing.

By **lexicon**, we mean two different families of databases :

- **dictionaries** ; their entries are of the forms shown in figure 3. These entries have to be inflected so that they match text occurrences. In our sample of nouns of figure 3, each entry generates a singular form and a plural one, with the corresponding grammatical tags. In other terms, the plural form is lemmatized, that is, marked as equivalent (up to number) to the entry form (generally in the singular);

- **graphs**; each graph contains a set of related strings. For example, the graph of figure 1 represents the morphological family the words built on the root *tax*-.these word forms are often subdivided into :

- sets of inflected words, namely, conjugated verbs, nouns and adjectives inflected in masculine and feminine forms, they correspond to the rightmost column of boxes, each box contains corresponding grammatical endings;
- sets of derivationally related words, that is, relations between verb forms and noun forms, noun forms and adjective forms, etc.

Elsewhere (e.g. the verbs of figure 2), we represents each set of inflected words by a canonical word form (i.e. an entry) put between angles;

The paths between initial and final state are equivalent, the name of the equivalence relation is the (arbitrary) name of the graph.

When we study the forms involving *health* given by the concordance, we find different types of compound nouns :

- compounds that differ by their syntactic form, we have for example: *health minister, ministry of health, public health delivery system*,
- compounds that differ semantically, for example a *health minister* is an individual human, marked **Hum**, a *ministry of health* is a collective human body, marked **HumColl**,

We construct classes of terms based on these two parameters, all terms contain the word *health* modified to its left and/or right by various terms, these terms are listed in common boxes, they may be factors of other terms and can be reused in the same family of graphs or in others. For example, in figure xx, we find a box of terms starting with *<administration>* and ending with *task <force>*,<sup>4</sup> all these terms marked **HumColl** could be found combined with *food* instead of *health*, with perhaps the exception of *corps*.

At this point, the construction of graphs becomes quite empirical, the number of subclasses depends on each word, there is no reason why *food* and *health* should have the same number of meanings, hence similar grammars. The number of graphs depends on the number of meanings and terms found in dictionaries and in texts, it varies widely. Also, the decision to

<sup>3</sup> Concordances can be sorted according to either right or left contexts, which corresponds to searches of terms of different syntactic structures (e.g. *health resort, poor health*).

<sup>4</sup> We recall that the words between angles stand here for both the singular and the plural forms.

include a term in a dictionary rather than in a local grammar may be difficult to take, it is partly based on the number of observed related terms.

We distinguished several meanings for *health* (see § 1), they correspond to different sets of local grammars :

- the graphs of figures 5,6,7,8,9,10 correspond to the first meaning :
  - figure 5 contains the terms that refer to humans and collective humans,
  - figure 6 contains the terms that refer to political activities involving health,
  - figure 7 contains the terms that refer to medicine and related activities,
  - figure 8 contains the terms that refer to economic aspects of health,
  - figure 9 contains a variety of terms involving health : places, texts, people,
- the graph of figure 10 corresponds to the meaning found in *health food*,
- the graph of figure 11 corresponds to the meaning exemplified by *economic health*,
- the graph of figure 12 has a syntactic definition : it contains terms conjoined by *and*,
- the graph of figure 13 assembles all the previous graphs, each box corresponds to one of these graphs. Applying this graph, called **HEALTH**, to a corpus means parsing the text with the local grammar **HEALTH**. The result can be seen on the concordance of figure 14.

It is possible to count the number of different strings that the grammar **HEALTH** recognizes in its present state: over 33,000 strings (over 66,000 inflected forms). This number is due to the large number of combinations that occur between words. Extension of this grammar based on larger or specialized corpora could easily lead to sets of hundreds of thousand of strings, a combinatorial explosion fully mastered by the finite-state representation which contains dictionary entries rather than inflected forms and which factorizes the terms according to their common factors. In this way, the number of terms may grow, while the size of graphs grows at a much slower rate, that is, with the logarithm of the number of strings.

We have presented a method for constructing graphs around a keyword or equivalently around a semantic unit. The graphs can be used directly to parse texts and recognize meaningful strings, practically always disambiguated. The graphs can be further elaborated into transducers. A transducer is a finite state graph that associates outputs to recognized strings. Outputs are of a general nature, they can consist in a lemma associated to a family of strings, for example the longest string of the family, which, in general, is more explicit than abbreviations. Another type of output could be a translation into another language, either a single translation associated to the whole family (or to its lemma) or translations for each member of the family. The linguistic and computational power of such systems makes many linguistic applications realistic on a very large scale.

## REFERENCES

- Chomsky, Noam 1956. Three models for the description of language, *IRE Transactions on Information Theory*, IT-2, pp.113-124.
- Gross, Maurice 1972. *Mathematical Methods in Linguistics*, Englewood Cliffs N.J.: Prentice Hall Inc., 159 p.
- Gross, Maurice 1997. The Construction of Local Grammars, Roche, Emmanuel & Yves Schabes, eds. 1997. *Finite State Language Processing*, Cambridge, Mass. : The MIT Press, pp. 329-352.
- Gross, Maurice, Dominique Perrin eds. 1989. *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science, Berlin: Springer Verlag, 110 p.
- Silberztein, Max, 1993. *Dictionnaires électroniques et analyse automatique de textes*. Masson: Paris, 233 p.
- Vitas, Dushko & Svetana Krstev 1996. Tuning the text with an electronic dictionary, In *Papers in computational lexicography (COMPLEX 96)*, Budapest : Hungarian Academy of Sciences, pp. 267-276.
- Woods, W.A. 1970. Transition network grammars for natural language, *CACM*, 13(10), pp. 591-606.