

## The Construction of Local Grammars

Maurice Gross<sup>1</sup>

Laboratoire d'Automatique Documentaire et Linguistique

Université Paris 7

Grammar, or, as it has now been called, linguistic theory, has always been driven by a quest for complete generalizations, resulting invariably in recent times in the production of abstract symbolism, often semantic, but also algorithmic. This development contrasts with that of the other Natural Sciences such as Biology or Geology, where the main stream of activity was and is still now the search and accumulation of exhaustive data. Why the study of language turned out to be so different is a moot question. One could argue that the study of sentences provides an endless variety of forms and that the observer himself can increase this variety at will within his own production of new forms; that would seem to confirm that an exhaustive approach makes no sense. As a simple computation shows, there are more than  $10^{50}$  sentences having at most 20 words, a number which seems to deprive of any meaning the possibility of performing a systematic inquiry. However, the same could be said in Astronomy, Botany or Entomology, since the potential number of observations of individual stars, plants or butterflies is also limitless. Note that nonetheless, establishing catalogs of objects (and devising suitable criteria to do so) remains an important part of the activity in these fields. It is not so in linguistics, even if lexicographers do accumulate and classify words very much as in the other sciences. But grammarians operating at the level of sentences seem to be interested only in elaborating general rules and do so without performing any sort of systematic observation and without a methodical accumulation of sentence forms to be used by further generations of scientists. It is not necessary to stress that such an accumulation in any science is made possible by constructing and using suitable equivalence relations to eliminate

---

<sup>1</sup> I indebted to Morris Salkoff for important suggestions that improved this article.

what are deemed to be accidental variations, irrelevant to the specified goal of the catalog.

The approach in linguistics leads all too easily to overgeneralization. To take a well-known example, using the grammatical categories of Classical Greek (as taught in high schools) to describe exotic languages is more often than not utterly irrelevant. Another example is the way in which models of grammars have been introduced in linguistics. The earliest models of language dealt with sequences of grammatical categories, i.e. they formalized sentence forms where each word is replaced by its grammatical category. Such models succeeded in capturing in a natural way gross positional features such as the place of articles and adjectives on the left of their noun. Owing to its conceptual simplicity, this model has been repeatedly introduced under different names. It might be proper to call it the Markovian model, since its essential ingredients were introduced by Markov to study phonetic sequences. Such crude models do not go very far. At a more refined level, phrase structure models directly reflect the grammatical analyses taught in high school. N. Chomsky 1957 formalized them under the name of context-free grammars and demonstrated some of their fundamental inadequacies on the basis of carefully selected examples. In fact, as early as 1952, Z.S. Harris had proposed transformational grammars, which constituted a vast improvement over the Markovian and phrase structure models. But again, any of these types of grammar can be shown to have its validity restricted to the description of the linear order of words or grammatical categories with rather simple dependencies holding between them. Detailed attempts of systematic applications have revealed an endless number of subclasses of exceptions, each of them require a special treatment.

Short range constraints between words in sentences are crudely accounted for by Markovian models. But since Chomsky's mathematical proof of the inadequacy of the models (N. Chomsky 1956, M. Gross 1972), they have been totally neglected, and the interest has shifted to the essential problems of long range constraints between words and phrases.

An exception is the model of W. Woods 1970, which, however has not been used to attempt a full scale analysis of the language. This is precisely our present programme. It could be viewed as an attempt to revive the Markovian model, but this would be wrong, because previous Markovian models were aimed at giving a global description of a language, whereas the model we advocate, and which we call it finite-state for short, is of a strictly local nature. In this perspective, the global nature of language results from the interaction of a multiplicity of local finite-state schemes which we call finite-state local automata.

Our goal we repeat is very specifically to account for all the possible sentences

within a given corpus, and this, with no exception. The apparent obstruction evoked above to the realization of such a programme is avoided by the complexity of the various automata necessary for the description of the corpus. Examples will show what we mean by this admittedly loose presentation. It turns out that the long range constraints are taken care of automatically, so to say, by the interaction of the automata within the complete structured schema. We will see that these individual automata can be reused to describe other corpora. This is somewhat similar to the way small molecules combine to produce much larger ones in organic chemistry. To start with, we give elementary examples where the finite constraints can be exhaustively described in a local way, that is, without interferences from the rest of the grammar.

Consider some examples of adverbs. The following sentence form, where an adverb and an elementary sentence are combined, is not accepted:

*\*Democratically, Bob is authoritarian*

but the same form with an adjunction to the same adverb is accepted:

*Democratically speaking, Bob is authoritarian*

Many adverbs derived from adjectives are systematically accepted in the left context of *speaking* and of no other forms. The same adverbs are forbidden in the context of *saying*, *calling*, *talking*, although such words are morphologically and semantically similar to *speaking*. Alongside these productive forms, we observe combinations that are considered as frozen, such as:

*(broadly + generally + roughly) speaking*

These two phenomena are clearly of a finite-state nature.

Another analogous phenomenon involving adverbial contexts is found in the pairs:

*(Stupidly + Surprisingly), Bob drank his beer*

*(Stupidly + Surprisingly) enough, Bob drank his beer*

The word *enough* optionally modifies some adverbs in a constrained way. For example, the combination is forbidden with the adverbs *initially*, *actually*, etc. We also observe frozen combination such as *sure enough*, *true enough*. Representing such families of constraints by finite automata is quite natural.

In the same way, a noun phrase such as *an English speaking student* can be generalized in the following way:

- in the position of *English* one finds the name of any language,
- in the position of *student* any human noun may occur, whether nouns of individuals (e.g. *child*, *grocer*) or of groups (e.g. *Parliament*),
- the word speaking is obligatory; neighbouring words such as talking, discussing are not allowed.

Once the nouns of the lexicon of the language have been classified by features such as: **Language**<sup>2</sup>, **Human**, **HumanGroup**, the combinatorial productivity of these phrases is captured by a simple finite automaton in a most natural fashion.

We have grouped these three examples in figure 1.

---

**Figure 1**

---

### Notations:

1. *The graphs* of figure 1 represent finite state automata. Each has one initial state (left-most arrow) and one final state (right-most square). This representation has been devised in order to facilitate the effective construction of grammars by linguists<sup>3</sup>. It departs from classical representations in that states are not overtly represented: the nodes are not the states, there is no symbol for them. Arrows are labelled by the alphabet of the automaton, that is English words and/or their grammatical categories. The notation  $\langle N:Hum;s \rangle$  corresponds to any human noun (**Hum**) in the singular (**s**),  $\langle E \rangle$  is the null string. A word between angle brackets corresponds to all the members of its inflectional class:  $\langle a \rangle$  corresponds to the variant articles *a* and *an*. Inflected words and grammatical categories are defined by an electronic dictionary (EDELA) of about 50,000 entries (simple words), inflected in a set of about 100,000 words with their grammatical attributes. To avoid multiplying parallel paths with words having identical roles, labels are grouped in boxes, hence a box of  $n$  words labels  $n$  arrows in the classical representation. Shaded boxes contain subautomata that are called into the graph by their name.

2. *Structures*. We note sentence patterns in the following way:  $N_0 V N_1 Prep N_2$  represents the structure subject-verb-two complements; the  $N_i$  s are noun

---

<sup>2</sup> This category should be subdivided in **Modern Languages** and **Ancient Languages** (no longer spoken).

<sup>3</sup> M. Silberztein 1993 has written a graph editor for this purpose: FSGRAPH.

phrases. But in the graphs, phrase boundaries are not always marked. One way of formalizing phrases is by describing them in separate automata, which requires attributing a name to each automaton, hence to the corresponding phrases. This name is used in the automata where the phrase occurs. The unit of description is the sentence, and as a first step, declarative simple sentences, i.e. the sentence forms subject-verb-complement(s). We will discuss their transformational equivalence.

We will describe here a more complex example than those in figure 1 and we will use it to show how a general method of representation can be developed for precise and complex data. The corpus we have chosen is the description of the activity of a Stock Exchange, as reported daily in newspapers. The texts are short and they seem to be repetitive, using fixed phrases that recur constantly, differing only in numerical variations.

Such a point of departure is highly subjective. Firstly the choice of the domain is completely semantic and secondly, it is determined by the intuition that the set of expressions is restricted, perhaps closed. The perusal of texts, over a period of several months, has given the impression that the vocabulary, the constructions and the style of the domain are limited. Such a hypothesis needs to be verified carefully, and can only be confirmed experimentally. After all, it might be the case that the family of texts considered special contain in fact all the sentences of English. Namely, the general sentences of the language appear rarely, but by accumulating them, albeit slowly, the whole of English would be covered. We did not perform any a priori study, consisting for example in building the lexicon of a series of texts and comparing them chronologically. Instead, we decided to analyze syntactically the sentences and the phrases of the texts and to classify them in order to fit the representation. Once these local grammars are built, it is easy to use them to parse the texts and to verify their rate of success<sup>4</sup>.

---

<sup>4</sup> M. Silberztein's FSGRAPH program incorporates a generator that provides a parser for each graph. D. Maurel 1990 has written an extensive f.-s. grammar for time adverbials. E. Roche 1993 has built general parsers for full sentences and E. Laporte has used related transducers to resolve various types of ambiguity.

## 1. Linguistic modules

Typical sentences dealt with are the following:

- (1) *Advancing issues outnumbered decliners, 1,016 to 861*
- (2) *The Dow Jones Industrial Average fell 15.40 points Friday*
- (3) *The Dow Jones industrial average closed below 4,000*

It is clear that (1) on the one hand and (2)-(3) on the other describe entirely different facts. Hence, they will be described by disjoint local grammars. Such separations are crucial in the sense they allow a modular construction of the grammar of the field. In this case, the separation is obvious, but we will see that in other situations, one must introduce both semantic and syntactic criteria to obtain separations. Moreover, ergonomic limitations such as the size of computer screens also intervene as boundary conditions in the effective construction of local grammars.

---

Figure 2

---

### 1.1. Example 1

In figure 2, we give a local grammar of the sentences that are used to express the meaning of (1). The graph contains independent modules which we discuss now.

**Module 1.** The shaded right-most box is called *AdvUnchanged*. It represents an embedded local grammar that describes phrases such as:

*with 230 stocks left unchanged*

We provide this subgrammar in figure 3.

---

Figure 3

---

**Module 2.** In the right lower part, we have another subgrammar, which is clearly isolated by arrows corresponding to empty (flattened) nodes of the automaton.

For example, the forms represented are the adverb or apposition:

*Dnum to Dnum specified as: 1,016 to 861*

Both shaded boxes *Dnum* represent an automaton of numerals in two forms, digits as above or literal as in:

*one thousand and sixteen to eight hundred and sixty one*

**Module 3.** The upper part of the graph is occupied by sentences of the form *There are X*:

*There are twice as many decliners as advances*

In these sentences, the numerical information is given by a comparative expression. Note that other sentences should be added to this subgrammar, namely other comparative sentences of the type:

*There were (more + less) decliners than advances*

**Module 4.** The lower left-most part contains the core of sentences (1). They all have the syntactic form: subject-verb-complement, noted  $N_0 V N_1$  for the exact structure of (1), and  $N_0 V Prep N_1$  for the three prepositional cases given in separate boxes. Modules 1 and 2 are adverbials that can be added to these structures. They provide precise numbers, whereas numerical information was rounded in the *There are* sentences. This is a feature we will observe in other situations.

The constructions are symmetrical, in the sense that subjects and objects are identical from a morpho-syntactic point of view. Some are simple words: *decliners*, *winners*, others are compound nouns, more or less elliptical: *declining shares* vs. *share-price declines*. All are in the plural. Identical phrases are observed in module 3. Semantically, these phrases are separated into two groups: ***share-prices which gain*** and ***share-prices which lose***. This description results from a decision taken about incorporating semantics into the grammar. As a consequence, the subgrammar is composed of two independent submodules, one for each sentence group:

*Decliners topped advancers*

*Advancers topped decliners*

The same is true in module 3, except that a common part *There are ...* is factored

out to the left<sup>5</sup> .

We could have decided to limit ourselves to a syntactic description, ignoring the two semantic types. In this case, we would have considered only one type of phrase, in which would be grouped *advances* and *decliners* in the same distributional class; the content of this class would then appear both in the subject and complement positions. On the one hand, this representation is more economical since there is no longer any need for distinct submodules. On the other hand, a unique module such as this generates forms of the type:

- (4) *Decliners topped decliners*  
 (5) *Decliners topped share-price losses*

which are forbidden as nonsensical. An often-heard argument in favor of the limitation of the description to strict syntactic data consists in claiming that forms such as (4)-(5) will never occur in texts, hence will not have to be recognized by a parser. We oppose this stand for two reasons:

- forms such a (4)-(5) may indeed be found when, in the process of parsing a sentence, systematic hypotheses about the words and phrases of the sentence are made;
- the local grammars we build are neutral with respect to the parsing and synthesizing of sentences. For sentence generation, a grammar where all paths are meaningful is certainly easier to use.

However, the semantic adjustment we have just argued for is not sufficient, for there are syntactic differences. In each of the two semantic groups we have distinguished three separate types of nouns: simple nouns (e.g. *decliners*) and compounds of two types: *share price declines* and *declining shares*. This last type of compound has the pronominal form *declining ones* which is allowed in complement positions only when the subject is one of the two compound forms:

- (6) *Declining shares topped advancing ones*  
 (7) *\*(Decliners + declines) topped advancing ones*

Hence, we cannot simply add *ones* to the boxes which contain *issues*, *shares*, *stocks* in complement positions, in which case the grammar would generate (7). To adjust this subject-object dependency, we have to duplicate the four corresponding subgraphs, doubling the size of this local grammar. We must

---

<sup>5</sup> Notice that the situation is similar in module 4 where the adverbials are right factors common to both submodules.

realize that the size of the computer screen is such that a duplication of the given graph cannot 'physically' be performed: two separate graphs (with two names) will be needed. At this point, we would separate the *There are* sentences from the other type.

## 1.2. Example 2

Our second example is a grammar of the sentences that express the variations of a Stock Exchange Index, say one of the Dow Jones indexes. Examples of sentences indicating positive and negative variations are:

- (8) *The Dow Jones industrial average (gained + lost) 15.40 points at 3,398.37*
- (9) *The Dow Jones Industrial average finished with a (gain + loss)*
- (10) *The Dow Jones Industrial average broke an all-time record of 5,000 points*

The sentences present common syntactic features:

- (i) the subject is an *Index*: (cf. figures 5-6-7),
- (ii) a verb (figure 4) or a verbal phrase (figure 8) expresses the direction of the variation,
- (iii) two complements contain numerals which provide:

- a relative variation, namely the difference with the previous quotation day: 15.40 in (8). The variation is always a positive number, the sign being expressed by the verb,
- and then, the full value of the Index: 3,398.37 or 5,000.

In (9), the complement of relative variation (**RelChange**)<sup>6</sup> is obligatory, and the complement of Index value is optional. In (8), both complements are optional.

A full description of these sentences requires at least 50 graphs corresponding to different sentence types. We will discuss here the two types given in figures 4 and 8.

---

Figures 4, 5, 6, 7, and 8

---

Let us comment on some of the features of these graphs.

---

<sup>6</sup> This subgraph corresponds to forms such as *15.40 points or 0.50 %*, etc.

### 1.2.1. Subjects

The box **Index** may be filled by any index name from any stock market. In the figures 5-6-7, we give the names of the main indexes used in New York City, London and Tokyo. These graphs are typical of compound nouns, whether technical terms or proper names. Such utterances have abbreviations of various types: acronyms, omission of parts, and they also have lexical variants, either limited to parts of the term or morphemically unrelated synonyms. Finite automata represent these variations in a natural way. Note that, depending on the content of the automaton, we may want to name them grammars or lexicons. Beyond the representation of strings, incorporation into the same automaton constitutes a statement of equivalence for these strings. In many cases, semantic equivalence is the natural relation that holds between the strings and at the same time, it is the most useful relation for our descriptive program. However, there is leeway for refinements linked to the discussion above (§ 1.1, Module 4) about the amount of semantics we want to include in graphs, under the general proviso that finite-state models are appropriate.

We shall consider the various numerals involved. **Dnum** is the name of the graph that describes numerals. The numerical value of an index is given by the variable **Dnum** appearing in six shaded boxes with the same interpretation. **Dnum** has already been used in figure 2 and in figure 3 (cf. § 1.1 Module 2), but there, **Dnum** corresponds to numbers of stock names and as such, ranges between a few units and the thousands<sup>7</sup>. When **Dnum** corresponds to volumes of trading (i.e. number of shares sold), it ranges in the millions, and when **Dnum** corresponds to the Dow Jones Average, numerals oscillate around the 4.000 (in 1995), whereas the FST index and the Nikkei have different ranges. The grammar **Dnum** covers all of these numbers, the question is then whether we want to adjust the numerals to the terms they bear on.

The solution given in figure 4 consists in having a unique graph **Index**, which is a union of the various Exchange graphs. Since the numerical range of all the indexes is wide, the general grammar **Dnum** covers all cases, except that numerals in the millions are not relevant. A different solution consists in having as many graphs of the type of figure 4 as there are indexes and in using one specific grammar of numerals for each index (e.g. **DnumNikkei** for **IndexTokyo**). This dilemma has no solution within present linguistic and formal frameworks. The choice may depend on applications (e.g. for banks, for brokers) and will vary accordingly.

Another way of discussing this issue is in terms of the modularity of the

---

<sup>7</sup> This number depends on the number of issues quoted in each stock market.

subgraphs. Adjusting numerals to indexes amounts to introducing constraints between the box **Index** and the boxes **Dnum**. These constraints are superimposed on the existing paths that link these boxes, but they are independent of these paths. If we represent them directly, we change the formal method of representation (e.g. to graphs with colored edges), and this goes beyond the natural use of the finite state model. In the other solutions evoked, we use a different method for representing the constraint:

- in the solution of figure 4, all shaded boxes are independent; no fine-tuning of the numerals is performed,
- in the solution where we refine the lexicons involved (i.e. lexicons of indexes and lexicons of numerals), subgraphs such as **IndexTokyo** and **DnumNikkei** become autonomous, that is, the modularity of the various components is preserved.

Finally, let us mention another solution for this adjustment problem. We could describe the combinations index-numerals as free at the syntactic level, which is roughly what we have done in figure 4. The adjustment of the numerals would be treated in a separate semantic component. We hinted at this solution for the percentage numerals by appending a subscript:  $< 100$ . In this case, the variation range is  $0 < Dnum < 100$ . Such information could be either 'manually' introduced into the graph or in some cases constructed from the context (i.e. the paths involving the box **Dnum**). Then a separate component of the system would use this indication to restrict **Dnum** to the relevant range of variation.

### 1.2.2. Verbs.

In principle, the verbs appearing in figure 4 are polarized, indicating an upward movement of the index. This semantic feature often has a syntactic consequence, for sentences without any complement informally indicate the trend, as in:

*The Dow Jones (advanced + jumped + grew)*

as opposed to verbs indicating the opposite trend:

*The Dow Jones (slid + declined + slumped)*

In figure 4, complements are adverbials, close to locatives; in a sense they are not essential whereas in figure 8, they are similar to objects and the verbs are not polarized.

### 1.2.3. Complements

The growth of the index is made explicit in complements which provide a

numerical value of the index:

*The Dow Jones advanced to 3,425 points*

This minimal information is often enhanced by recalling the former value of the index; various forms can then be used:

*The Dow Jones advanced from 3,213 to 3,425 points*

*The Dow Jones advanced to 3,425 points up from 3,213*

Other complements or parts of these numerical complements are more stylistic than informative. For example, nouns such as *level*, *peak*, *record* or *psychologically important high* are classifiers for the numerical value of the index. They are semantically redundant, and as a consequence (Z.S. Harris 1988) can be zeroed in certain contexts.

The graph of figure 4 contains additional information:

- relative changes, including percentages of variation, which indicate indirectly the value of the index on the previous quotation day (cf. 1.2.1);
- time indication of duration: the subautomaton **AdjTime** corresponds to phrases such as *six week* (e.g. *a six week record high*), indication of date: the subautomaton (i.e. lexicon) **Day's** contains the five working days of the week (e.g. *from Tuesday's close*).

### 1.3. Verbal compounds

The graph of figure 8 named **NVNUpDown** corresponds to sentences describing both upward and downward movements of an index. The motivation for having such a graph distinct from the graph of figure 4 is both syntactic and semantic: the verbs in figure 4 all carry a meaning of directed movement. In figure 8, the same movements are expressed by combinations of verbs and complements and the verbs by themselves are not polarized. For example in:

*The Dow Jones hit a new (high + low)*

the verb *hit* does not carry any information, it is the nouns *high* and *low* that are significant. This situation is common with support verb constructions (M. Gross 1994) introduced in the nominalizations of verbs as in:

*to (have + register + show) a (decline + gain + loss + ... )*  
 = *to (decline + gain + lose + ... )*

or with support verbs and stand-alone nouns as in:

*to (hit + reach) a record*

Most of these verbal compounds take the same numerical complements as those of simple verbs. There are however a few differences:

- with simple verbs, numerical complements are all of an adverbial type;
- in support constructions the numerical complement sometimes becomes a noun complement of the supported noun:

*The Dow Jones hit a record high of 4.000 points*

The subgrammar of figure 9 bears similarities to that of figure 8 with respect to localization of meaning. The meaning of variation is even less localized in figure 9, for it is given by metaphorical and idiomatic expressions:

*The Dow Jones ended on a firm note*

*The Dow Jones gathered steam*

---

Figure 9

---

The sentences we have listed are all different, except for a few variations for a small group. Most of them can receive additional information, for example the general numerical appositions:

*The Dow Jones ended on a firm note, at 3,425 points up 1 % from 3,213*

Such examples hint at the definition of a numerical module which would appear in several graphs, avoiding duplication. But it should be noticed that two very similar modules of this kind may be needed:

- one for upward movements:

*The Dow Jones advanced to 3,425 points up from 3,213*

*\*The Dow Jones advanced to 3,213 points up from 3,425*

- one for downward movements:

*The Dow Jones fell to 3,213 points, down from 3,425*

*\*The Dow Jones fell to 3,425 points, down from 3,213*

The adjustment realized by these modules is semantic; the two values of the index have to be ordered correctly. But a lexical feature is also involved, for the adverbs *down* and *up* depend on the verbs:

*\*The Dow Jones advanced to 3,425 points, down from (3,213 + 3,645)*

\_Some verbs are not polarized (*to stay*, *to trade*) and accept both complements. They are described in a different graph.

In figure 10, we describe a non polarized range of variations expressed by an adverbial complement **BetweenPoints**, as in:

*The Dow Jones hovered between 3,213 and 3,425 points*

*The Dow Jones hovered between 3,425 and 3,213 points*

where both constructions, with reversed order of the numerals, are accepted. Hence there is no need here for an arithmetical constraint between the two values. Polarized complements are of a similar type and share many of the components of this family of phrases.

---

Figure 10

---

#### 1.4. Practical limitations

The complexity of the graphs of figures 2, 4 and 8 is maximal, from an ergonomic viewpoint:

- first, the format of the screen of the graph editor does not allow many more boxes,
- second, the complexity of the chains of elements is high, to the point where the linguist<sup>8</sup> who builds the graph becomes prone to errors.

These practical limitations can be overcome in various ways:

---

<sup>8</sup> Grammars should not be individual pieces of work. Their construction is sufficiently explicit to allow specialists other than the author to use and modify graphs.

- by using larger computer screens and appropriate software;
- paths in figure 4 are composed of straight segments. Hence the reading of a sentence from the initial state to the final state is kept left-to-right, that is, natural. The graph editor allows right-to-left reading of paths, as in figure 8. It then becomes possible to lengthen the paths, and at the same time more dependencies may be introduced, adding to the perceptual complexity of the graph.<sup>9</sup>

As a general solution to these problems, we use several techniques:

- firstly, we systematically have recourse to modularity, that is, to **semantically** defined subgraphs which are embedded into a graph, and then occupy one small box of the graph;
- secondly, we divide the sets of utterances according to **syntactic** criteria. We then construct separate subgrammars for specific syntactic forms.<sup>10</sup> In § 1.3, we discussed examples of this approach,
- thirdly, we attempt to draw graphs in a way that preserves syntactic similarities between the sentences of the graph. For example, sentences are mostly analyzed into sequences of categories such as:

*Determiner* followed by a *Noun*, (for the subject)

*Verb*

*Preposition*, *Determiner* followed by a *Noun*, (for the complements)

Nouns can be modified on their left or on their right. Since most sentences contain these basic elements, we attempt to place them in the same vertical zones. Although such zones are not materially indicated in the graph, they can be clearly observed for verbs and for some complements in most of our examples. Such a display, when feasible, introduces linguistic clarity for the dependencies among the various parts.

Not totally independent of such attempts is a more subjective notion of elegance or beauty of the graph. It is based on local and global symmetries, sometimes those of classical typography. For example, we avoid cutting boxes by arrows, and in general, we try to reduce the number of intersections of paths. In some cases, such results are achieved through the use of empty nodes whose only function is to redirect paths outside of an encumbered area of the graph. In other

---

<sup>9</sup> Of course, in the case of loops, right-to-left reading is a necessity.

<sup>10</sup> Since different syntactic forms may involve different lexical items, the separation of graphs should also be viewed as based on lexical criteria.

situations, we duplicated certain paths to avoid a web of intersecting arrows going to one area of the graph. These procedures are uneconomical in terms of number of states,<sup>11</sup> but general algorithms of determinization and minimization can be applied to these redundant graphs in order to provide compact forms for use by parsers.

Graphs possessing the qualities discussed are definitely more readable and easier to maintain.

## 2. Transformations

Many transformations affect word order. Finite automata can represent compactly sets of strings that differ by variant substrings including the null variant, but they cannot well represent pairs of strings that differ by a permutation. In other terms, the two substrings  $uv$  and  $vu$  of the strings  $AuvB$  and  $AvuB$  have to be considered as totally distinct, hence represented by two different paths with common factors  $A$  and  $B$ . This observation has consequences for the description of sentence forms.

In some cases, duplications are not costly. For example, in figure 8 we had to duplicate paths that include boxes with the nouns: *decline*, ... , *upward move*, because of variants such as:

*a 3 % decline, a decline of 3 %*

The situation is different for inserts, that is adverbials and sentential inserts such as: *at the end of the session*, *as it seemed* or *as confirmed by Federal authorities*. Given the syntactic form:

(1)  $N_0$  Aux V  $N_1$  Prep  $N_2$

that is, a typical sentence form with an auxiliary and two complements, most inserts may occur either at the beginning or at the end of the sentence, or at any of the four spaces separating the constituents:

***Without any reason***, the Dow Jones has lost 100 points at 3.000  
 The Dow Jones, ***without any reason***, has lost 100 points at 3.000  
 The Dow Jones has, ***without any reason***, lost 100 points at 3.000  
 etc.

---

<sup>11</sup> A general way of representing the ambiguity of a given string is by generating it through as many different paths as there are meanings, very much in the way constituent trees are used. Duplicating paths prevents the use of this convention.

One way of handling this situation is by making six copies of the subgrammar (e.g. **NVNUpDown**) corresponding to (1) and by introducing a box **Sinsert** in each of the six mentioned positions. Since these six subgrammars are semantically and syntactically equivalent, they could be put in the same graph. Merging these graphs does not have to be a trivial union, for common parts exist that can be factorized out, for example, as in figure 11.

---

Figure 11

---

Passive forms are another example of the problems raised by the representation of permitted forms. The following sentence is a Passive transform of one of the active sentences of figure 8:

*An all-time record of 4,000 points was reached by the Dow Jones Index*

One of the problems we have to solve is the systematic derivation of Passive forms from active ones. The transformational rule:

$$N_0 V Prep N_1 = N_1 be Vpp Prep by N_0$$

(with *Prep* possibly zero)

is not general, and its application depends on the lexical choice of *V* and on the nature of *Prep N<sub>1</sub>*. Hence, it does not seem possible to construct a Passive graph automatically from an Active one such as the graph of figure 8, which contains well-identified complements, that is a priori passivizable forms. However, the sentences with main verb *to have*, *to register* behave differently:

*\*A steep decline was had by the Dow Jones*  
*A steep decline was registered by the Dow Jones*

With the sentences of the type:

*The Dow Jones continued its fall*

the situation is more complex because of the pronoun *ITS* obligatorily coreferent to the subject. Passive forms are all unacceptable:

*\*Its fall was continued*

Note that the Middle transformation that also brings the object into the subject position is allowed, but only if the pronoun is replaced by its source:

*The fall of the Dow Jones continued*

As a consequence of the numerous irregularities observed, the only possibility is to build the graph of the Passive forms 'by hand'. This observation is true for all unary transformations.

By definition, the unary transformations are those that preserve an invariant of meaning. In this respect, a sentence and its transforms belong to the same class of equivalence (Z.S. Harris 1968). We can construct this class by taking the union of the corresponding automata, say Active, Passive and Middle. Among others, a subclass that needs to be added to this class should be mentioned. The sentences described in figure 8 and 9 are semantically simple. Alongside them there exist similar sentences of a higher complexity: the corresponding causative sentences. For example, associated with:

(2) *The Dow Jones moved up to a record 4,000 points*

we find the sentence with a causative subject:

(3) *The fall of interest rates sent the Dow Jones up to a record 4,000 points*

This sentence has a passive form, where the agent can be omitted, yielding:

(4) *The Dow Jones was sent up to a record 4,000 points*

This is a sentence equivalent to (2), that is, which will have to belong to the same class as (2).

These remarks show that the coverage of a local grammar for an initially simple notion may become considerable. But at the same time, it should be clear that many of the modules built for such a special purpose will be of use in a general grammar.

### 3. Conclusion

For obvious reasons, grammarians and theoreticians have always attempted to describe the general features of sentences. This tendency has materialized in sweeping generalizations intended to facilitate language teaching and recently to construct mathematical systems. But beyond these generalities lies an extremely

rigid set of dependencies between individual words, which is huge in size; it has been accumulated over the millenia by language users, piece by piece, in micro areas such as those we began to analyze here. We have studied elsewhere what we call the lexicon-grammar of free sentences. The lexicon-grammar of French is a description of the argument structure of about 12,000 verbs. Each verbal entry has been marked for the transformations it accepts (J.-P. Boons, A. Guillet, C. Leclère 1976; A. Guillet, C. Leclère 1992; M. Gross 1975, 1994). It has been shown that every verb had a unique syntactic paradigm. The lexicon-grammar has been extended to frozen sentences, that is, to sentences with at least one constant argument (e.g. the idiomatic form: *N take the bull by the horns* has two constant arguments). We have shown that the lexicon-grammar of frozen sentences is several times larger than the one for free sentences: so far it covers 25,000 idiomatic-like sentences and it is far from having the coverage the lexicon-grammar of free forms has. Moreover, we exclude from this count an even larger number of sentences with main verbs *être* (to be), *avoir* (to have, to get), *faire* (to do, to make).

What we have presented here is the natural generalization of lexicon-grammar. The enormity of the number of dependencies between words is itself a compelling reason to consider the sort of fixed-string free-slot theory that finite state local grammars suggest. Most of all, the notion of local grammar constitutes a generalization of the notion of equivalence classes of transformed sentences and allows the practical construction of classes of semantically equivalent utterances. We leave to another discussion the implications for theoretical linguistics of the need, and hopefully, of the validity of such a model.

## REFERENCES

Boons Jean-Paul, Guillet Alain, Leclère Christian, 1976. *La structure des phrases simples en français: Les verbes intransitifs*. Droz: Geneva.

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 1975. *Méthodes en syntaxe*, Paris: Hermann, 412 p.

Gross, Maurice 1994. *Constructing Lexicon-Grammars, Computational Approaches to the Lexicon*, B.T.S. Atkins and A. Zampolli eds., Oxford: Oxford University Press, pp. 213-263.

Guillet Alain, Leclère Christian, 1992. *La structure des phrases simples en français: constructions transitives locatives*. Droz: Geneva.

Harris, Zellig 1968. *Mathematical Structures of Language*, New York: Interscience Publishers, John Wiley and Sons, 230 p.

Harris, Zellig 1988. *Language and Information*, New York: Columbia University Press, 119 p.

Laporte, Eric 1994. Experiments in Lexical Disambiguation Using Local Grammars, *Papers in Computational Lexicography (COMPLEX)*, Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences, pp.163-172.

Maurel, Denis 1990. Adverbes de date: étude préliminaire à leur traitement automatique, *Lingvisticae Investigationes* XIV:1, Amsterdam-Philadelphia: J. Benjamins Pub. Co., pp. 31-63.

Roche, Emmanuel 1993. Une représentation par automate fini des textes et des propriétés transformationnelles des verbes, *Lingvisticae Investigationes* XVII:1, Amsterdam-Philadelphia: J. Benjamins Pub. Co., pp. 189-222.

Silberstein, Max 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson, 233 p.

Woods, W.A. 1970. Transition network grammars for natural language, *CACM*, 13(10), pp. 591-606.

## ANNEX

### *List of figures*

- Figure 1 **Three adverbials**
- Figure 2 **IssuesTop**
- Figure 3 **AdvUnchanged**
- Figure 4 **NVUp**
- Figure 5 **IndexNY**
- Figure 6 **IndexLondon**
- Figure 7 **IndexTokyo**
- Figure 8 **NVNUpDown**
- Figure 9 **NVNUpDownIdiomatic**
- Figure 10 **BetweenPoints**
- Figure 11 **SInsert**