Local grammars and their representation by finite automata

Maurice Gross

SINC.DOC 921023

LOCAL GRAMMARS AND THEIR REPRESENTATION BY FINITE AUTOMATA

Maurice Gross

University of Paris 7

Laboratoire d'Automatique Documentaire et Linguistique_

In the study of collocations and of frozen sentences (idioms, clichés, collocations, many metaphors and figurative meanings, etc.) one often encounters sets of similar forms that cannot be related by formal rules of either type: phrase structure or transformational. We present examples of such situations and we show how the formalism of finite automata can be used to represent them in a natural way.

1. Introduction

Transformational grammars are global grammars whose aim is to describe the sentences of a language at a formal level, that is, in strictly combinatorial terms. The descriptions are intended to be complete, namely to attain a coverage as extended as possible of the language. Moreover, the grammar of a language should be such that non sentences should fall outside of its range in an explicit way. To achieve this goal, Z.S. Harris and N. Chomsky have proposed combinatorial systems specialized in the following way:

- elementary sentences are described by formation rules,
- complex sentences combine sentences (starting from elementary ones) into more complex forms.

Accordingly, transformational rules are of two types:

- unary, transforming an elementary basic sentence into another elementary form,
- binary, transforming a pair of sentences into a more complex one.

Transformations preserve some semantic invariant carried by elementary sentences, implying that related sentences are similar in shape and lexical content. In some cases, a transformational relation is a synonymy relation, but it should be clear that it is a transformation that introduces the negation and that it preserves the basic meaning but not synonymy. Hence the relation of antonymy preserves the invariant of meaning. In the same way, the oddity of the sentence:

Your generosity discussed this mandoline is preserved in its passive form:

This mandoline was discussed by your generosity

In Z.S. Harris' transformational framework, transformations are equivalence relations that operate between two sentences and thus define equivalence classes.

The decision to introduce a given transformation is based on empirical observations subjected to a precisely defined methodology: intuitions about relatedness of sentences, especially intuitions of synonymy of sentences are numerous, but in order to be formalized into a transformation, such intuitions must be either confirmed by formal arguments that justify a transformational link between them or else left with the intuitive status of synonymous sentences or paraphrases, not representable by grammatical methods. This is the case for distributional relations. Consider a sentence such as:

(1) Bob worked on this problem

and other nouns in the complement position:

(2) Bob worked on this report
(3) Bob worked on this question

Sentences (1) and (3) appear as synonymous, although it is not immediately obvious that the two nouns problem and question are synonymous, whereas (2) has a different meaning. The relation (1) = (3) is not a transformation. Some synonymy relations will have to be defined independently in order to relate both sentences, and exclude the same relation with (2).

In certain situations there are formal reasons to introduce such relations. Consider the two idiomatic sentences:

(4) This meeting is a pain in the neck
(5) This meeting is a pain in the ass

and the equivalent free sentence:

(6) This meeting is a pain

they are clearly synonymous and the two locative noun phrases by which they differ do not contribute to their meaning. Moreover, the parts common to sentences (4) and (5) have identical syntactic and semantic properties, very specific properties since the combination of words is unique. In such circumstances, the argument of idiomatic invariance (J. McCawley 1976, M. Gross 1988) justifies the formal equivalence relation: (4) = (5) = (6), and these three sentences constitute an equivalence class which can be represented by the graph of figure (1).

Such a graph is read from left to right, that is, from the initial state to the final state. Each path represents a sentence.Such finite state graphs or automata are restricted here to finite paths (i.e. no loops are allowed), they are called Directed Acyclic Graphs or DAGs. They represent in a natural way local variants of a given sentence or phrase. Their computational nature makes them efficient in parsing procedure (M. Silberztein 1989).

Figure 1

We propose here to describe certain families of utterances by means of finite automata. We will see that such descriptions serve simultanously several purposes:

- they are special formation rules, they generalize the notion of elementary sentence, as such they constitute departure points for the application of transformations (unary and binary),
- they constitute a means of representation for equivalence classes built by means of transformations (E. Roche 1992).

At this point, we must make more explicit our representation of sentences in order to define clearly the new relation we have just introduced in the formation component of elementary sentences.

We discuss elementary sentences, that is, sentences of the shape subject-verb-essential complements (if any). We write N0 V W for such general shapes. Since each verb governs a specific string of complements, we will specify structures in the following way.

N0 V W =: N0 give N1 to N2

The symbols Ni (i = 0, 1, ...) stand for noun phrases.

The question of the attachment of a preposition (here to) to its noun phrase is open. Indeed, some device will have to be added to such representations, in order for example to distinguish English from French where preposition Stranding occurs much less often.

The complement noun phrases N1 and N2 are numerically indexed in order to precisely define the combinatorial effect of transformations. For example, we write:

N0 give N1 to N2 = N0 give to N2 N1

for the stylistic transformation that tends to order both complements according to their length:

Bob gave the book to one of the students
Bob gave to the student the book he went through during the semester

The transformation:

N0 give N1 to N2 = N0 give N2 N1

is different, since it changes the status of the second complement, which may undergo Passive:

Bob gave the student the book he went through during the semester
= The student was given the book Bob went through during the semester

In order to indicate this change of function, we can write:

N0 give (N2)1 (N1)2

meaning that the second complement has become first (authorizing passivization) whereas the direct object has become second complement. The stylistic length permutation cannot apply to this form, as in:

?*Bob gave the book the student who has asked so many question about its author

By specifying information in this way, we define complete classes of equivalence on a formal basis, such as:

Bob described the volcano
Bob did not describe the volcano
Bob is a describer of the volcano
Bob made a description of the volcano
The volcano has been described
The volcano is not describable
The volcano is undescribable
The volcano has a certain description by Bob
The volcano has no description
The volcano is without description, etc.

It is important to stress the use of Nominalization and Adjectivization relations. By introducing the notion of support verb (to be, to have, etc., Z.S. Harris 1964, M. Gross 1981), we account for the full derivational morphology of the verb by syntactic methods. In this way, we eliminate the so-called morphological level from the synchronic description.

2. Examples of local Grammars

Example 1

Let us now consider another example of the equivalence relation between sentences at the formation level. The following idiomatic expressions are synonymous:

Bob lost his cool
Bob lost his nerve
Bob lost his temper
Bob lost his cork
Bob lost his self-control

Bob blew a fuse
Bob blew a gasket

The direct complement is frozen, namely the determiners are frozen (e.g. the possessive adjectives are obligatorily coreferent to the subject, no modifier is allowed for the complement nouns. We observe variants for these forms, but they have the same meaning:

Bob blew his cool
Bob blew his temper (up)
Bob blew his top
Bob blew his cork
Bob blew his stack

Since these sentences share many features, we will represent their similarities, which leads us to structure this list of sentences into a local grammar for these utterances.

First we can factor out the sequence_:

N0 (=: Bob) lose Poss0

which is shared by sentences differing only by the frozen noun complement.

We can proceed in the same way with the verb to blow, however, the determiners are more varied, we have two new cases: a (fuse + gasket) and the lid.

If we compare the complements of both verbs lose and blow, we find a common set of 3 nouns with Poss0, the noun stack must be left out of this group since:

*Bob lost his stack

This set of similarities and differences is represented in the graph of figure 2.

Figure 2

But we can refine the description and extend it. First, the nouns are not all frozen to the same extent, in other words the combinations Verbs-Nouns are not all idiomatic in the same way. Thus the noun temper has a range that goes beyond the two verbs discussed, but it does not have the full autonomy of the synonymous noun self-control:

Bob lost his self-control
Bob lost the remarkable self-control he has always displayed
*Bob lost the remarkable temper he has always displayed

Other modifiers are common to these two nouns, but excluded for others:

Bob lost his proverbial self-control
Bob lost his proverbial temper
*Bob lost his proverbial cork

The nouns cork, fuse, gasket, stack, top do not appear to be used in the same idiomatic way outside of the sentences of figure 2, but this is not the case for temper, nerve and cool. We observe:

Bob (kept + controlled + held) his temper
Bob (is in + is out of) temper
Bob (got + flew) into a temper
Bob's nerve failed him
Bob kept (E + his) cool

All of these sentences contain a support verb, and we observe here common restrictions on the combinations between support verbs and their supported nouns. Moreover, we observe new frozen forms:

Bob flipped his lid

and synonymous forms with similar structures:

Bob (was in + flew into) a rage

In order to include these sentences in the grammar of figure 2, one has to defactorize some of the paths:

- we can isolate temper, in order to add the other support verbs, and authorize Modifiers,
- we have to isolate cool in order to introduce the support verb to keep.

The resulting grammar is shown in figure 3.

Figure 3

Example 2

Let us now discuss a different example: adverbial phrases that correspond to dates.

The graph DateRounded of figure 4 represents a family of date forms that are semantically similar, in the sense the years are rounded to the nearest tens. The tens are excluded. These figures are given in the two forms alphabetical and numerical.

Figure 4

The range of ten is losely divided into three periods by the terms beginning, mid, middle and late, the adjective swinging belongs to another semantic register. We have added two idiomatic phrases.

Prepositional variants are represented.

The positions of the adjectives early and late are represented by two different, hence independent paths, hiding a possible syntactic relation representable by the permutation rule:

in the (early + late) sixties
= (early + late) in the sixties

Example 3

Let us undertake a description of more precise dates_, namely forms that occur in sentences such as:

(1) The incident took place on Tuesday May 2nd, 1969

There are variants for this date form:

- the name of the day is redundant, it can be derived from the numerical date by using a calendar. Hence the date in (1) is equivalent to the abbreviated form:

(2) on May 2nd, 1969

- in a similar way we can argue that the numerical name of the day can be reconstructed from the expression:

(3) on the first Tuesday of May 1969

by using the arithmetical condition: 2 is smaller than 7, but in a first stage of description, we will not attempt to describe (3) as a formal variant of (2).

It is interesting to observe that the phrase of date is naturally described as a prepositional noun phrase, since it has the properties of this grammatical notion. But the internal structure of this phrase cannot be analyzed in traditional terms:

- Tuesday and May seem to be nouns, perhaps proper names, but of such a specific nature that no terminology has been made up for them,
- 2nd and 1969 are numerals, 2nd is an ordinal number but is 1969 a cardinal number?

Moreover these numerals qualify 'nouns', but the nature of the relation between 2nd or 1969 and the names of months is extremely specific and terms like 'adjective' or 'determiner' are not relevant to their description. Meanwhile, the pattern of dependencies between all these lexical elements is strictly defined and obeys regular rules that we are going to state:

- in the position of Tuesday any of the seven names of day may occur,
- in the position of May, any of the twelve names of month can occur,
- in the position of 2nd, we find different forms:

1st, 2nd, 3rd, 4th, 5th, ..., 21st, 22nd, ..., 30th, 31st

and we observe numerals from 1 to 31 without the mark of order. In the position of 1969, a whole range of numerals is allowed. Numerals of years are specific but depending on the calendar or on the range of time: historical times, geological times, specific modifiers may have to be appended to them (e.g. 800 AD, 400 B.-C.). We will use the symbol NumYear to represent this set of numerals.

The name of the day is entirely optional and, as mentioned above, its omission does not change the meaning of the phrase. Depending on the context of the sentences that include dates, other parts of (1)-(2)-(3) can be omitted. Hence, the following form is accepted:

The incident took place on May 2nd

implying that the year is either the year when the sentence was uttered or a year already mentioned in the context. Hence omitting NumYear is exactly like omitting a pronoun, a coreference effect is created and all the problems of attaching the truncated date to a full date are the general problems of the location of an antecedent for a pronoun.

The following forms of date are also possible:

(4) on Tuesday May the 2nd of 1969
(5) on Tuesday the 2nd of May of 1969

The distributions of names of day and month and NumYear are identical, the mark of order must be attached to the numeral of day. The name of the day and NumYear can be omitted exactly as in (3). The shape (4) can be considered as related to (3) by the omission of the determiner the. The shape (5) presents a different word order that could perhaps be related to that of (4) by a transformation. Nowtheless we will describe the sequences of type (5) independently from those of type (4).

In (4) and (5), the name of the month can be omitted, and the context should allow for the interpretation of the truncated form:

(6) The incident took place on the 2nd

When the month is omitted the day is allowed but NumYear is forbidden:

The incident took place on Tuesday the 2nd
*The incident took place on the 2nd 1969

Notice that (6) can be seen as the result of the omission of May in (4) or of of May in (5). We will choose the solution: omission in (5), for the purely formal reason that May is contiguous to 1969, making the dependency between both abbreviations easier to state, as will be seen below.

Omitting the numeral of day is not possible:

*on May 1969

but we do have the form:

(7) in May 1969

where NumYear can be omitted: in May.

Attempting to include (7) in our paradigm of dates involves substituting in for on. However, unlike the other substitutions we considered, this substitution is restricted to a particular substring of the departure string, hence on this basis only, we see no benefit in including the forms (7) in the grammar of dates we can now write. On the other hand, the situation is similar with the forms:

(8) on Monday

Since in (8), the name of the day is obligatory, whereas it was optional in all other forms, and since NumYear is forbidden whereas optional elswhere, the question arises whether to describe it as a separate form of date.

We mentioned that NumYear needed further description, this description could take this same form, and the corresponding automaton could be appended to the automaton of dates.

We could introduce further details on the date by adding the time of the day:

on Monday, May 2nd, 1969 at noon
at four o'clock
at 4 p.m
at 16 hours and 32 minutes

The utterances found to the right of at can be described exactly by the same method, the construction of the automaton will raise exactly the same problems of substitution, abbreviations and compatibilities between strings and substrings.

Again the resulting automaton can be optionally appended to the right of the final state of the automaton of dates given in figure 5.

Figure 5

3. Transformations of finite state grammars

So far we have dealt with a family of strings formally defined by two operations on a given sequence of words:

- allowing substitutions of words,
- allowing omission of words_.

The various examples of descriptions we just described are satisfactory only to a certain point. We have been mainly using a formal principle of factorization of the word sequences that are common to several utterances. We have been forced to repeat some sequences within the same graph (i.e. within the given local grammar), in other terms, the principle fails in such cases.

In one case at least, the source of the failure is clear, and we already mentioned that the permutation rule:

> (early + late) in the sixties = in the (early + late) sixties

which could save a subgraph in figure 4. In a more general way, we face a broad limitation: permutation rules cannot be handled in a natural way by finite-state grammars.

This observation holds for the grammar of figure 5: various boxes (subgraphs) have been duplicated:

- two for the names of day and of month,
- two boxes for NumYear,
- three boxes for the numerals of day.

Can we save such duplications by introducing permutation rules, that is transformational rules? The answer is clearly no in the case of the date adverbials starting with in, since no numeral of day is allowed.

If we examine the two subgraphs that include the names of day, we immediately see that the determiner a is in complementary distribution with the Modifiers of the names of day placed to their right. This type of complementation is linguistically significant and found in other contexts. Hence, some equivalence rule such as:

> a = Modifier (=: June 6, 1969)

holds, up to a permutation rule since we have:

> a Monday and Monday june 6, 1969

Thus, introducing such transformations would save the duplication of the names of day.

In the same way, a transformation that has the following effect:

on the 6th of May = on May 6

could also save duplications of subgraphs.

On the whole, from the point of view of savings that we have developed, a notion of non redundant grammar can be outlined as a system of two components:

- formation rules constituted by finite-state graphs generating elementary sentences and phrases,

- transformation rules that modify the initial graphs, introducing variants, mainly variants of word order.

A typical and general example is that of adverbial permutations. Consider the sentence:

People lost their cool

and the adverbial phrase in the sixties. The following combinations of these two utterances are accepted:

In the sixties, people lost their cool
People, in the sixties, lost their cool
?People lost, in the sixties, their cool
People lost their cool in the sixties

Such sentences are easily described in a general way by specifying their constituent structure:

N0 V N1

and by stating that the adverbial complement can occur at any constituent boundary of this structure.

Such a formulation of the rule cannot be made directly on the strings defined by our finite automata because there is no indication of constituent boundaries: transitions between states, that is word boundaries, are all of the same nature.

Various formal solutions for handling are possible this situation. For example, we can modify figure 2 by indicating the places where the adverbial strings may occur. We present such a modified grammar_ on figure 6.

Figure 6

However, the grammar of figure 6 has a standard interpretation: in a shaded box the name of an automaton is indicated which is to be inserted in the automaton of the sentences and which is of the same nature as this main automaton._ For example, an adverbial such as in the sixties appears four times in the sentences of the corresponding grammar, thusgenerating unacceptable sentences. Our solution does not allow for restrictions of the form:

(R) DateRounded can occur only once in the grammar

Hence, a special device must be added to the finite-state formalism. The device (R) is equivalent to a transformation that would move the adverbial to any of its indicated positions.

To sum up these observations and proposals, we have to modify the classical component of formation rules in the following way:

- finite state grammars are used to generate sentences as strings of words,

- a constituent analysis must be superimposed on these strings.

Then the transformational component must be generalized so as to operate on finite-state graphs with marked constituent structure:

Taking into account the fact that lexically frozen structures are more numerous than free ones, the generalization we propose should substantially modify the shape and scope of current parsers.

REFERENCES

Gross, Maurice 1981. Les bases empiriques de la notion de prédicat sémantique, Formes syntaxiques et prédicats sémantiques, A. Guillet et C. Leclère eds., Langages, No 63, Paris: Larousse, 7-52.

Gross, Maurice 1988. Methods and Tactics in the Construction of a Lexicon-Grammar. In Linguistics in the Morning Calm, Selected Papers from SICOL 1986, Seoul: Hanshin Pub. Co., 177-197.

Gross, Maurice 1990. Grammaire transformationnelle du français. 3-Syntaxe de l'adverbe, Paris: ASSTRIL, 670 p.

Harris, Zellig 1964. Elementary Transformations, Philadelphie: University of Pennsylvania, TDAP No 54. Reprinted in Papers in Structural and Transformational Linguistics, 1970, Dordrecht: Reidel.

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

McCawley, James 1979. Adverbs, Vowels and Other Wonders, Chicago: University of Chicago Press, 84-95.

Roche, Emmanuel 1992. Une représentation par automate fini des textes et des propriétés transformationnelles des verbes, Lingvisticae Investigationes, XVI:2, Amsterdam-Philadelphia: John Benjamins.

Silberztein, Max 1989. Dictionnaires électroniques et reconnaissance lexicale automatique, Thèse de doctorat, Université Paris 7: LADL.

_

_. Institut Blaise Pascal, CNRS. I am indebted to M. Salkoff for substantial improvements.

_. Poss0 represents possessive adjectives coreferent to the subject N0.

_. For detailed descriptions of dates in French, see M. Gross 1990, D. Maurel 1990.

_. We could consider omission as substitution of the null word for a given word.

_. For the sake of simplicity we did not do it for the more complete grammar of figure 3.

_. In fact, the system of program InTex (M. Silberztein 1992) compiles automatically the complete grammar (made of the main sentences and the adverbials) into a recognition procedure that locates in texts each sentence generated by the grammar.