Lexical comprehension and production in Alexia system

Thierry Selva, Fabrice Issac, Thierry Chanier, Christophe Fouqueré

Lexical comprehension and production in Alexia system

T. SELVA[1], F. ISSAC[2], T. CHANIER[1], C. FOUQUERÉ[2]

Selva T., Issac F., Chanier T., Fouqueré C. (1997) : "Lexical comprehension and production in the ALEXIA system". Conférence "Language Teaching and Language Technology", University of Gronigen, avril

1 Laboratoire d'Informatique de Besançon
Université de Franche-Comté, France
2 Laboratoire d'Informatique de Paris-Nord
Université Paris 13 - Villetaneuse, France

1. Introduction

In language learning, vocabulary is very important. Studies have shown that the dictionary is used very often in a written comprehension task. However, its utility is not always obvious. In this paper we discuss the improvements electronic dictionaries can provide compared to classical paper ones.

In lexical access, they help the learner by making the relevant information selection and research easier and then improve the efficiency of usage. Our system, Alexia, contains specific lexical information for learners.

In lexical production, computers gives us large possibilities with automatic processing. We will see how we use an analyser and a parser in order to make pedagogical new style activities.

2. Lexical access

In this part, we will see studies about the lexical access with a dictionary and a proposed model on the use of a dictionary. Then we will describe the Alexia system and our model of lexical access.

1 Studies and model

Several works (Hartmann, 1983, Bogaards, 1988) have shown that the dictionary is used very often during the translation or the reading of a text in a foreign language. Indeed, the learner is faced with the problem of the meaning of the words he or she is reading. If he cannot deduce the meaning from the context, the dictionary remains the only resort. However, its usefulness is not always obvious: according to Bogaards (1995), experiments have shown that the dictionary does not

seem likely to contribute to a better understanding of the texts. He puts forward several reasons:

- Learners do not like to use a dictionary, they consider it as a required and restrictive step which put them away from their reading.

- They are unable to use a dictionary. They have difficulty in finding the relevant piece of information. They acceptthe slightest indication which is in the line with their hypothesisby considering it as conclusive in order to shorten the "ordeal". Moreover, they have to read other entries in order to understand the first one, either because of an explicit reference or because the first definition contains unknown words. This leads to get them lost...

- The dictionary is detrimental to the reading process: tests (Benssoussan & al, 1984, Nesi & Meara, 1991) show that students using a dictionary often need more time to complete their task, without necessasrily having best results. According to Müllich (1990), the longer a learner has to search for the piece of information, the fewer chances he or she has of getting it.

As a result, Bogaards deduces that to take advantage of a dictionary requires a advanced level of knowledge of the language as well as some energy and tenacity. In order to show the complexity of the access process he has proposed this model for the use of the dictionary (figure 1):

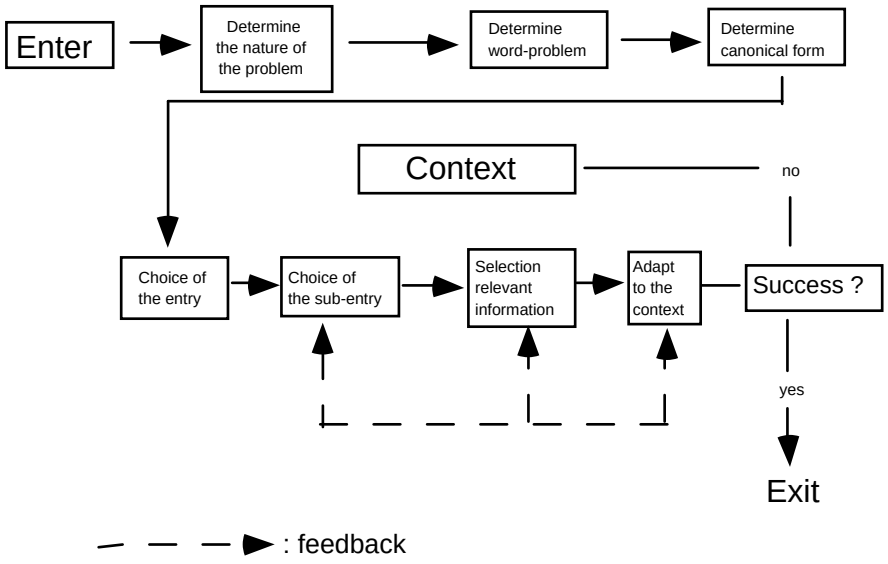

figure 1: Model of the dictionary use

Without solving all the problems, in particular the linguistic quality of the dictionaries, it seems to us that computers may be very helpful for the learner in his

or her comprehension work, by providing a greater user-friendliness and a more selective presentation.

We will describe briefly two units of Alexia, then we will see the difficulties a learner may be confronted with when referring to a classical dictionary and the ways our system tries to get around them.

2 Alexia: Corpus of texts and general dictionary

Alexia is a computer assisted lexical learning environment of French as a foreign language. It is composed of several units : a corpus of texts, a general dictionary, a personal dictionary and a lexical activities unit. It also comprises a learner's model which indicates how the learner uses the system so as to be able to follow the learner, to assess his or her learning and to give the learner advice (Chanier & al, 1995, Issac & Selva, 1996).

Alexia has a corpus of approximatively 400 texts available, all related with the work, employment and unemployment field, the vocabulary of this field is supposed to be known and mastered by every native. This corpus is only available for reading and will serve for every piece of written comprehension work.

We have also used it for the extraction of the more representative words and phrases of the field, that is, words and phrases which appear the most often in the texts.

These words and phrases, and their syntactic derivatives, synonyms, antonyms, actors, collocations and terms built from lexical functions (Mel'cuk 1992) form a glossary of about 200 entries, which constitutes the general dictionary. An entry will be called a lexical item later, that is, either a " simple " word (a group of letters enclosed by two blank spaces, ex : travail (*work*), or a collocation (ex : travail au noir (*moonlighting*)). For more details, see Issac & Selva (1996).

These entries, related to each other by semantic relations such as synonymy, hypernymy, antonymy, etc. form a lexical network. The system is able to generate automatically from the database a graph representing the network concerning a selected word.

3 Model of the lexical access with a dictionary

From these studies, we have conceived a model of lexical access. We did not test it (see part 2.4). The stages for the understanding of a lexical item from a context can be summerized in the figure 2.

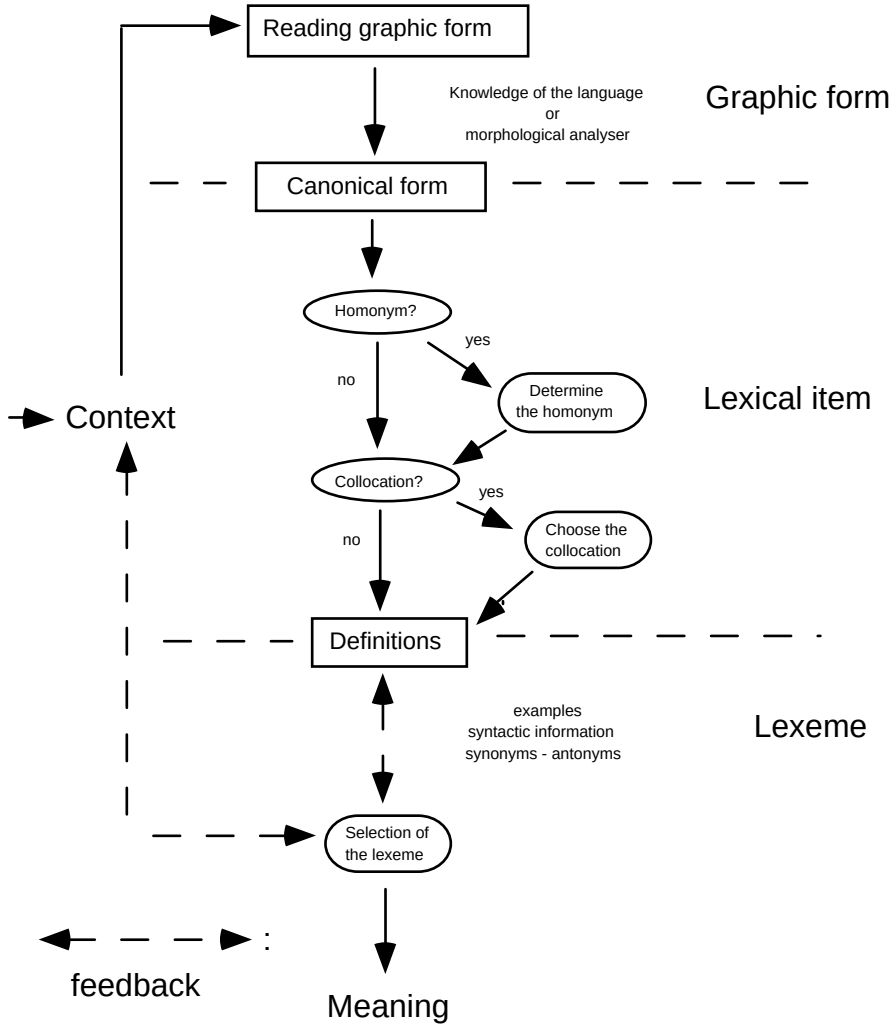

Figure 2: Model of lexical access with a dictionary

Now we will see the different stages.

.1 Stage 1: canonical form and inflection

The learner must be able to find the canonical form of a word he or she does not understand in order to take advantage of a dictionary. Like other electronic dictionaries Alexia presents the list of the lexical items in a special field. An entry can be selected either by clicking on it or by typing its first letters. This process of selection allows to find the entry very quickly, which is very much appreciated by the learners (Guillot & Kenning, 1994a & b). Because of this easy access, the learner is more likely to use more often the dictionary for extra information on other entries related with the first one.

To pass from the inflected form of the text to the canonical form is not always obvious in a morphological language like French which has many irregular forms (irai pour aller (*to go*), yeux pour œil (*eye*)). Even if certain irregular forms are cited at their place in the alphabetical order (for instance, very irregular plurals), it is not always the case (for instance, most of the time conjugated verbs are not in the dictionary) and then, the learner can only rely on his or her knowledge of the language.

In Alexia, a part of the difficulty is removed by the use of a morphological analyser which gives the canonical form(s) of an inflected one, as well as its part of speech.

.2 Stage 2: the homonymy

The homonymy is one of the most important problems the learner is confronted with during his or her search of information on an entry. Indeed, amongst several lexical item which present the same written form, precisely picking the correct one out is essential in order to have later subtle and reliable pieces of information on the meanings, the syntactic structure, the synonyms, etc. The homonymy between lexical items of different parts of speech (ex : boucher (butcher), noun and boucher (to fill up), verb) is not generally a problem. In Alexia, the difference is made by the addition, at the end of the word, of its part of speech. That is not the case for words with the same part of speech. Only a semantic criterion allows us to make the correct selection.

In a synchronic hypothesis, by describing the language for the year 1997, it is important to make clear to the learner that homonyms, formerly united by semantic links, have now become different lexical items. Dictionaries which group all the homonyms in the same entry (by taking care to differentiate them by a special notation) do not make this point clear. This grouping can be mixed up with a broader polysemy. Thereby, the work of dissociation is done by the learner. This stage does nothing but increase the time spent away from the text.

In Alexia, we point out this phenomenon to the learner by making an intermediate window appear in the case of homonymy. The lexical item is then followed by a piece of semantic information which permits the differentiation (ex : contracter : passer un accord (*to enter into an agreement*), contracter : raidir (*to stiffen*)). This window forces the learner to make up his or her mind on the choice of a item.

.3 Stage 3: the collocations

There remains a final problem to consider when selecting the lexical item. Is this item isolated or is it a part of a group of words inside which its meaning is modified? In other words, is the learner reading a collocation?

Although it is an important one (there is a higher proportion of phrases than simple words in the language), classical dictionaries do not stress this phenomenon. As collocations are not full entries and as they may have internal lexical variations, they are not listed in the alphabetical order and it is therefore difficult to locate them (for instance, is coup de barre (avoir un coup de barre, *to feel tired all of a sudden*) found in the entry for coup or the one for barre ?).

In Alexia, a collocation is a full entry with its definition, example, syntactic structure, etc. A part of the collocations has been automatically extracted from the corpus and the rest has been added from other dictionairies or by our intuition, the corpus being not large enough. They are handcoded, as well as their lexical variations. The system helps the learner to pick the collocations out by showing, when he or she selects a lexical item, all its collocations (for instance, if the learner selects travailler (*to work*), he or she will be shown travailler au noir (*to moonlight*)). He or she can then quickly decide whether the problem comes from a collocation or not.

In order not to overload, collocations are only listed with the other lexical items in their main form (the one which occurs the most often). Variations are part of the syntactic information. Nevertheless, the access in the list to a collocation is made from each of its constituants (words which compose it), as well as the constituants of its variations. Links may be easily multiplicated whereas information should not[1].

In most of electronic dictionaries (bilingual of monolingual), the access to collocations is done by searching, in the full text, occurrences in the same entry of their different constituants. The research can be improved by the use of logical connectors such as *or*, *and* or *close to*. This is a convenient process but it changes nothing as regards the statute of the collocation. It is only cited in the entry of an other word, or there is only its explanation. Nothing is given for instance about its syntactic structure, its lexical variations, its synonyms, etc.

.4 Stage 4: the definition

According to Bogaards (1995), reading and understanding of the different meanings of a lexical item may cause many difficulties. Depending on the entry, the learner may be discouraged by the quantity of information he or she has to read. Without reducing the number of meanings, the computer can show them in a gradual way. In Alexia, like in the electronic dictionary Le Robert Electronique, we use abridged definitions. These display the minimum of pieces of information necessary for selecting the appropriate meaning. Of course, if the learner needs more than the abridged definitions for understanding, he or she can then refer to “ normal ” definitions, either one by one - the leaner then ignores those he or she is not interested in - or all together, like a paper dictionary.

The last problem, and not the least important, is the understanding of the definitions.

[1] However, this happens in paper dictionaries (Bogaards, 1991).

In this paper, we will not raise the main problem which is the linguistic quality of the definitions. Do they have to be full sentence or le more like classical ones, i.e. substitutable for the lexical item in the text? Do the meanings have to be presented in a frequency order or in a logical order? These are questions of lexicography we will not answer here. We will only point out that, regardless of the writing of the definitions, an electronique dictionary has to make them easier to understand.

Even if every learner's dictionary is written using a controlled vocabulary (the 2000 or 3000 most frequent words of the language), this constraint may not be sufficient enough for understanding, either because the learner's level is not high enough, or because dictionaries do not always respect it. The learner must then read other entries in order to understand the first one, which leads to a waste of time.

One of the most appreciated advantages of electronic dictionaries is to go easily, by clicking, from a definition to another one (Guillot & Kenning, 1994b). In Alexia, this process is improved by allowing firstly, to keep the first definition the learner is trying to understand on the screen, and secondly by giving the particular meaning of the word not understood in this definition. This is done by clicking on the word. A small window appears and gives the meaning of the word in the sentence instead of displaying the full entry. The particular meanings cannot be found automatically. They are coded in the database and are a part of the definition of a lexical item. Of course, we can only reference the already existing items of the glossary. In this way, we hope to avoid the reading of extra definitions as much as possible and thus shorten the time needed for understanding.

The system presents example of use, syntactic information and synonyms of the lexeme (sub-entry). He or she gets them by clicking on the different definitions. This information and the feedback with the context help the learner to differentiate between several meanings.

4 Test

We will soon carry a method of testing with foreign learners. It will be a formative evaluation in order to observe the way the learners use the system. They will have to solve several linguistic tasks by using the lexical resources of the dictionary. Will they find the relevant piece of information? How efficient would they be?

We will not try to validate the model, for the number of learners will be restricted. But from traces and observations, we will able to know whether the interface is suitable and whether we have to install a guide. Moreover, we will be able to modify the model in order to improve it for next tests.

3. The production stage

Although dictionaries for lexical production exist (Longman, 1993), very few works have been carried out and contrary to the lexical access, there is no model of production when using a dictionary. Therefore, in this paper, we will only see the way a dictionary can help for the production of lexical item.

In this domain, computers and their capacity of automatic processing (pedagogical activities) give us large opportunities. We will now see what could be automatic lexical activities, the formalism we use and the parser.

1 Production-based lexical activities

The goal of the production stage is to allow the learner to re-use the lexical knowledge he or she achieves in the understanding stage.

This production stage is composed by series of activities articulated around a syntactic parser based on the Tree Adjoinning Grammar formalism (TAG). We will see below the motivations that have incited us to choose this formalism.

.1 Limits

Systems based on TALN have their limits. We cannot leave the learner produce everything, and this for three reasons at least:

- Formal grammars of the natural language, whatever the formalism is, have a cover very limited in the current state of research. Consequently if the productions of a learner are not constraint, the analysis will likely fail.
- In our system there is no semantics treatment (except some semantic features of the formalism). Consequently productions which had no sense ("the green ideas...") could be judged correct.
- A too great liberty do not encourage the learner to produce. On the contrary they have tendency to re-use resources offered by the system.

.2 Activities

As example, we propose therefore series of activities strongly guided by the system.

* Metarule : We use possibilities of the grammatical formalism that associates to each structures sentence some of these syntactic derivatives (nominal form, passive form,...). From words studied by the learner we will select sentences in the corpus, that we know we able to analyze, and we will ask the learner to give one of these derivatives. For example: "What is the passive form of: <<the law fix several conditions>>".

* Expression : Expressions hold a great importance in our system and the use of a parser allow us to propose complex tasks.

- Replace in a sentence a word by the correct expression. We propose to the learner a sentence where the use of an expression is more adapted. Then the learner had to rewrite the sentence with the correct expression.
- Study the frozen degree of an expression. Let an expression in a sentence, the system proposes to the learner to rewrite the sentence by moving a word on the paradigmatic axis. For example: "travailler au noir -> bosser au noir".

.3 Output of the parser

We can distinguish three representation levels for a sentence after a parsing:

- The syntagmatic tree (derived tree see figure 4) which represents the syntactic structure of a sentence, totally when the sentence is correct, partially when is not.
- The derivation tree (see figure 4), which is build in parallel with the syntagmatic tree, specifies how a derived tree was constructed.
- To each node (syntactic category) is associated unification features which characterize it. These features constrain the formation of a new derived (and derivation) tree.

According to a given sentence we can distinguish four kinds of parse results.

- The grammar is useless , i.e. the lexical units chosen by the learner cannot be combined to form a correct sentence. Then the system can identify the way the different units are incompatible.
- The production is correct. The learner can consult the derived, the derivation and the elementary trees (see below).
- The production is incorrect with a mistake detected at some unification stage. The cause could be some agreement mistake (wether morphological or semantic). A correction/explanation can be proposed in some cases.
- The production is incorrect with a syntactic mistake. The system offer to the learner the smallest sequences of trees corresponding to the sentence.

2 The TAG formalism

We have chosen for AlexiA to use the lexicalized tree adjoinning grammar formalism. This formalism, that situates in the lineage of unification grammars, is not based on rewriting rules but on elementary trees. The analysis is built on an specific operation: the adjunction, and have an important constraint: the lexicalization of the linguistic information. All elementary tree has a lexical anchor at least at its leaves.

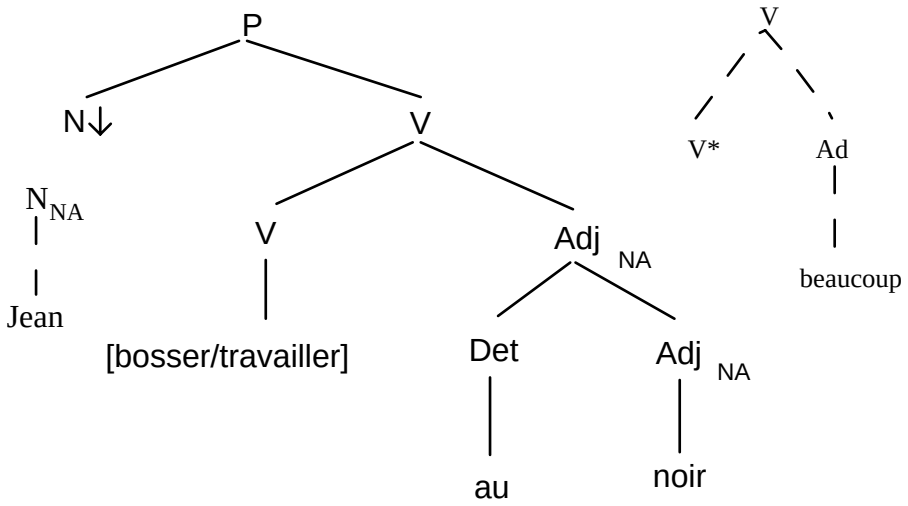

figure 3: elementary trees. One of them is the collocation travailler au noir (*to moonlight*)

This lexicalization allows us to correctly describe the linguistic process at stake for, in the grammar, each lexical item, including expressions, is defined with its syntactic context.

There are two operations defined in the TAG formalism to build a new tree (called a derived tree), adjunction and substitution.

The adjunct operation allow TAGs an extended domain of locality, while keeping a reasonable analysis complexity (TAGs are only slightly more powerful than context-free grammars). Technically speaking, the substite operation is a specialized version of adjunction, but linguistic work in TAG grammar development argue for separating these two notions.

We obtain a derived tree by application of the two operations on elementary trees or on derived trees. The derivation tree is a tree which specifies how a derived tree was constructed.

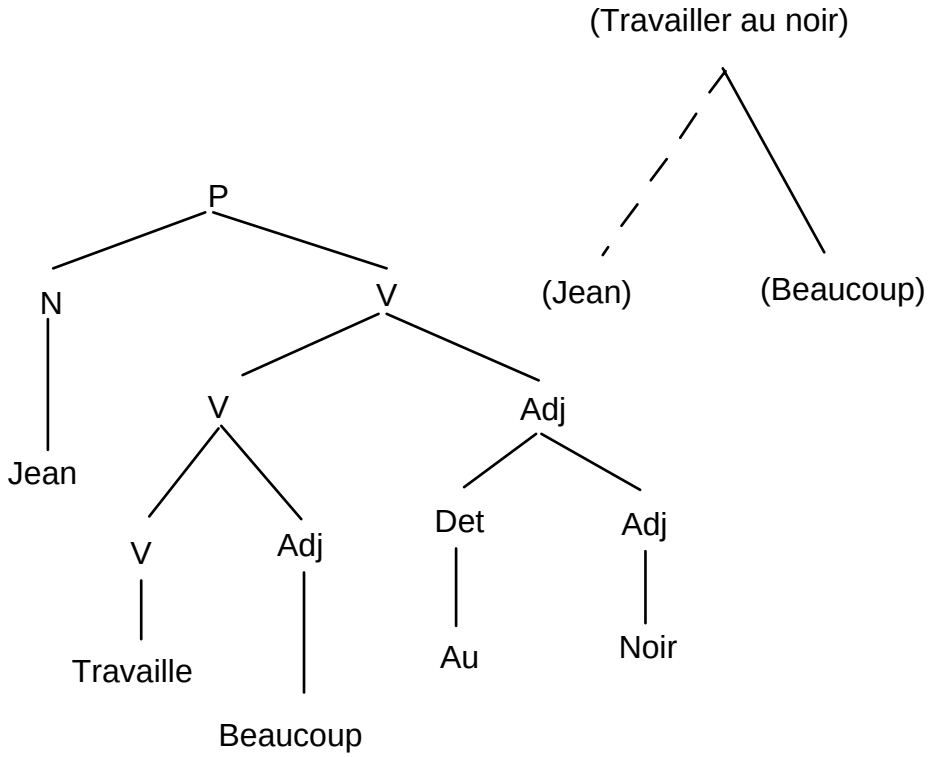

figure 4: derived and derivation tree

A unification operation takes place during these two operations in the following way. In each node of an elementary tree two feature structures are attached: a top (resp. bottom) feature structure contains information about the top (resp. bottom) of the tree rooted at this node (Vijay-Shanker, 1987).

This formalism is entirely adapted to apprenticeships objectives that we have fixed (Abeillé, 1992). Indeed, this formalism possesses, by definition, qualities that we research for our system:

- Each lexical item being defined in a sub-categorization framework, we will be able therefore to directly re-use information contained in the grammar. They will be exploited by the learner during the comprehension stage.
- Very important expressions in foreign language apprenticeship will be easily represented.
- Each element of the grammar being a tree, this particularity allows us to offer to the learner an intuitive representation (graphic) of a syntactic structure since it is not necessary to define notions such as transitive, idiomatic,...

3 The parser

The parser properly speaking consist in two stages: initialization and roundup, described in (Issac, 1994). During the initialization stage, we skim the grammar in

order to create a minimal sub-grammar. Then we determine for each tree the different positions of possible adjunction. The roundup stage consists in adjunct or substitute according to cases trees corresponding to contiguous sub-strings of the string to parse.

The parsing is essentially bottom-up in order to obtain, in case of incorrect sentence, the most partial information. For instance if the learner produces the sentence *"beaucoup Jean travaille"*[2] then the parser will return both *"beaucoup"* and *"Jean travaille"*. In case of failure, partial trees will be presented to the learner in order that he or she could rectify the construction of its sentence (see figure 5).

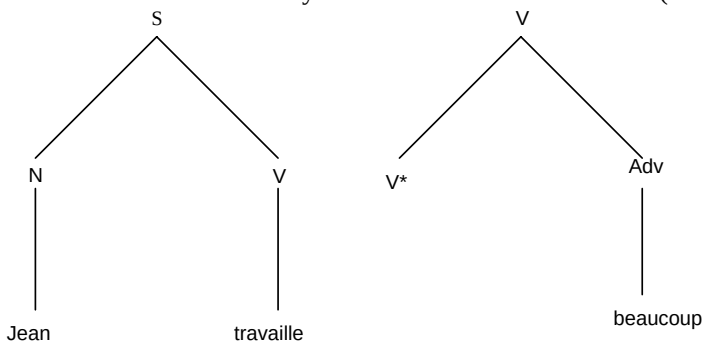

figure 5: output of the parser for the sentence "beaucoup Jean travaille"

Thus the learner has simultaneously the possibility to correct its production, and to understand inherent constraints to the expressions or words he has chosen.

The grammar we use to parse French sentences for our application consists of a morphological parser, a morpho-syntactic lexicon, and a set of tree. A tree-family is a set of elementary trees corresponding to different syntactic structures that share the same subcategorization type. For example there is a family for each kind of verb according to its subcategorization (one or two complements, introduced by some preposition or not).

We use a morphological parser we previously developed to select in the tree-family database the families corresponding to the morphological information of some word (55000 lexical entries, 199 morpho-syntactic features).

This method cannot take into account all the lexical units, especially for idioms, thus we use a specialized morpho-syntactic lexicon to select these. An idiom is defined as a pre-build tree.

4. Conclusion

From studies and models, we have proposed our own model for the lexical access. It has been conceived from the specific lexical resources of the dictionary. We will soon test the system with foreign learners in order to improve it.

As to the lexical production, we did not present a model. In this domain, the gap between paper and electronic dictionaries is the most obvious. We are building

[2] Which can be translated by "a lot John work"

pedagogical activities in which the analyser and the parser will be very useful for the production. Although we use a standard formalism (TAG), needs for learning are very specific and we had to create a full parsing system.

5. References

A. Abeillé (1992): A lexicalized tree adjoining grammar for French and its relevance to language teaching. Intelligent tutoring systems for foreign language learning, Vol F80 NATO ASI, pp 39-50.

M. Bensoussan, D. Sim, R. Weiss (1984): The effect of dictionary usage on EFL test performance compare with student and teacher attitudinal expectations, *Reading in a Foreign Language*, 2, pp 262-276.

P. Bogaards (1988): A propos de l'usage du dictionnaire de langue étrangère, *Cahiers de Lexicologie*, 52, pp 131-152.

P. Bogaards (1991): Word frequency in the search strategies of French dictionary users, *Lexicographica*, n° 7, pp 202-212.

P. Bogaards (1995): Dictionnaires et compréhension écrite, *Cahiers de Lexicologie*, 67, 1995-2, pp 37-53.

T. Chanier, C. Fouqueré, F. Issac (1995): AlexiA *: Un environnement d'aide à l'apprentissage lexical du français langue seconde*, EIAO 95, pp 79-90, Eyrolles, Paris.

M.-N. Guillot, M.-M. Kenning (1994a): Electronic Monolingual Dictionaries as Language Learning : a Case Study, *Computers Education*, Vol 23, No 1/2, pp 63-73.

M.-N. Guillot, M.-M. Kenning (1994b): Le Robert Electronique : a Reassessment of the Case for Dictionary-Based Work, *Computer Assisted Language Learning*, Vol 7, No 3, pp 209-225.

R. K. K. Hartmann (1983): The bilingual learner's dictionary and its users, *Multilingua*, 2-4, pp 195-201.

F. Issac (1994): Un algorithme d'analyse pour les grammaires d'arbres adjoints. *Colloque international sur les Grammaires d'Arbres Adjoints (TAG+3)*, Paris.

F. Issac, T. Selva (1996): *Représentation et utilisation de connaissances dans un système d'aide à l'apprentissage lexical*, Actes du 2e Colloque Jeunes Chercheurs en Sciences Cognitives, Giens, pp 192-201.

A. Joshi, L. Levy & M. Takahashi (1975): Tree adjunct grammar. *Journal of the computer and system sciences*, *10*(1), pp 136-163.

A. Joshi (1985): Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural description ?. *Natural language parsing*. pp 206-250. Cambridge University Press.

Longman (1993): Language Activator, the world's first production dictionary, Longman Group UK Limited.

I. Mel'cuk (1992): *Dictionnaire Explicatif et Combinatoire du français contemporain. Recherche lexico-sémantique III*. Les presses de l'université de Montréal.

H. Nesi, P. Meara (1991): How using dictionary affects performance in multiple-choice EFL tests, *Reading in a Foreign Language*, 8, pp 631-643.

Y. Schabes, A. Abeillé & A. Joshi (1988). Parsing strategies with lexicalized grammars: application to tree adjoining grammars. COLING'88. pp 578-583.