$\mathcal{A}_{\text{lexi}}\mathcal{A}$: a computer based environment for french foreign language lexical learning

Thierry CHANIER, Nathalie COINTE
Christophe FOUQUERÉ, Fabrice ISSAC

Abstract

The lexical acquisition, and more precisely the lexical utterance acquisition, takes an important place in foreign language learning. In this paper, we present a learning system called $\mathcal{A}_{\text{lexi}}\mathcal{A}$. It proposes an adapted lexicalized help in production and in understanding. It takes into account both lexical access strategies usually used to build lexical databases, and user strategies. The system proposes also lexical exercises based on the lexicon studied by the learner during the understanding and/or production stage to help the memorization. Note that our lexical database offers many linguistic information. The system gives to the learner a lexical mistake correction help during the production stage. It also proposes lexical utterances or collocations more adapted than what is supposed to be produced.

Keywords:
Foreign language, Lexical utterance, Production/understanding, TAG.

1 Introduction

The lexical acquisition, and more precisely the lexical utterance acquisition, takes an important part in foreign language learning. Indeed, the fact that a learner doesn't master a sufficient vocabulary level is considered as an important barrier to communication. Studies in theoretical and applied linguistic, as in psycholinguistic, show that the "word" is a complex linguistic unit carrying morphological, syntactic, semantic and also pragmatic information. Moreover, the choice of a word in the production of a sentence can carry a structure on all the sentence. Knowing a word requires at one and the same time to know the context where it's used orally or in written (of frequency use, associated collocations), its usage limitations according to variation of function or situation, and eventually to be able to combine it in the mental lexicon and to establish relations with other words in the associated net [14].

The vocabulary acquisition rate in a second language is very low. Some researchers think that, for many learners, the acquisition of 2000 words within 5 years is impossible[1]. Learning a second language is therefore difficult for many reasons: learning time very short, non intensive, low output, therefore bad use of strategies applied in first language. Consequently it's a major challenge in the field of computer assisted language learning (CALL).

Some learning systems offer to the learner *traditional* lexical resources (electronic alphabetical dictionary, concordancer and text database more or less homogeneous) together with lexical exercises and particularly activities where the learner can collect and organize his/her lexical information. From our point of view, according to what we know about the mental lexicon [3] and our definition of word acquisition, it appears important to us to organize these lexical activities around an *active* database which contains all the necessary information (morphologic, syntactic, semantic and pragmatic). *Active* means that the system can use its knowledge for:

- help in understanding, offer multiple access to the learner, show in the form of multiple net association between words,
- help in production, diagnose learner sentences so that it is not only able to correct some mistakes, but also offer paraphrases.

Finally, we also decide not to limit the lexicon to simple words, so we include collocations and semi-frozen utterances (including idiomatic utterances). These utterances are not sufficiently considered in CALL. However these linguistic units are numerically much more important than the simple lexical entries[2]. They are frequently used for native people, but are very difficult to acquire for non native learners. For didactitians in L2[3], learning justifications are multiple [13]: these utterances supply a basic material for analysis and segmentation of speech, mastering it should allow the learner not to transgress some lexical restriction, avoid him to make mistakes of register in the speech production, they can also make the oral and written production easier because the learner considers words more globally, finally they permit the learner to look for the social aspect of speech.

Some computer assisted learning systems try to remedy this state of things. The help is either in production or in understanding but not both [6, 7]. This help will

[1] In first language, the acquisition is 3000 words within a year during the school period.

[2] LADL works and other works have shown their important proportion in the language (20000 frozen verbal utterances against 8000 or 12000 free verbs; 6000 frozen adverb against 2000 free; 80000 simple noun against 300000 or 400000 compound noun) and their important occurrence probability in text. When there is potential ambiguity, its almost always the idiomatic interpretation which is the correct one, except for play on words.

[3] First foreign language.

be sometimes *imposed*: the action of the learner isn't taken into account. If the memorization and access cognitive strategies to vocabulary are undertaken, it is during the elaboration of the system database and not during its use [2].

Our system takes up another way, it proposes a lexicalized help to the learner both in production and in understanding. It proposes lexical access strategies currently used to modelize lexical database, as well as those used by the learner. The latter can build a lexical database according to its own criteria. The system offers also lexical activities for the memorization based on the lexicon studied by the learner during the understanding and/or production stage. Our lexical database offers a lot of linguistic information. Moreover, the learner can use a lexical mistake corrector during the production stage, which indicates that a word is misplaced and proposes utterances or collocations more adapted to what is supposed to be produced. Finally, we choose two different representation patterns: the net and the tree. The lexical data representation and the lexical access become then of first interest in such environments.

The first part of the paper presents briefly our system $\mathcal{A}_{\text{lexi}}\mathcal{A}$. The second, third and fourth parts concern the accessible data formalism and the relevant treatments. The last part shows the various aspects of access and lexical learning in our system and concludes with the perspective offered by such an environment.

2 $\mathcal{A}_{\text{lexi}}\mathcal{A}$: a presentation

The $\mathcal{A}_{\text{lexi}}\mathcal{A}$ system applies to advanced learners in french as second language. The aim is to help him, in a personal way, to memorize and to structure the lexical information, and more specifically as far as utterance and collocation are concerned.

Figure 1: $\mathcal{A}_{\text{lexi}}\mathcal{A}$ reduced architecture

The $\mathcal{A}_{\text{lexi}}\mathcal{A}$ system is composed of different units: lexical access, lexical activities, model of the learner, and implies two databases: a text corpus and a lexical net. We'll describe the system through typical use scenarii of the learner. It should be observed at first that all tasks of the learner are recorded in a unit called *model of the learner*. It associates the strategies used for the lexical access and the creation of the personal lexical database. These unit data are exploited in order to propose to the learner *lexical activities* for the purpose of helping the learner to memorize the new lexical element met and studied. The lexical database is described in detail

in section 4. It is implemented as a net as the learner frequently makes associations between words according to different criteria (same syntactic category, same semantic field, ...) improving the memorization of the information. This follows studies in mental lexicon organization [3, 15].

The user is proposed two scenarii. The *understanding scenario* allows the learner to consolidate his vocabulary and the *production scenario* validates it.

2.1 Typical understanding scenario

The **understanding** involves to two complementary tasks: to understand properly the vocabulary, and to memorize this vocabulary (i.e. to increase the learner lexicon). The learner works on one of the texts in the text corpus. This corpus is described in the section 3. The user can specify a text category (i.e. specify the lexical field corresponding to his main interest). The classical approach consists then in looking for all new words and utterances he has never met or all whose he already met but didn't understand completely.

The lexical access unit allows the learner to ask different sort of linguistic information on a lexical item he has noticed: definition, example of use, synonym, ... At this stage, he can also ask the system to indicate him utterances and collocations corresponding to a particular meaning. He can also have access to a graphical view of the semantic net and to the syntactic representations of words and utterances he studies (i.e. syntactic structures in which the word can be used). Moreover these data bring him information on lexical variations and possible syntactic transformations.

After this first stage where the learner asked for some lexical help, he completes his own database by using one of the method below:

- Annotation: in this case the learner *takes notes* on each word or utterance, he asked for. So is it for a translation or for some information extracted from the semantic net during previous lexical access.
- Association: personal association between words and/or utterances is a productive learning approach. It allows the effective appropriation of information. Example: creation of a group *tromperie (deceit)*, in which we find *monter un bateau (to hoax)* , *tromper (to deceive)*, *berner (to toss)*.
- Graphical representation: the learner can also *draw it's own lexical net*, nodes represent the words he studied and links translate for instance his personal association, or some part of the basic net.

The different operations done by the learner are recorded and constitute the model of the learner.

2.2 Typical production scenario

The learner can either do a summary of a text chosen in the basic corpus or produce a text on a particular topic. The text or topic selection is essential for the grammar parser (cf. section 5). This offers correction of grammatical and lexical mistakes. The system can then supply to the learner a sentence more adapted than what he produced and, in particular, indicate a more adequate utterance. In this case, the same lexical access functions can be found as in the understanding level. The production help is then both dynamic and static. The dynamic help corresponds to the various lexical databases (lexicon personal or provided by the system). The static help corresponds to the syntactic and lexical improvements proposed by the parser.

3 Corpus and Lexical data

In order to describe precisely the lexical knowledge we fixed a semantic field. We have chosen to work on the following fields: Work, Employment, Unemployment. Since the appraisment must refer to the manner in which the language is really used, we built an electronic text corpus in standard french, from which linguistic studies have been made. These studies serve to create an active lexical database.

3.1 Corpus

In order to build our lexicon on the semantic field *search for a job*, we began with an electronic corpus, which proposes today 250000 words in standard french. So we have selected article newspaper, review, magazines of all sorts: economics or social press, daily or not working-class periodical publications published by unemployed people association, etc. This corpus, composed of written text, contains nevertheless interviews.

We built this corpus for many reasons, it must: be used to build a dictionary for the field; provide real examples of these words in context, by association of lexical items and sentences; be a text database for the constitution of lexical exercises to help the lexical acquisition in L2.

3.2 The extraction of the words and utterances of the field

To exploit the corpus data, we have built an automatic extraction procedure for lexical items (simple words, collocation and terminological utterances) characteristic of the field. To do so, we chose to use statistic and probabilistic methods. For the extraction of the basic words we compute the lexicometric order extracting the list of each corpus shape associated to its frequency, in decreasing order. The first elements of the list correspond to grammatical words, but we find also lexical words, the exceptionally frequent use of which was conditioned by the corpus thema (for substantive *travail*, *embauche*, *emploi*, etc.). These lexical words, associated to some support verbs (*faire*, *mettre*, etc.) often used in utterances, form the set of basic words.

This basic word list has served as a reference to extract utterances, a component of which belonging to this list. This way to proceed eliminates some utterances, like metaphoric utterances (because it's possible that no word of such an utterance belongs to the lexical field we studied), but offer the advantage to reduce interferences (structures not characterizing the field in particular, but french). These coocurences have been built from lemmas and not from bended forms, so as to observe the syntactic variation and frozen degree of these utterances.

From this list of coocurences and the lexicometric order of the corpus, we apply a probabilistic model. We study more specifically lemmas of the basic words and we eliminate grammatical words in association. We then deduced the probability for each collocation and fixed the acceptability step, we intervened a second time to eliminate manually collocations[4] which don't present interest with regard of their use in the *recherche d'emploi* field. These two manual interventions give to reduce the corpus resulting of statistic and probability. Finally each utterance is indexed with regard to sentences in which it appears, so as to study the syntactic variations.

[4] 10 to 15 % of the collocations.

3.3 Linguistic studies

We kept 150 lexical units and described them *à-la* Mel'cuk, as in the DEC[5]. The text-meaning theory developed by I. Mel'cuk gives with no doubt the finest description of french, even if only some restricted part of french has been studied in the DEC. Moreover this description is oriented generation and paraphrase.

1. Basic Information
   - Desambiguisation of the meaning. A lexical unit corresponds to only one meaning, the different acceptations of a word or utterance are indexed. Example:
     (a) X avoir un métier
     (b) X faire Y avec un certain effort
     (c) X modifier Y par action suivie
   - Definitions. They are either computed from data in the network, or built from primitives.
     Example:
     (a) *bosser*: plus(*travailler*) $\Longleftrightarrow$ *travailler dur*
     (b) *travailler*: *faire une chose avec un certain effort*
   - Examples in context. They are extracted from the corpus and associated to the entry.
   - Language registers. We chose four language registers: trim, current, familiar, coarse.
2. Syntactic information
   We indicate syntactic category, gender and number for each lexical entry. We give also some syntactic constructions (passivation, pronominalisation, question, modifier introduction, lexical variation, ...) (cf section 4).
3. Network information
   The net consists of nodes, representing lexical entries, and links representing different associations.
   - Computation of the semantic links. Each lexical item is linked to different others. Some kind of links are defined below. We manage the lexical entry with the meaning, so utterances and words can be linked between us, as also utterances among themselves.
   - The utterances are linked by
     – their meaning: utterances with a close meaning or having semantic link between themselves,
     – the words they have in common: utterances sharing words are linked between themselves by these words.

     We divide the list of links in four groups to improve the design of the network:
     – Semantic type: synonymy and antynomy, inclusion of one meaning w.r.t. the other.
     – Derived form, actant or circumstancial forms.
     – Functions: we choose three general functions we can apply in many cases.
       Example: Oper1(*travail*) $\rightarrow$ *dénicher[un travail].*
     – Meaning component: less, more, very/intense, good, ...

[5]Dictionnaire Explicatif et Combinatoire [10, 11, 12].

4 Analysis

We have chosen the Tree Adjoining Grammar formalism (TAG, [1]) as the syntactic framework in $\mathcal{A}_{\text{lexi}}\mathcal{A}$. The grammar written in this unification-based formalism has no rewriting rules but consists only in a set of elementary tree structures (one such tree for each lexical entry). The analysis is based on two specific operations: the adjunction of a tree into another one and the substitution of a tree at an 'open' node. Each elementary tree must have at its terminal nodes at least one lexical term. Thus the linguistic processes are more easily understandable as a structure is necessarily associated to each group of words.

We specifically designed a parse process in two steps: *initialization* and *consolidation* [9]. During the initialization step, we parse the grammar so as to create a minimal sub-grammar. Then we determine for each tree the different possible adjunction positions. The consolidation step consists in adjunctions and substitutions, the trees correspond to contiguous substrings of the entry string. The analysis is mainly bottom-up in order to get, if the sentence is incorrect, the most possible partial information. If the analysis fails, the partial trees are presented to the learner and allows him to correct himself the sentence. So the learner has simultaneously the possibility to improve his production, and to understand the constraints inherent on the utterances or words he has chosen.

5 cognitive strategy and learning

5.1 evaluation

Considering the objective of $\mathcal{A}_{\text{lexi}}\mathcal{A}$, we need to evaluate:

- Strategies adopted by the learner in text understanding tasks and those which are adopted in production tasks.
- The way the learner build his personal lexicon.
- The different kinds of lexical access and lexical resource help he uses to the understanding, the production and the vocabulary carrying.

Research in psycholinguistic and in applied linguistic don't bring answers to these questions for the following reasons. The experimentations lead in psycholinguistic on the lexicon (cf for instance [16]) consist generally to discover the structure of the mental lexicon from word or segment of sentence recognition tasks, done in a very short time and in very limited contexts. The subject answers are limited to predefine choice selections. Even if we take into account results of these experimentations, there is no possibility to judge subject performance in lexical production. The context, too much limited, does not take into account the linguistic combinatorics met in the text and ignore the high level choice done by subjects. Finally the stimuli answer speed gives no information on strategies used when there is a longer time to choose, in other words when the subject has to integrate a set of information issuing from the textual context and to do a selection on the information.

The experiments, numerically limited, lead by researchers in applied linguistic used only paper dictionary [4]. Results of these experiments are not easily transposable because the task given to the subjects are those generally encountered in second language communication situation. Moreover, these experiments do not take into account computer based environments, and all specific strategies induced by those environments on learning.

So we prepare an experimentation to evaluate the underlying hypothesis in $\mathcal{A}_{\text{lexi}}\mathcal{A}$ and that could modify the environment. Note that, for these computer

based environments, this kind of evaluation is often done after the implementation. It is generally used to justify the work and not to improve it.

As $\mathcal{A}_{\text{lexi}}\mathcal{A}$ is based on a lexical net, it's necessary to do the experimentation on this kind of support. We already have two electronic dictionary which allows to do a lexical research based on semantic link between lexical items: one in english, WORDNET, and the other in french, DICOLOGIQUE [5]. As far as we know there was no experiment done on these two environments. Our system is oriented to french lexical learning, justifying then the DICOLOGIQUE choice as experimentation frame. The experimentation allows us to observe how subjects take advantage of this type of net and answer the following questions:

- How learners access a lexical item ? (ACCESS)
- Which strategies did he use to understand the meaning of a new word ? (UNDERSTANDING)
- Which strategies in face of a new word ? (PRODUCTION)
- Which methods are used by learners to remember vocabulary ? Does a computer based environment using lexical net promote this remembering ? (LEARNING)
- Does a lexical net allow a better approach of production and understanding tasks, what are the advantages or the disadvantages of such a representation: does it suit to learners strategies and does it correspond better to their mental representations ? (ELABORATION OF THE LEXICAL NET)
- Which information are more frequently used in a lexical research (synonymy, antonymy, syntactic structure, example, definition, ...) ?

We do the experimentation with learners in french of different levels. Subjects are then in a lexical acquisition phasis, but they have the same semantic field knowledge than french people. We take subject with different levels, this allows us to compare the different strategies, the aim is to introduce in $\mathcal{A}_{\text{lexi}}\mathcal{A}$ a reflexive dialogue level.

5.2 Personal lexicon and lexical activities

The first pedagogic principle defended by Goodfellow [8] is to help the learner to build his own dictionary. Three interdependent criteria justify a lexical item to belong to his personal dictionary: an utterance would have been taken up in understanding, production or introduced in regrouping several utterances. As a lexical net promotes the acquisition, Goodfellow [7] shows that a computer environment allows the learner to do his own associations. The question stay open to know if the link type between items has to be chosen by the user or, on the contrary, among those used by the lexical database. The status of a lexical item can consequently be different according to its introduction mode, and if it had been validated by appropriate lexical activities.

Mac Whinney suggests, starting on his works on the competition model, some pedagogic principles according to the lexical acquisition step of a learner (we omit the phonological restructuration and the initial transfer step):

- the learning, by heart, of lexical items is important at the beginning, but must wear away,
- during the acquisition of a new lexical item, its syntactic and casual structure has to be clarified and there must be links with other lexical items. At this step L1 transfer mistake must be clearly corrected,

- even simple transfer procedure cause mistakes, it must be better to ignore such mistake in production. The understanding, essential at this step, must be worked up from rich and difficult materials,
- in a more advanced step, it is necessary to correct mistakes so as to avoid *fossilization* phenomena and to help the functional restructuration of the learner knowledge.

From this general description, according to the level of the learner, it is possible to build a succession of traditional lexical activities in understanding and in production on words or utterances: reconstitution of a part of the net, word grouping following different criteria, exercises with blank, paraphrase of an utterance, choice between paraphrase in context, determination of the meaning of a word or an utterance in a text where it is introduced in a redundant way, retrieval of the meaning of a word or an utterance from partial information (the tip on the tongue concept), etc. These activities need for a large part the use of the text corpus, the lexical database and the TAG parser, when it will be necessary to diagnose the productions of the learner. We add to this linguistic work the possibility to dialogue with the learner about his current task, e.g. the strategies he uses, and first of all about his lexical access strategies.

The construction of such a personal dictionary and the introduction of learning activities correspond to the recommandations given by psycholinguistic and applied linguistic researchers, note that it cannot be realized without important linguistic resource settings, organized with respect to the learner.

References

[1] A. Abeillé.
*Quand l'arbre ne cache pas la forêt, analyse du français à l'aide d'une grammaire d'arbres adjoints.*In TA informations, Vol 31, n2, pp. 51-70, 1990.

[2] E. Agirre, X. Arregi, X. Artola, A. Diaz de Ilarraza, F. Evrard, K. Sarasola.
Intelligent dictionnary help system. Proceedings of *EURALEX'90.*

[3] J. Aitchison.
Words in mind. Oxford: Blackwell, 1987.

[4] P. Bogaards.
A propos de l'usage du dictionnaire de langue étrangère, in **Cahiers de lexicologie**, n^{o}52, vol. 1, pp 131-152,1991.

[5] D. Dutoit. *A set theoretic approach to lexical semantics*, Actes de *COLLING92*, Nantes, 1992.

[6] N. M. Fontana, S. M. Caldeira, M. Cristina, F. De Oliveira, O. N. Oliveira Jnr.
Computer assisted writing : Application to english as a foreign language. In **Computer assisted language learning**, Vol 6, Part 2, 1993.

[7] R. Goodfellow.
Call for vocabulary, requirements, theory and design. In **Computer assisted language learning**, Vol 6, Part 2, PP 99-122, 1993.

[8] R. Goodfellow.
Design principles for computer-aided vocabulary learning, in **Computer assisted language learning**, Vol 23,1/2, pp53-62, 1994.

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

[10] I. Mel'cuk.
Dictionnaire Explicatif et Combinatoire du français contemporain. Recherche lexico-sémantique I. Les presses de l'université de Montréal, 1984.

[11] I. Mel'cuk.
Dictionnaire Explicatif et Combinatoire du français contemporain. Recherche lexico-sémantique II. Les presses de l'université de Montréal, 1988.

[12] I. Mel'cuk.
Dictionnaire Explicatif et Combinatoire du français contemporain. Recherche lexico-sémantique III. Les presses de l'université de Montréal, 1992.

[13] J. Nattinger.
Some current trends in vocabulary teaching. In **Vocabulary and language teaching**. Carter R., McCarthy M. (Ed.) Longman, 1988.

[14] J. C. Richards.
Lexical knowledge and the teaching of vocabulary. In **The context of language teaching**, Richards J. C. (Ed.), Cambridge University Press, 1985.

[15] D. Singleton, D. Little.
Le lexique mental de l'apprenant d'une langue étrangère : quelques aperçus apportés par le TCD Modern Language Research Project. In Acts **Acquisition d'une langue étrangère : perspectives et recherches**, Grenoble, pp 395-402, 1991.

[16] I. Taylor.
Psycholinguistics : Learning and using language, Prentice Hall, 1990.