# ITOLDU : from technical vocabulary learning in a technical English resource pooling environment to contributing to the Papillon lexical data base

**Valérie BELLYNCK**
équipe STG, LGP2
461, rue de la Papeterie - BP65
38402 Saint-Martin d'Hères - France
Valerie.Bellynck@efpg.inpg.fr

**John KENWRIGHT**
Cellule TICE, bureau 2.12,
701, rue de la piscine, BP81
38402 Saint-Martin d'Hères - France
John.Kenwright@inpg.fr

**Abstract**

The first stage of this ITOLDU project aims to facilitate technical English teaching, especially for vocabulary acquisition. We are pursuing two immediate goals: maximizing positive student contributions, even outside of the classroom, and minimizing teacher intervention. With ITOLDU, we will be able to combine different kinds of usable access to find French technical specific expressions equivalent to English ones. The teaching context allows for a reproduction of the same situation with different classrooms and will enable the collection of statistics extracted from logs and surveys. Building versions of ITOLDU specialized to various types of language students could be the best way to produce various types of lexical contributions for the Papillon multilingual lexical data base.

## Introduction

Human manipulation of digital dictionaries firstly helps users to use new ways of accessing words, and secondly to take their actions into account as "unconscious contributions".

The use of software to access data is completely constrained: each way needs to be anticipated by software developers. A similar situation appears with "yellow pages" web services for phone number requests: when someone does not perfectly remember the spelling of a name, but can remember something about the street (part of its name, or district area…), s/he may never find it. With a book, any user can gain access to all data at the same time, and may find the desired phone number in the paper yellow pages (book). Hence, s/he can construct her/is own means of access, based on displayed order. With a web service, s/he can access only a small amount of data at the same time. A simple user may imagine other ways of access, but cannot extend the web service to try them out.

A digital dictionary has to be accessed via computing applications, but is also for human use. For software, the information relating to each entry must be as complete as possible, but for humans, if too much data is displayed at the same time, it becomes unreadable. Hence, users should be allowed to customize not only the access to dictionary content, but also the display of retrieved lexical data. We will present the first version of a technical English vocabulary learning environment, already in use by students. In the future, we plan to use it for experiments on accessing, displaying and teacher customizing, starting each year with the same initial conditions.

In the first section, we will describe the user context and how to efficiently collect measurements. Then we will present the first experimental web application, ITOLDU (Industrial Technical On Line Dictionary for University, http://www.pagesperso.laposte.net/kenwright/ITOLDU for the extranet version). In the last section, we will explain how a user can transform an access form into another to personalize the access methods, and why ITOLDU should be generalized to elicit other types of lexical contributions to Papillon from other types of students.

## 1 Teaching and learning context

### 1.1 English learning at EFPG

The context of English learning allows us to use the same experimental contexts for variants of experiments. Basic vocabulary needs are covered as well as specific technical ones shared by some communities.

The teaching learning context leads us to divide vocabulary in domains of use (business, basic, or technical English for different specialities), left to the teacher's choice.

Asking students to look for the French translation(s) of an English technical term may reveal the need for a different strategy for that used in the case of basic English, particularly in our case, where French students don't yet know the technical terms in their own language well enough.

The actual version could be used with other languages, but our learning context concerns only the direction from French to English.

To investigate the modalities of access, we need voluntary and motivated users. In a learning context, the teacher can simply motivate the user of a digital dictionary to contribute by taking into ac-

count the quality of the contribution with specific bonuses awarded in return for the student's evaluation of a given translation. But the teacher often can't spend a lot of time checking up on each contribution of each student: the work is in addition to normal working hours. Our solution will be to let the system take up this function.

### 1.2 Size and types of classes

At the EFPG engineer school, we train each year about 200 students in 10 groups, 3 years of study for each class. We have to manage different initial English levels, some students having learned English as a second foreign language (LV2). Next year, the ITOLDU web site will be used via the EFPG intranet.

The technical specific fields cover pulp and paper science, fiber chemistry, packaging, rheology, digital printing, and colour management.

As a primary use, for preliminary experiments to evaluate usability, the ITOLDU web site has been accessed by a class of 6 "sandwich course" students doing a technical degree.

The experiment has taken place between 15th May and 30th June 2004, and was divided into :

- 2 two-hour lessons,
- three weeks later, one 3 two-hour lessons,
- and finally, three weeks later, 2 two-hour lessons followed by a final exam.

The interest in testing ITOLDU in this way lies in the imposed spacing between the lessons and the opportunity for students with varying levels of English to contribute to vocabulary acquisition and share findings with their "community".

### 1.3 About the vocabulary to be learned

- Learning technical English is heavily sought after by French institutions.
- The most important direction is English – French: the tool should help remembering English terms to express accurate technical concepts.
- The students don't know the technical terms.
- There are probably 10000-20000 terms.
- The basic part is to be learned by all students and represents about 10% (1000-2000 items).
- Each student should choose and learn a small fraction of the remaining 90%.

## 2 ITOLDU

### 2.1 The first version

Recall that, in this first stage, we first want to maximize student positive contributions, even out of courses, and minimize teacher intervention. The idea is simple: through the English courses and between two courses, each student has to collect or create the lexical data for her/his own digital dictionary based on findings to do with texts or other sources the teacher has given them. The student can also add other words or findings s/he comes across in their own pursuit of language acquisition. S/he can choose from existing propositions that s/he finds and correct or create her/his own proposition. Selecting an existing proposition generates a vote for the student who has created it.

### 2.2 Teacher side of ITOLDU

ITOLDU offers teachers the opportunity of supervising student groups, encouraging involvement thanks to bonus marks, and livening up vocabulary via playful word hunts.

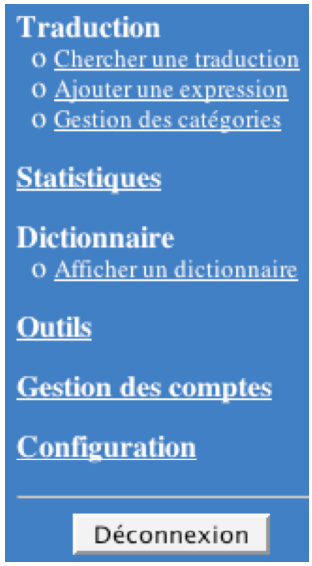


*Figure 1: teachers' summary*

Figure 1 shows the summary of a teachers' session. One can customize general web service properties (title of the site, language), broadcast learning things to do, contributing to the digital dictionary's construction (search a translation, add a new expression and create new technical domains – "categories"), manage student groups ("Gestion des comptes"), and look at each student or classroom contribution shown in Figure 4 ("Statistiques", "Afficher un dictionnaire"). Teachers never have to look inside the source of a html page (or worse in code!).

### 2.3 Student side of ITOLDU

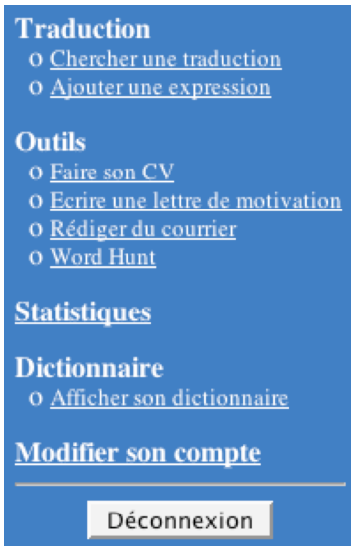


*Figure 2: students' summary*

ITOLDU allows students to gather words or expressions, and to contribute consciously with a proposition of translation or unconsciously with a selection of someone else's translation. When a student connects to her/his own digital dictionary, s/he finds a summary (Figure 2) to access the digital dictionary (search translation and add a new expression), use the teachers' prepared "to-do" tools ("Outils": CV, application letter, word-

hunt …), look at her/his ows statistics, print the current digital dictionary (Figure 5).

### 2.4 Scenario

Let us imagine that a teacher will prepare his course for a classroom and create groups and log-ins. S/he will then give the students some technical English text to study, which includes unknown technical words and expressions. Students will be shown how the ITOLDU tool works, how contributions affect part of their final grade and the concept of sharing knowledge and mutual aid. The teacher can also include an initial "word hunt" (list of targeted vocabulary) to set the ball rolling and encourage users to regularly check the site so as not to be the last to find a word.

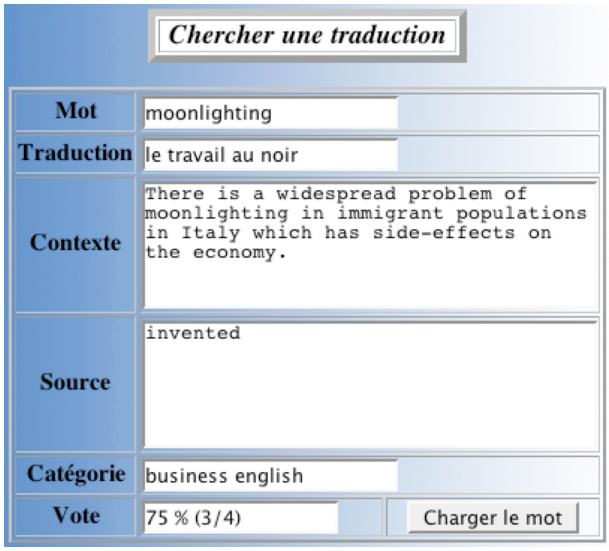


*Figure 3: basis search access form*

When reading a text, a student can be confronted with an unknown word, s/he uses the ITOLDU search tool (Figure 3). In this first version of the application, the access form is minimal: one can only enter an expression or the first letters of an expression in the first input field. But this form has been designed to be easily replaced or combined with richer ones later.

If there is no entry for the word or expression, the student can enter a translation proposal, with an example of use, the context where s/he has found it, and its bibliographical reference. Each voluntary contribution is cumulated for the statistics and the grades of each student.

If there are one or more entries for the targeted word or expression, the student can select the one which seems to be the best and add it to her/his own dictionary. This action results in an involuntary or unconscious contribution: a vote for the student who suggested this translation (the author). Each vote is cumulated in the statistics of the author (Figure 4).

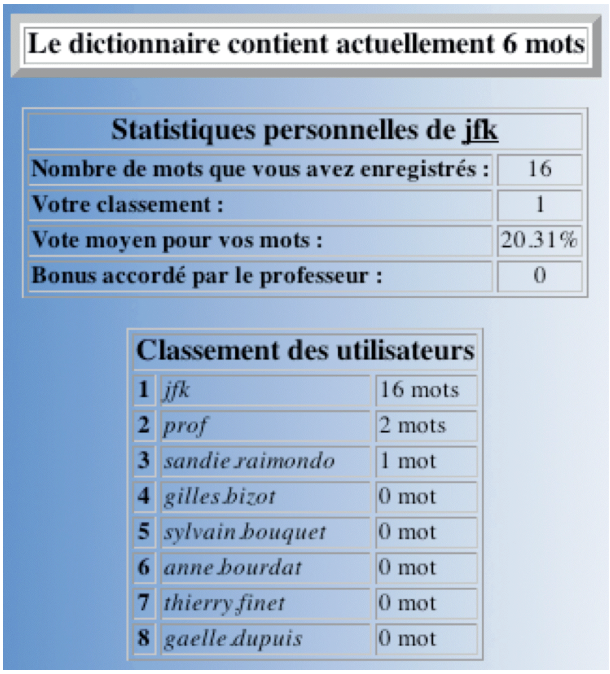


*Figure 4: resource pooling statistics*

This method of using selections as implicit votes, and further as "unconscious contributions", is the kernel of the system. As a matter of fact, it will replace teacher mediation. Students can't enter wrong definitions on purpose, because they would be incorporated in their own dictionaries (Figure 5), and teachers can trace contributions.

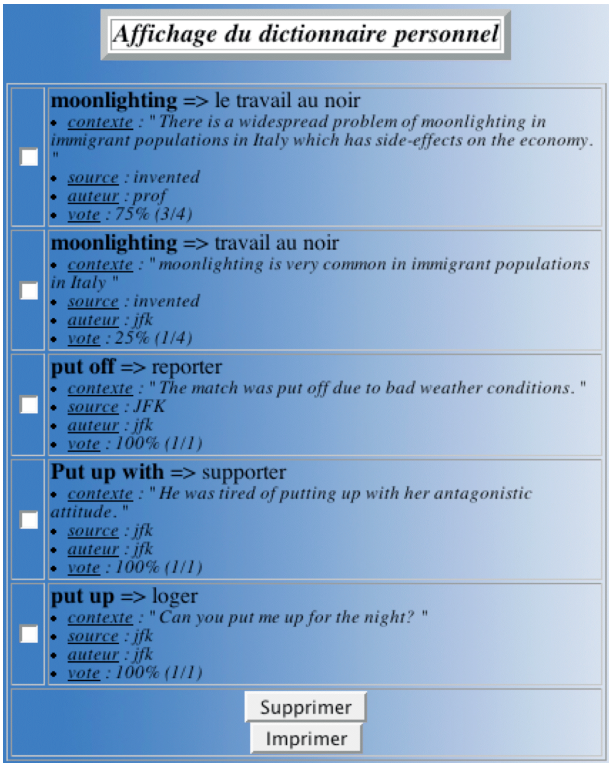


*Figure 5: taking over dictionary*

For word hunts, the student who finds the word first "win the game" and has her/his score published on a score board – just like in a computer game.

Initial experiments were beginning at the same time as the first version of this paper was written so findings could not be included for the moment.

## 3 Access personalization

### 3.1 Basic idea

To access words via a dictionary, people can start from synonyms they have in their head, look up their definitions, choose the one which seems the nearest, and then move again to words used in that definition. But one can also begin to read the dictionary from any page, trying to find some related idea (“linear” access).

To access words through a discussion with someone else, one can begin by expressing an idea, and then stop if that person can’t find the word, ask people around to help find an expression or a word that could take the place of the sought after expression, and continue.

To access a digital dictionary, one is usually limited to entering a lemma (or wordform if there is a “lemmatize” option), and to filtering via a small number of constraints (part of speech/clause, domain, variety such a GB/US). The usual methods are already closed to the book access, but without its “linear” extension, which, would anyway be limited by the screen-window. To extend the access to more “human” ways, there are two problems. Firstly, how to express the request (how to specify the word looked for)? Secondly, how to solve this problem and transcribe the request in the digital access?

A proposal for a few modalities of access has been presented in a paper on "Sensillons for the Papillon project" (Bellynck, 2002).

### 3.2 Proposing access through a Sensillon

A "Sensillon" is a web service development project to interface between "common" human users (as would have been our grandmothers) and Papillon, a multilingual collaborative lexical database. The basic idea is that accessing a word one has "on the tip of the tongue" (denoted by "xxx" in the following) is similar to entering a sense.

In that project, we proposed users to combine a few access methods by dragging and dropping their icons (left column in Figure 6) to placeholders (here, the two blank and empty horizontal thin fields), each method implementing a different way to enter the sense of the desired word. If the user wants to use more than two methods, another such field appears.

For example, suppose a user wants to find "a word to say a name for a computing interface object able to collect other objects". This first describing sentence pertains to the "definition" kind. To produce an access tool for this method, we need an English analyser able to transform a sentence into a formal or semiformal definition, and match it against a terminological database.

For the same word or expression, suppose that the user begins a sentence like "someone can combine tools by placing them in a …", and then can't go on because s/he can't find the next word (xxx). This access method could produce a request to find an example in a large sentence data base, using some kind of fuzzy matching. The user would then get some "near" sentences, and try to find the desired word in one of them. Here, the matching functions might use some semantic indexing of content words (like Wordnet synsets).

Another possibility is thinking about words like "receiver", "to collect", "form", "pattern", "input field" as words which senses are close, and "to pass", "receptionnist" as words which senses are far apart. Here, we would probably use a method based on conceptual vectors (Lafourcade 2002).

The results of each of these four searching modalities should be displayed in a personalizable view. For example, the paraphrase access results will be definitions, the example access results will be examples of usage, the conceptual access results will be a graph, or a 2D space with 2 opposite clusters, etc., and the user should be allowed to ask for highlighting words corresponding to the words in the query, or, on the contrary, words which are possible answers to the query.

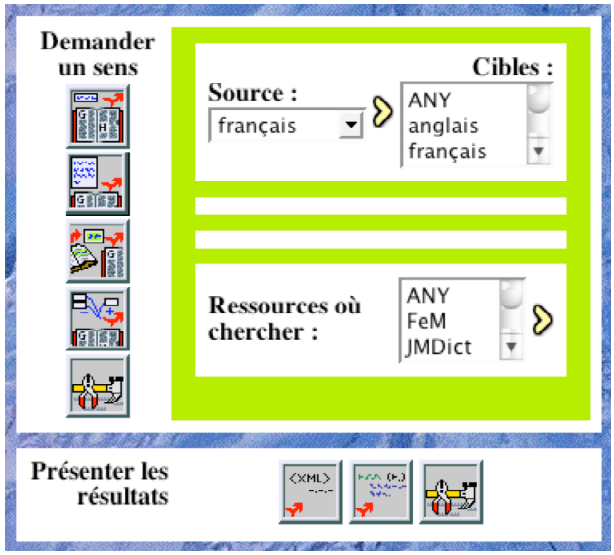


*Figure 6: construction of a sensillon*

In the "sensillon constructor", each way of displaying a result is presented alongside the corresponding input modality. This allows the user to select a part of the result to search again.

Here are the basic access methods proposed:

Classic access: with this tool, users can input source terms with their domains of use, and linguistic properties (part of

speech, level of language, reflexivity, transitivity…)


Paraphrase access: users can input a sentence that tries to define or describe the targeted expression. This tool needs to analyze the sentence.


Example access: users can paste an example of usage of the expression, maybe without the expression, that is, displaying a "hole" in its place (a tool for … small trees in a garden —> plant, water…).


Conceptual tool: users can input approximate/realted terms (words or synonymous of words that may be found near the targeted expression, in the same context, or by opposition).


More suitable tools: this button should permit a user to enter more imaginative tools. (A more detailed specification is still needed here.)

### 3.3 From Sensillon to ITOLDU to Papillon

With ITOLDU, we will experiment step by step the access modalities made available as sensillons. Real use in our teaching context will help us to collect feedback.

We plan also to build a direct "bridge" to Papillon, both ways: to access it as any lexical resource available on the web, when students (or teachers) look for an equivalent, and to contribute to it new bilingual terminology of guaranteed quality.

There is a very difficult problem encountered by Papillon and similar projects aiming at building data by (numerous) contributions of (numerous) volunteers: if the "contribution web site" does not offer any kind of service from which the contributions can be extracted, there are very few contributors and very few contributions.

In the case of OKI site http://www.yakushite.net, the service is an access to OKI machine translation system, and users are professional translators who contribute bilingual equivalents because they are then integrated in the MT system and become useful for their work.

Our idea, then, is that we should try to build a lexical data base like Papillon with *unconscious contributions from language students,* which are actually in a constrainned environment, rather than with volunteer contributions.

According to the type of students at hand, we can then "elicit" different types of contributions. English students in engineer school can contribute technical terms, as in our case. Students of lexicography and lexicology could contribute information about collocations, especially about values of lexico-semantic functions, in their mother tongue and possibly in their major foreign language, as part of their assignments. After all, that is exactly how Mel'čuk and Polguère have built the DEC and the DiCo!

But, to return to our sensillons, that supposes that the user interface be as simple as possible, limited to the task at hand (to the particular student assignment at hand), and as "playful" and iconic as possible.

### Conclusion and perspectives

ITOLDU is implemented in a first version actually in use, with only one access form. Next, it will enable to elaborate experiments of various facilitated means of access, in the case of users who want to learn technical English terms.

We plan to report on our first experimentations with our students in a few months. The second version of ITOLDU will should contain several sensillons, and some kind of communication with the Papillon database, to contribute paper-industry related terms in English and French, and with various kinds of web-accessible corpora related to this domain.

Seeing the ease with which we get lexical contributions in our ITOLDU framework, specialized to engineer students learning technical English, compared with the near impossibility to get contributions to Papillon through its general purpose, not task-oriented interface, we propose to generalize our method and build versions of ITOLDU specialized to various contexts of language learning, to elicit various kinds of information about various kinds of words or terms in a variety of languages, and pass them to Papillon.

### Acknowledgements

Our thanks go to Cédric Sintes and Sébastien Duvillard-Charvaix for their contribution to the first developments of the ITOLDU software in the framework of a student project, and Christian Boitet for his "volunteer contribution" to this paper.

### References

V. Bellynck. 2002. *Bases lexicales multilingues et objets pédagogiques interactifs : Sensillon pour Papillon*. In "Proceeding of Papillon 2002 Seminar", NII, 13 p., Tokyo.

M. Lafourcade, 2002. *Lens effects in autonomous terminology learning with conceptual vectors*. In prococeeding of COLING-02, 7 p., Taipeh.