Bilingual Lexical Data Contributed by Language Teachers via a Web Service: Quality vs. Quantity

Valérie Bellynck, Christian Boitet, and John Kenwright

***Abstract*—IToldU is a light web service which, in its first year of use for teaching technical English in French engineering schools, has enabled the contribution of just over 17000 English terms in about twenty technical domains. These terms are associated with their French translations (95% of which are correct) and examples of use (about 85% correct). In the second year, emphasis has been on quality rather than on quality: about 6000 high-quality entries have been contributed by the same number of students and classes. Some desirable extensions are in progress, e.g. to add English when this language is not included in the original language pair, and to synchronize with off-line contributions prepared on a PDA or a hand-held calculator.**

***Index Terms*—Collaborative dictionary construction, examples of use, technical English teaching**.

I. Introduction

THE collaborative construction of free lexical resources has been hampered by the difficulty of obtaining many individual small and voluntary contributions. IToldU (Interactive Technical On-Line Dictionary for Universities) is a light web service which can be used for the collaborative construction of a bilingual lexicon by a small community (typically, a group of students) while learning a foreign language in technical or specific domains. Contributions are freely offered, but are also constrained in that part of the students' English grades are computed by IToldU itself.

For the first two authors, the initial objective in building this site was to collect the produced lexica in order to populate the multi-usage multilingual lexical database (MLDB) Papillon (see http://www.papillon-dictionary.org/). For the third author, an English teacher of ICTE (Information and Communication Techniques for Education) at INPG (Institut Polytechnique de Grenoble), the objective was to improve the teaching of technical English vocabulary to French engineering students.

Manuscript received November 25, 2008. Manuscript accepted for publication August 15, 2009.

Valérie Bellynck is with équipe STG, LGP2 461 rue de la Papeterie, BP 65, 38402 Saint-Martin-d'Hères, France (e-mail: Valerie.Bellynck@efpg.inpg.fr).

Christian Boitet is with équipe GETA, laboratoire CLIPS, 385 rue de la Bibliothèque, BP 53, 38041 Grenoble Cedex 9, France (e-mail: Christian.Boitet@imag.fr).

John Kenwright is with Cellule TICE, bureau 2.12, 701 rue de la Piscine, BP 81, 38402 Saint-Martin-d'Hères, France (e-mail: John.Kenwright@inpg.fr).

In its current state, IToldU addresses mainly the instructional objective rather than the lexicographical one. Moreover, its use has led to a third interesting possibility, that of teaching the structure of simple sentences of English through examples in use: it turns out that students are not satisfied with copying and pasting sentences containing the terms they translate, but prefer to create their own examples.

In the following sections, we will: present IToldU; evaluate its first two full years of use (describing its pedagogical impact on students and teachers and the quantitative and qualitative lexicographical results obtained when varying the desired quality level); and describe plans for increasing contributions, for extending collection to other languages and types of information, and for synchronization with the Papillon online multilingual lexical database.

II. The IToldU Web Service

A. Teaching Context and Goals

The teaching context is as follows:

- Acquiring and using technical English.
- The most important translation direction is English-French.
- Students don't yet know the technical terms in English and have only recently encountered them in French.
- There are probably 10,000-20,000 terms with which the teacher is not necessarily familiar (either in French or in English).
- The teaching goals of the English courses, over the three years spent in the schools by students, are twofold:
- The base technical vocabulary that is to be learned by all students represents about 10% (1000-2000 items) of the terms.
- Each student should choose and learn a small fraction of the remaining 90%.

Students know how to use between 150-300 specific English words or terms associated with their technical field (paper industry) by the time they leave in the third year. Of course, they know many more general terms, and terms in all other domains encountered during their courses (including other technical fields, work placements, themes and skills seen in traditional English classes, job hunting, etc.).

B. Initial Requirements

During the English courses, each student must collect or create the lexical data for his or her own dictionary, based on texts or other sources given by the teacher. Other words or findings encountered during pursuit of language acquisition can also be added. Students can choose from existing found examples and can correct or create their own. Contributing a translation or selecting an existing example generates a vote for the responsible student.

Teachers and students can restrict their views to the elements most useful to them: students and visitors can search for, create, and memorize translations of technical (or thematic) English expressions, and teachers can run quantitative statistics, control student contributions, and enliven the site using "word hunts," etc. The coordinating teacher is the only one allowed to manage the site (through lists of teachers, students, classes, etc.).

The objective of collecting lexical data is not mentioned to the students and teachers, who are only aware of the pedagogical objectives enunciated by the coordinator:

- Motivating the students to do "lexical" work outside of the class room,
- Minimizing the supplementary workload of the teachers.

C. Implementation

IToldU associates a MYSQL database with each group of students for their three years at the school. It contains the teachers, the students, and the groups of students, with their access rights. It also contains the current dictionary of the group, with students associated with created or adopted entries.

IToldU is programmed in HTML/SQL/PHP, and installed on a free Internet provider (laposte.net, then grenet.fr). It is easy to clone, to install on other sites, and to adapt to other languages, because all messages and menu items are contained in text resources, and can be edited without any special knowledge of programming.

Users have passwords and access rights. The global parameters can only be set by the coordinating teacher. Other teachers can consult students' accounts and direct them. Students can only capture data and consult their personal dictionaries and the dictionary of their group.

D. Usage by Students

Students must seek technical expressions in English and propose correct French translations. For each term, they must include (by citation or creation) an example in context and its source (e.g. from class, booklets, lab sessions, magazines, press, or web or bibliographic sources).

In the examples, the interface is in French, because the students are French speakers. But, as said above, IToldU is easily localizable to other interface languages.

When a student connects to his or her own digital dictionary, he or she finds a summary (Fig. 1) page providing access to the digital dictionary (to search for translations and add new expressions). Also on the page are useful teachers' tools ("Outils") for preparing CVs, application letters, or word hunts. A user can look at his own statistics, measure his knowledge against that of fellow classmates, or print the current dictionary (Fig. 5).

The current access form is minimal: one can only enter an expression or the first letters of an expression in the first input field. However, it has been designed to be easily replaced or combined with richer ones later.

If there is no entry for a word or expression, the student should enter a translation proposal, with an example of use, the context where it was found, and its bibliographical reference. Each voluntary contribution by a student counts toward its statistics and grades.

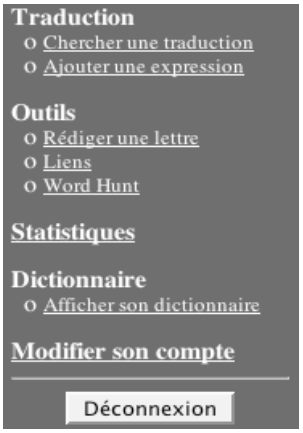

Fig. 1. Students summary.

The principle used for motivating the students and regulating their contributions is simple: the student begins by checking, before introducing a term of interest, whether it has already been handled by a groupmate.

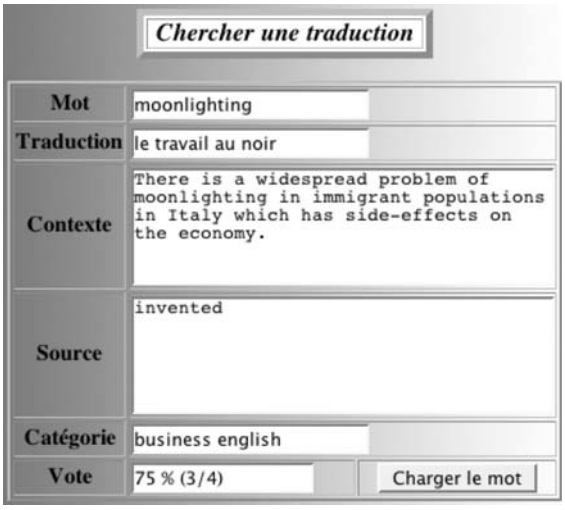

Fig. 2. Form for adding a term in IToldU.

If so, and if the translation and the example look acceptable, s/he can (but does not have to) "adopt" it by adding it to his/her personal dictionary. S/he can also create a new entry, Fig. 2.

Students receive a point for uploading an entry onto their dictionary (effectively "voting" for it). However, if the entry is wrong, the student will lose a point later. In both cases, IToldU motivates students via the possibility of gaining or losing points. This incentive instills in them a positive learning attitude. Moreover, the publication of the "top ten" best scores on the web site motivates them to participate more and more often, creating a healthy competitiveness among individuals and groups.

E. Teachers

IToldU offers teachers the possibility of supervising student groups, encouraging involvement through the use of bonus marks, and livening up vocabulary acquisition via playful "word hunts". Fig. 3 shows the summary of a teacher's session.

S/he can customize general properties (e.g. the title of the site, or its language), broadcast learning activities, contribute to the digital dictionary's construction (by searching for a translation, adding a new expression and creating new technical domains – called "categories"), manage student groups ("*Gestion des comptes*" – account management), and look at the contribution of each student or classroom, as shown in Fig. 4 and Fig. 5 ("*Statistiques*", "*Afficher un dictionnaire*" – display a dictionary).

Fig. 3. Teachers' summary.

A particular blessing is that teachers never have to look inside the source code of an HTML page or (even worse!) other program code. Another important point is that the time constraints of the teachers are taken into account: teachers have almost no time to follow students' work outside the classroom (perhaps 1-2 minutes per student). The use of IToldU should not increase their work time, but if possible reduce it.

That seems to be the case now, as the grading system has been designed to optimize the teacher's time. During the first few weeks of use by a new group, the teacher systematically goes online and deletes any incorrect words. This supervision encourages rigor at of the start of the program.

Fig. 4. Resource pooling statistics.

During the second year, evaluations of contributions are scheduled (every five months) in which teachers check a few dictionary samples from each student in their class. Students don't know which sample will be checked, and are hence motivated to check and improve their entire dictionary. Owing to lack of time and for pedagogical reasons, teachers do not correct mistakes, but simply mark that a translation or an example is wrong. IToldU supports such error marking on fields. Then students must make the corrections before a certain time elapses, or IToldU will subtract the corresponding points.

Fig. 6 shows an example of a "word hunt" screen. "Word hunt" is a challenging but enjoyable part of IToldU for both teachers and students. The first student to find a translation wins a point! Thus students log on as often as possible to see if there are words up for grabs!

III. Evaluation

A. Pedagogical Aspects

Reactions of teachers and students. The current complete version of IToldU (http://opus.grenet.fr/itoldu/ITOLDU) was used for the first time in 2004-05 by all the students of EFPG, an engineering school that is attached to INPG, with a clear positive pedagogical impact. A total of 250 students were involved in the beta test, spread out over the three years of engineering school and one year of professional BA (licence) work. As far as English teaching was concerned, there were 17 groups, 6 teachers, and 1 coordinating teacher (the third author).

IToldU already addresses quite well the need felt by the coordinating teacher for a computer tool improving management of training, teachers' work, and students' learning of specialized English technical vocabulary.

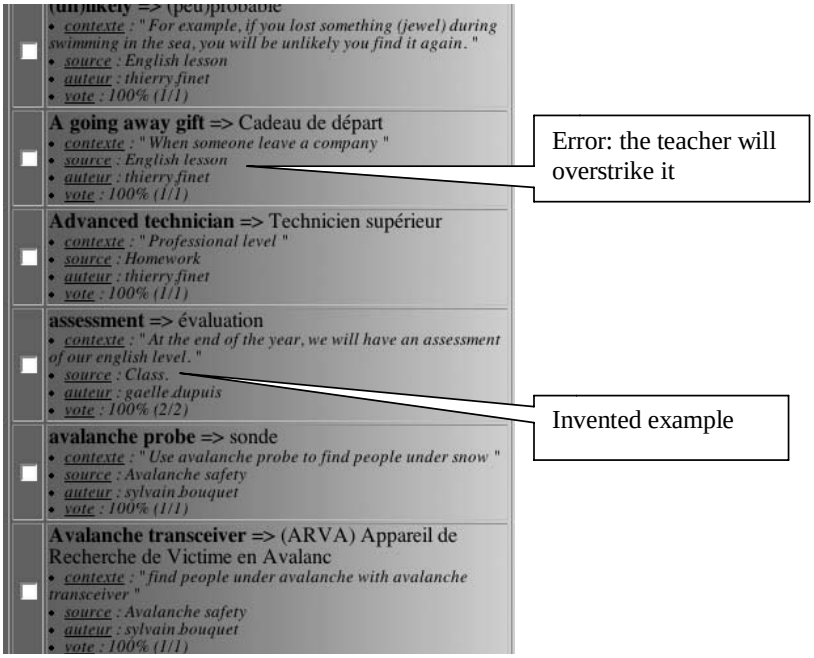

Fig. 5. Fragment of the (sub)dictionary of a class.

perks	Avantages	gaelle.dupuis	DUTpromo13
jobless	Au chômage	gaelle.dupuis	DUTpromo13
Employment agency	Agence de placements	thierry.finet	DUTpromo13
nine-to-five job		Ajouter	
Hire and fire		Ajouter	
Corporate culture		Ajouter	
Long-hours culture		Ajouter	
Casual Friday		Ajouter	
going rate		Ajouter	
cash in hand		Ajouter	
job with scope		Ajouter	

Fig. 6. Word hunt prepared by a teacher.

The use of IToldU has changed the behavior of most students for the better: they are more interested in taking notes. Further, using IToldU outside of classes is seen as a supplementary learning process in the acquisition of technical English vocabulary, and not only as a receptacle into which students are forced to put translations and examples, and which they later ignore.

IToldU not only motivates students by computing part of their grade as a function of their (correct) use of the site. It also allows teachers to establish a spirit of cooperation and emulation among students. On the one hand, as we have seen, students cooperate by "voting" for those whose entries they adopt. On the other hand, the system shows the students who have contributed most on a "scoreboard". Finally, word hunts give rise to a healthy and playful emulation.

Students now consider the long term, because they know they will be allowed to take ITOLDU with them in their professional life as an active copy of their personal dictionary (which can be installed and maintained on a Web site). If they wish, they can take along the entire dictionary built by their classmates.

However, it must be noted that not all teachers were as involved in the adoption and use of IToldU as the third author due to the difficulty of working conditions and lack of time; hence the inequality of the contributions of different classes.

Contributive aspects. The problem of motivating students to contribute and of automatically regulating the global contribution process is a particular case of a more general problem widely recognized as very difficult: that of motivating voluntary and free contributions to the population of knowledge bases. That problem is difficult because there are very few specialists in any field who are willing to give their hard-won the knowledge without return or reward.

Beyond such of rare contributions (which, even if they are large for individuals, represent only a small fraction of the desired knowledge), it is necessary to rely on large numbers of non-specialists, each contributing small, and even fragmentary, knowledge elements. However, in reality, it has always been difficult to obtain numerous individual voluntary and free contributions from a "community of interest".

If contributors gain something by contributing, then the contribution is not "free" in the strictest sense. For example, translators using Oki Electric http://www.yakushite.net/ web site put words in dictionaries because they use freely available online tools for translators (bilingual editor, online dictionaries, proposals from translation memories and from the MT system Pensée), in which contributed words become almost instantaneously active.

If, on the other hand, contributions are truly free, contributors are motivated in some way – of course, as discreetly and pleasantly as possible. That is the case of IToldU, in which almost all users – both teachers and students – are "strongly invited" to use the tool.

B. Dictionary Evaluation (First Year)

1) Quantitative aspect

In the first semester, about 12,000 English-French entries were entered into IToldU by the students, along with about 8,000 usage contexts.

At the end of the academic year, IToldU contained 17,062 English-French entries, and about as many usage contexts (only 157 entries lacked contexts).

2) Qualitative aspect

The second author quickly revised all the contributions of the first year, and about 10% in detail, thereby correcting them. Apart from errors arising from problems in inputting diacritics on the Web, the French translations of English terms are almost all correct. By contrast, 15% to 20% of usage contexts are not examples of use. Following are some details on these two types of contribution.

Translations. 95% of the translations seem correct to us. An interesting point is that only about 30% of the English terms chosen by the students concern a purely technical lexical field, one linked with students' studies (of manufacturing paper pulp, paper, cardboard, color processing, inks, rheology, etc.) while 70% concern "paratechnical" fields, such as business or job hunting, or general English.

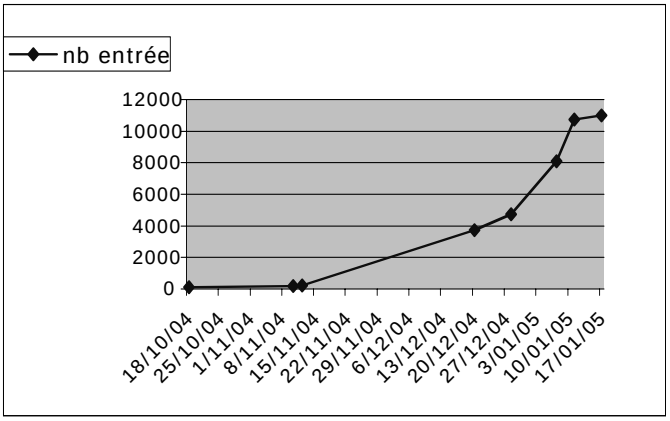

Fig. 7. Evolution of the number of entries in the first semester.

From usage contexts to examples of use. "Contexts" merit some comments. In the mind of the teachers, contexts should be citations of sentences in which the English terms had been encountered. But several unexpected things happened.

Certain students understood that they were being asked for the "domain" of the citation, selected from a list provided by IToldU. One finds for example:

5024	opportunity	possibilité, débouché	society
5025	to put up	ériger, construire	society
5026	to fulfill	accomplir, réaliser	society
5027	fulfilling	profondément, satisfaisant	society
15009	gas-fired	chauffé au gaz	used in paper mill

Others thought that they were being asked for definitions.

15049	a wind mill	une éolienne	an energy-producing facility
15065	a light bulb	une ampoule électrique	energy-related equipment
4632	TCF (totally chlorine free)	sans chlore	stade de blanchiment

The coordinator accordingly modified his description: he asked for "examples in use", and created some himself, putting "invented" in the source field. The students then understood that they, too, could invent examples, and did so. At the level of content, several cases arose:

- Some students created or adapted sentences containing the English terms in question, but in such a way that the word meaning could not be discriminated.

16070	collude	s'associer	they colluded last year
16990	telematics	télématique	it s telematics
16998	darts	fléchettes	he throws the darts
17003	potoling	spéléologie	the potoling is dangerous
17006	chiari-oscuro	clair-obscur	the is a chiarioscuro effect
17026	heir	héritier	you heir to your mother

- At the other extreme, other students used long sentences as examples.

12956	Falsification	Falsification	Some various documents to be protected from counterfeiting and falsification like service vouchers, security label and certificates of authenticity have special features.
12957	service vouchers	Tickets de prestation	
12958	security label	Etiquettes sécurisées	
12959	certificates of authenticity	Certificats d' authenticité	
12960	anti-counterfeiting features	Eléments anti-contrefaçon	
12961	anti-falsification feature	Eléments anti-falsification	

- Many proposals are intended as "honest examples", but are not in correct English.

6619	carriageway	chaussée	the carriageway is destroy by the cars
7073	union	syndicat	an union for help employees
7098	pythonesque	humour absurde	this joke are very pythonesque with his very absurd humor
9183	(to) insulate	isoler	insulating materials can be very useful in electronic

- A small percentage of students vented their frustration by putting "garbage" (silly examples or obscenities) in their examples.

In total, about 15% of the examples are incorrect with respect to content, again not counting input errors, and many

more are incorrect with respect to language, grammar, and spelling.

Hence, there is the origin of the idea to use IToldU not only for learning vocabulary, but also for language learning. Interestingly, students used some of these examples in class during oral performance.

IV. Perspectives

A. Encouraging More Contributions

Other possible ways to encourage more contributions:

- Generalize the "scoreboard" idea to show credits for each entry part.
- Introduce personalization facilities (i.e. automatic or semi-automatic user profiling), so that the system can suggest personalized lists of "things-to-do" or new contributions in the user's domain of interest.
- Allow users to self-organize in groups and groups of groups, each group having certain access rights and a profile.
- Give users access to tools that can extract potential translation pairs from related corpora (texts on the same domain in two or more languages, usually not parallel).
- Let users contribute directly through an "active reading" interface (translated words or idioms appear in annotations of read text).
- Make the importing environment accessible to users wishing to upload sets of translation pairs from any format (Excel, Word, FileMaker, XML, etc.).
- As the ultimate objective, integrate the lexical contribution function as an add-on (plug-in) in as many applications as possible, to be used by the general public.

B. Synchronize Papillon with IToldU

Since the Papillon platform (in particular, its CDM part) accepts any kind of dictionary, provided it is formatted in XML and can be mapped to the CDM DTD, the first problem in linking IToldU and Papillon is to define the mapping of information: are IToldU entries Papillon "lexies", or lemmas, or *vocables*? As seen in the examples above, they are in fact only *vocables* – citation forms without any disambiguating part-of- speech tags.

The second problem is maintenance: the periodic updating of information from IToldU in Papillon.

The fact that the information can be modified under Papillon as well as under IToldU should not be a major problem, as Papillon is designed to keep the contributions of each contributor in his or her private work space, and to allow the creation of groups of contributors. It should then suffice to create one IToldU contributor. Alternatively, if one wishes to keep track of the student contributors in Papillon, one could create a Papillon user for each IToldU student. Papillon groups would correspond to IToldU classes, with one main group for IToldU itself.

The basic idea for maintenance, found to be valid in other contexts, is to compute the differences between two successive states of the IToldU database, and then to compute an update program which can be executed by the Papillon API as if modifications had been made interactively using the Papillon web interface.

C. Extension to Other Language Pairs or Triples

Nothing in IToldU is specific to the English-French language pair, and the software is easy to localize: a language teacher with no programming skills can do it by editing text files.

However, one necessary change is that IToldU should be able to handle three languages in parallel (thereby integrating a second foreign language that a student may also be studying as a course requirement): the two languages used in the classroom and English if it is not one of these.

D. Other Information Types

In the current context of engineering schools, it does not seem possible to obtain sophisticated types of information beyond the lexicographical, such as DiCo semantic formula, definitions, regimes[1], lexico-semantic functions, and other types of collocations. Perhaps the parts-of-speech could be contributed by our students, but nothing more.

Hence, we are trying to find other learning contexts in which such advanced information types are more likely to be contributed by users, such as language schools and translation or interpretation schools.

V. Conclusion

The collaborative construction of free lexical resources is currently hampered by the difficulty of obtaining many small unpaid contributions. IToldU is a light web service which, in its first year of use for teaching technical English in French engineering schools, has led to the contribution of more than 17,000 English terms, in about 20 technical domains, with their French translations (95% correct) and almost as many examples of use (about 85% correct). The quality level has been raised in the second year. In 2 years, 22,000 entries have been created.

IToldU should now be extended to other language pairs, and to language triples. It is also a testbed for a user-friendly method to localize the interface to any language.

It remains to be seen whether IToldU can be synchronized with Papillon, a much more ambitious multilingual lexical database, and to what other contexts of use it could be extended to obtain other types of information, such as regimes, semantic formulas, lexico-semantic functions, or free collocations.

[1] Melchuk's term for the syntactic-semantic valencies, *aka* subcategorization frames.

ACKNOWLEDGMENTS

We would like to thank our reviewers for many useful comments, and Mark Seligman for a very detailed revision of the paper and improvement of its language. All remaining errors are of course ours!

REFERENCES

[1] V. Bellynck, “Bases lexicales multilingues et objets pédagogiques interactifs : Sensillon pour Papillon,” in *Proceedings of Papillon 2002 Seminar*, NII, Tokyo, July 2002, 13 p.

[2] V. Bellynck, C. Boitet and J. Kenwright, “Resource pooling for technical English learning via lexical access,” in *Proc. Papillon-04 seminar*, UJF, Grenoble, 30 Aug.-2 Sept. 2004, 5 p.

[3] V. Bellynck, C. Boitet, and J. Kenwright, “ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases,” in *Computational Linguistics and Intelligent Text Processing (Proc. CICLING-2005)*, A. Gelbukh (Ed.), Springer, LNCS 3406, pp. 319-327.

[4] M. Mangeot-Lerebours, “Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue,” PhD in Computer Science, Université Joseph Fourier, Grenoble I, 280 p., Grenoble, France, 2001.

[5] M. Mangeot-Lerebours, G. Sérasset, and M. Lafourcade, “Construction collab-orative d'une base lexicale multilingue, le projet Papillon,” *TAL*, 44/2, pp. 151-176.

[6] T. Murata, M. Kitamura, T. Fukui, and T. Sukehiro, “Implementation of Collaborative Translation Environment `Yakushite Net’,” in *Proceedings of MT Summit VIII*, New Orleans, Sept. 2003.

[7] N. Tokuda and L. Chen, “An Online Tutoring System for Language Translation,” *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 46-55, July-September 2001.