European Commission (DG Education and Culture) (2004). *Study on innovative learning environments in school education*. Retrieved May 2007, from http://insight.eun.org/ww/en/pub/insight/misc/library.cfm.

Grasha, A. F., & Yangarber-Hicks, N. (2000). Integrating teaching styles and learning styles with instructional technology. *College Teaching, 48*, 2–10.

Griffiths, D., & Blat, J. (2005). The role of teachers in editing and authoring units of learning using IMS learning design. *International Journal on Advanced Technology for Learning, 2* (4).

Hoyles, C. (1993). Microworlds/Schoolworlds: The transformation of an innovation. In C. Keitel & K. Ruthven (Eds.), *Learning from computers: Mathematics education and technology*(pp. 1–17). New York: Springer.

Jones, K. (2005, April). *The shaping of student knowledge with dynamic geometry software*. Paper presented at the Computer Assisted Learning Conference. 2005 (CAL05), Bristol, United Kingdom. Retrieved September 2006, from: http://eprints.soton.ac.uk/18817/.

Kent, P., Hoyles, C., Noss, R., Guile, D., & Bakker, A. (Eds.). (2007). Learning technologies at work [Special issue]. *Mind Culture and Activity, 14*(1–2).

Lagrange, J. B., Artigue, M., Laborde, C., & Trouche, T. (2003). Technology and mathematics education: A multidimensional study of the evolution of research and innovation. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 239–271). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lee, C. P. (2007). Boundary negotiating artifacts: Unbinding the routine of boundary objects and embracing chaos in collaborative work. *Computer Supported Cooperative Work, 16*, 307–339.

Monaghan, J. (2004). Teachers' activities in technology-based mathematics lessons. *International Journal of Computers for Mathematical Learning, 9*, 327–357.

Noss, R. (1995). Computers as commodities. In A. A. diSessa, C. Hoyles & R. Noss (Eds.), *Computers and exploratory learning* (pp. 363–381). Berlin, Germany: Springer.

Noss, R., Bakker, A., Hoyles, C., & Kent, P. (2007). Situating graphs as workplace knowledge. *Educational Studies in Mathematics, 65*, 367–384.

Papert, S. (2006). Afterward: After how comes what. In K. R. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 581–586). Cambridge, United Kingdom: Cambridge University Press.

Pelgrum, W. J. (1996). The educational potential of new information technologies: Where are we now? In B. A. Collis, G. A. Knezek, K. -W. Lai, K. T. Miyashita, T. Sakamoto et al. (Eds.), *Children and computers in school* (pp. 118–119). Mahwah, NJ: Lawrence Erlbaum.

Rabardel, P. (1995). *Les hommes & les technologies. Approche cognitive des instruments contemporains* [Human beings and technologies: A cognitive approach of contemporary instruments]. Paris: A. Colin.

Sutherland, R. (2004). Designs for learning: ICT and knowledge in the classroom. *Computers and Education, 43*, 5–16.

Vakkari, P. (1999). Task complexity, problem structure and information actions – Integrating studies on information seeking and retrieval. *Information Processing and Management, 35*, 819–837.

Venezky, R. L., & Davis, C. (2002). *Quo vademus? The transformation of schooling in a networked world* [OECD/CERI, Version 8c, March 06]. Retrieved November 2007, from http://www.oecd.org/dataoecd/48/20/2073054.pdf.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Warfield, V. M. (2006). *Invitation to didactique*. Retrieved January 2008, from http://www.math.washington.edu/∼warfield/Inv%20to%20Did66%207-22-06.pdf.

Wilensky, U. (2003). Statistical mechanics for secondary school: The GasLab multi-agent modeling toolkit. *International Journal of Computers for Mathematical Learning, 8*, 1–4.

Chapter 6
Integrated Digital Language Learning

Georges Antoniadis, Sylviane Granger, Olivier Kraif, Claude Ponton, Julia Medori and Virginie Zampa

Abstract While the field of technology-enhanced language learning (TELL) is undeniably thriving, most technology-enhanced language tools are still relatively crude. One reason for this is that the field is disconnected from research in natural language processing (NLP) and corpus linguistics (CL), two fields which could greatly improve the effectiveness of most pedagogical tools. The research carried out within the framework of the Kaleidoscope Network of Excellence aimed to demonstrate that it is both possible and desirable to integrate insights from NLP and CL into TELL to produce more powerful and effective tools. In the article we give a general outline of NLP and CL techniques and highlight their relevance for TELL. We also describe two types of integration that were implemented within the framework of Kaleidoscope: (1) integration of NLP processing into the glossary of the *Moodle* Learning Management System; (2) integration of error-tagged learner corpus data into *Exxelant*, a web-based error interface for teachers and researchers. The chapter also argues the case for optimising the role of language in all technology-enhanced learning applications, whether language focused or not.

Keywords Natural language processing (NLP) · Language learning · Computer-assisted language learning (CALL) · Technology-enhanced language learning (TELL) · Corpus · Learner corpus · Learning Management System · *Moodle* · Glossary· Error · Error tagging · Error feedback · Error interface

6.1 Introduction

Technologies have never been as much in the forefront of language learning as they are now. They have admittedly played an ever increasing role ever since the introduction of audiolingual methods, but today we are truly witnessing a technological explosion in the field, with a host of new developments such as web-based

S. Granger (✉)
Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve, Belgium
e-mail: sylviane.granger@uclouvain.be

N. Balacheff et al. (eds.), *Technology-Enhanced Learning*,
DOI 10.1007/978-1-4020-9827-7_6,

learning platforms, computer-mediated communication, blogs, wikis, whiteboards and the use of mobile devices such as iPods, PDAs and mobile phones. In this technology-rich environment, one would expect close links with two highly relevant language-related fields, namely natural language processing (NLP) and corpus linguistics (CL). Both are clearly of high relevance for language learning and teaching. NLP provides tools capable of automating language analysis and providing feedback on learner productions. CL offers large quantities of text in electronic format and tools to explore them quickly and efficiently. However, the impact of NLP and CL on technology-enhanced language learning (TELL) is still very limited, as attested by the very small number of articles dealing with these issues in major scientific journals. It is symptomatic, for example, that Chun's (2007) survey of major topics tackled in the latest issues of CALICO does not contain a line on those research strands. Most TELL specialists are still not aware of the relevance of NLP and CL. The three scientific communities remain quite separate, each with their own paradigms, terminology, scientific journals and conferences. Although some special interest groups are very active,[1] integration is still minimal. There are several reasons for this. One major factor is that NLP techniques are not foolproof and language practitioners do not want to have to deal with errors due to the software used. The fact that corpus linguistics is still a very young field also plays a role. As demonstrated by Mukherjee's (2004) survey among English-language teachers in Germany, the majority of language teachers show little familiarity with corpus tools and methods.

In this chapter we focus on these two neglected but highly promising aspects and report on a small-scale project carried out within Kaleidoscope to demonstrate the contribution that they can make to TELL. Sections 6.2 and 6.3 of this chapter give a brief overview of NLP techniques and corpus linguistics methods and tools and highlight their respective relevance for language learning and teaching. In Section 6.4 we describe the obstacles to the integration of NLP and corpus techniques into TELL and suggest ways of circumventing them. In Section 6.5 we demonstrate the feasibility of integration by describing two prototypes designed within the framework of Kaleidoscope: an intelligent glossary and a web-based error interface. In Section 6.6 we widen the perspective and highlight the potential impact of this type of research on the general field of technology-enhanced learning.

6.2 Natural Language Processing

Natural language processing is a multidisciplinary research field, at the crossroads of linguistics, computer science and artificial intelligence. It deals with the problems of understanding and generating natural human languages. Among the many NLP techniques, the following are particularly relevant for TELL: tokenisation,

[1] For example, EUROCALL's NLP Special Interest Group and CALICO's Intelligent Computer-Assisted Language Instruction group.

morphological processing, syntactic processing, speech recognition and synthesis and concordancing.[2] Here is a quick review of these techniques, starting from the simplest ones:

- *Tokenisation* is the very first operation of text processing: it consists of segmenting a text, that is, a sequence of characters, to get a sequence of lexical units, or tokens (e.g. punctuation marks, numbers, words). This simple operation leads, for example, to spell checking, by comparing the resulting tokens with recorded lists of inflected forms.
- *Morphological analysis* aims at analysing the morphemes that compose lexemes, in order to determine their morphological category (part-of-speech or POS), features (inflections), components (affixes) and canonical form (lemma). In many languages, state-of-the-art tools allow automatic POS-tagging and lemmatising with a very good accuracy (over 95% precision). Such analysis allows for many interesting applications: error diagnosis, as when the learner uses a correct form with an erroneous inflection (Kraif and Ponton, 2007), or glossing inflected terms in a text, as in the intelligent glossary described in Section 6.5.1.
- *Syntactic parsing*, usually taking POS-tagged and lemmatised texts as an input, aims at extracting dependency relations between lexemes, or hierarchical relations between phrases (constituents). Parsing is required, for example, to detect the erroneous verbal inflection in the following utterance: *The inhabitants of this country *suffers from malnutrition*, where the head of the noun phrase bearing subject function is *inhabitants* and not *country*. Because of syntactic ambiguities and computational complexity limitations, this analysis remains a tricky problem for unconstrained utterances. The best parsers hardly get fewer than 25% errors for standard written language, without full coverage of the sentences. Improved parsing would be a huge step forward for error detection and analysis.
- *Speech recognition* aims at discriminating through an acoustic signal the sequence of phonemes – and then lexemes – that composes the oral message. It is a particular problem of form recognition: discrete structures must be extracted from a continuous signal where many variations occur (tempo, pitch, accent, voice, intensity) without being relevant. Although considerable progress has been achieved with probabilistic models of language, these techniques are highly problematic and get low results for unexpected messages in a noisy environment.
- *Speech synthesis* is the reciprocal process to recognition. Text-to-speech systems are designed to convert written utterances (sometimes with phonetic and prosodic indications) into their oral form, using various parameters such as pitch, tempo and voice tone. It is an easier problem than recognition and many everyday life devices, such as GPS and phones, already implement this technology. The final quality depends closely on prosodic processing, which is an essential component for communication.

[2] Other major NLP techniques, such as machine translation, will not be described here as they are arguably less relevant for TELL. See Mitkov (2003) for a comprehensive overview of the field.

- *Concordancing* is dedicated to the extraction of examples from a corpus, searching for a given expression and its surrounding context. Concordances are often presented in KWIC (keyword in context) format, where left context, key expressions and right context appear in aligned columns. Modern concordancers allow searching not only for character strings but also for lemmas, compound units and morphosyntactic features, including NLP formalisms such as finite state automata or regular expressions. By sorting the data in various ways, users have easy access to the typical use of words or phrases. For example, a search for the verb "argue" in a corpus of native English academic writing instantly brings out the typically passive use of this verb in patterns like *it can/could/might be argued that*. . . or *it has been argued that*. . .

Because it is as old as modern computer science, NLP has yielded many mature technological outcomes in various fields such as machine translation, dialog generation, spell and grammar checking, information retrieval, speech recognition and speech synthesis. Applications for language learning appear to be a natural extension of these technologies. As stated by Nerbonne (2003),

> NLP focuses on how computers can best process language, analyze, store, sort and search it. It seems natural that NLP should be applied to the task of helping people learn language (p. 678).

NLP techniques are indeed numerous and cover a wide range of needs in language engineering. More than 20 years after the beginning of the rapprochement between NLP and computer-assisted language learning (CALL), many prototypes or experimental systems have been developed. For instance, some systems make use of POS-tagged and lemmatised texts to generate gap-fill exercises where the gaps are selected on the basis of morphosyntactic and/or semantic criteria (e.g. only personal pronouns or only time adverbs are gapped) (Antoniadis et al., 2004; Selva, 2002). Other systems, such as the Exills platform (Brun, Parmentier, Sandor, & Segond, 2002), give the learner access to NLP-enhanced linguistic tools (conjugators, disambiguated dictionaries, tagging, language identification, etc.) as an aid to producing and understanding utterances in a virtual environment.

Surprisingly, however, commercial systems are extremely rare and research developments remain at the stage of prototypes. This is due to the following three factors:

- The lack of reliability of NLP technologies.
- The high cost of NLP research and development and the lack of system modularity.
- The lack of interdisciplinary communication (didactic/linguistic/NLP).

Concerning the last two points, the NLP community is currently striving towards standardisation and one sees more and more "generic" resources with free software development (concordancers, taggers, lemmatisers, etc.). Generally, these programs do not require any modification other than the adaptation of the input/output formats and of the basic parameters. In view of the current state of the art, using the simplest tools is likely to bring major improvements, which more than compensate for the

modest investment made (see Section 6.5). As for collaboration between language practitioners and NLP specialists, various projects or networks such as Kaleidoscope demonstrate that it is clearly underway even if there is still scope for greater synergy.

6.3 Corpus Linguistics

Corpus linguistics can be defined as a linguistic methodology that is founded on the use of large electronic collections of naturally occurring texts, namely corpora. There are many different types of corpus: spoken and written, monolingual and multilingual, diachronic and synchronic, etc. Some corpora are meant to be representative of a language as a whole and therefore contain texts from a wide range of written and spoken sources (fiction, journalese, academic writing, informal conversation, political speeches, etc.). A good example of this type of corpus is the *British National Corpus*[3] (Aston & Burnard, 1998). Others, like the *Micase* corpus of academic spoken English,[4] are more limited in scope and cover only one text type. One relatively new corpus type that is particularly relevant for language learning and teaching is the learner corpus containing written or spoken data produced by foreign-language learners (for a survey of learner corpus research, see Granger, 2008a,b). For example, the *International Corpus of Learner English* (ICLE) CD-ROM contains writing produced by learners from 11 different mother tongue backgrounds (Granger, Dagneaux, & Meunier, 2002).

The fact that corpus data are in electronic format makes it possible to automate the analysis of a large amount of data. First, the data can easily be quantified; second, it is easy to get accurate information on the preferred environment of linguistic items; and third, it is possible to enrich the data with a wide range of linguistic annotations, notably by means of NLP techniques such as lemmatisation or POS-tagging.

In the following, we illustrate the power of corpus techniques with reference to learner corpora.

1. *Frequency*. Text retrieval software tools such as *WordSmith Tools* (WST) (Scott, 2004) are language-independent programs that enable researchers to count and sort words in text samples automatically. Using these tools, researchers have immediate access to frequency lists of all of the single words or sequences of words in their corpora. Lists derived from learner corpora can be automatically compared to lists based on comparable native speaker corpora, thereby revealing the words or phrases that learners tend to over- or underuse. By way of illustration, Table 6.1 lists the 10 most underused verb forms in the ICLE corpus as

[3] A simple search service for the BNC is offered at http://www.natcorp.ox.ac.uk/index.xml.

[4] The online, searchable part of the Micase corpus is available at http://quod.lib.umich.edu/m/micase/.

Table 6.1 Top 10 underused verb forms in the ICLE corpus

	Verb form	Keyness
1	described_VVN	554,7
2	seen_VVN	423,8
3	suggests_VVZ	363,1
4	argues_VVZ	332,9
5	required_VVN	330,0
6	remained_VVD	287,2
7	obtained_VVN	249,4
8	shown_VVN	242,9
9	appears_VVZ	233,7
10	held_VVN	231,9

compared to a comparable native academic corpus ordered in decreasing order of keyness.

2. *Patterning*. Corpus tools included in packages like WST, in particular phrase (or chunk) extraction and concordancing, are very powerful heuristic devices for uncovering recurrent patterns of use, or to put it another way, words' preferred lexical and grammatical company. Applying the phrase extraction method to a corpus of EFL speech and a comparable native speaker corpus, de Cock (2004) shows that EFL learners significantly underuse discourse markers such as *you know* or *I mean* and vagueness markers such as *sort of* or *and things* and therefore prove to be lacking routinised ways of interacting and building rapport with their interlocutors and of weaving in the right amount of imprecision and vagueness, both typical features of informal interactions. On the other hand, concordancers make it possible to extract all occurrences of a given lexical item (single word or phrase) in a corpus and sort them in a variety of ways, thereby allowing typical patterns to emerge. The concordance of the verb *argue* in learner writing highlights a preference for active structures such as *people argue* or *some people may argue*, which differ from the typical passive pattern brought out by the native concordance.
3. *Annotation*. In corpus linguistics terms, the term "annotation" refers to "the practice of adding interpretative (especially linguistic) information to an existing corpus of spoken and/or written language by some kind of coding attached to, or interspersed with, the electronic representation of the language material" (Leech, 1993, p. 275). In learner corpus terms, this means that any information about the learner samples that the researcher wants to code can be inserted in the text. Although there is no limit in principle to the type of annotation that can be used to enrich a learner corpus, two are by far the most commonly used: morphosyntactic annotation and error annotation. While the first type of annotation is an NLP technique (see Section 6.2), the latter is still largely manual. It consists of marking each error in learner corpora with a standardised system of error codes together with the error correction. For example, the above-mentioned error *The inhabitants of this country *suffers* will be coded as a grammatical error affecting a lexical verb and belonging to the category of concord errors. The correct form

suffer is also included with the appropriate mark-up. Error-tagging is a highly complex and time-consuming process, but it is a necessary step for automatic error detection.

6.4 NLP, Corpora and TELL

Both NLP and corpus research have a major role to play in TELL. NLP makes it possible to analyse language in much more sophisticated ways and several widely available NLP tools could easily be integrated into TELL applications. This said, NLP technologies are not 100% foolproof and their relative unreliability is a major obstacle, as the didactic context precludes the integration of erroneous input or feedback. For this reason, learner production analysis remains a problematic task. The more promising attempts concern very constrained contexts, where production variability is finite. Heift and Nicholson (2001) describe "German Tutor", a tutoring system that involves syntactic parsing of learner answers, with a high accuracy. Kraif and Ponton (2007) give a global framework for short answer analysis and error diagnosis and present an experiment that shows how very simple NLP techniques may yield high accuracy when comparing the learner's answer with an expected one. As suggested by the latter authors, it is advisable to favour such modest integration of NLP tools.

More realistic NLP applications in TELL concern the use and processing of native and learner corpora. Corpora give language teachers a practically inexhaustible source of examples of "real" native language, the type of language that the students will have to use in communicative situations. NLP makes it possible to search not only for character strings, but also for linguistic forms, namely lemmas, morphemes, morphosyntactic features, functional relations or complex patterns. This vastly extends the potential of corpus analysis and enhances searching functionalities in monolingual or multilingual corpora (Kraif & Tutin, in press).

Native corpora can be conceived of as large repositories of examples that illustrate specific linguistic phenomena, ranging from lexicon to morphology, syntax, phraseology, terminology and even translation (in the case of a multilingual corpus). NLP techniques are useful for adding comprehension aids to these texts: lemmatisation allows linking of inflected forms with entries in a dictionary (Antoniadis et al., 2004), and the results of automatic annotation may be directly displayed to the learner in order to help him understand the lexicon and grammar structure (Dokter & Nerbonne, 1998; Dokter, Nerbonne, Schurcks-Grozeva, & Smit, 1998).

Another promising development is the possibility of searching for new examples at each query (by a random selection of the parsed texts). By dynamic retrieval of examples, new activities can be generated every time the system is accessed. This is the case for Alfalex (Selva, 2002), where gap-fill exercises allow practicing of French inflectional and derivational morphology, conjugations, prepositions, collocations, etc., with sentences that are extracted on-the-fly from a corpus. The data-driven learning approach has given rise to a large amount of work, resources

and systems (Tribble & Barlow, 2001), which could be greatly enhanced by the addition of simple NLP techniques.

In their error-tagged format especially, *learner corpora* constitute an unparalleled resource that provides a very accurate profile of learners' degree of accuracy, complexity and fluency in the target language. They lend themselves to two types of pedagogical uses: direct and indirect (Römer, 2008):

- *Direct use*. Learners can compare data extracted from learner corpora and compare them with similar data from native corpora to discover differences between the two. Data-driven learning activities of this type may contribute to raising learners' awareness of their own difficulties and promoting learner autonomy (Bernardini, 2004).
- *Indirect use*. Materials designers can use learner corpora to draw up catalogues of learners' attested difficulties and thereby ensure that the pedagogical materials meet learners' needs. Learner corpus insights can be integrated into TELL in two different ways:
  - Non-NLP based: production of remedial TELL resources that tackle recurring errors (cf. Granger, 2003: CALL exercises targeting attested errors produced by learners of French as a Foreign Language; Chuang & Nesi, 2006: web-based resource called *GrammarTalk* which tackles recurring errors made by Chinese students).
  - NLP based: use of NLP techniques to design automatic error detection and feedback systems (cf. Izumi, Uchimoto, & Isahara, 2004; L'haire, 2004; Vandeventer, 2001). The main weaknesses of these techniques are their low precision and recall rates: results are disappointing for a wide range of error types and more corpus analyses are needed to improve the overall success rate. Learner corpora can be used as a benchmark to assess the efficiency of various NLP techniques. As demonstrated by Metcalf and Meurers (2006), different types of word order errors call for different processing: those involving phrasal verbs (e.g. *they give up it*) can be handled successfully by means of instance-based regular expression matching, while errors involving adverbs (e.g. *it brings rarely such connotations*) require more sophisticated parsing algorithms. A corpus containing learner errors is useful in determining which errors fall within the scope of which technique.

6.5 NLP-Enhanced TELL Applications

Two prototypes have been designed within the framework of Kaleidoscope[5] with a view to demonstrating how simple NLP techniques and learner corpus insights can be used to enhance TELL:

[5] The prototypes have been developed in the framework of the Integrated Digital Language Learning (IDILL) project, funded within the framework of the Kaleidoscope Network of Excellence. http://www.noe-kaleidoscope.org/group/idill/Home/.

- Integration of a POS-tagger into the *Moodle* glossary function.
- Design of a web-based error interface, *Exxelant.*

6.5.1 Intelligent Glossary

Glossing consists of providing additional information on words (definition, translation, additional examples, grammatical information, etc.). Several studies have demonstrated that computerised reading with full glossing may promote vocabulary acquisition. Constantinescu (2007) studies the benefits of CALL for vocabulary acquisition and reading comprehension and comes to the conclusion that "one great way to increase vocabulary acquisition and retention is the use of computerised reading passages enhanced with various types of glosses". The use of electronic glossing is supported by other studies such as Lomicka (1998), Al-Seghayer (2001) and Yoshii (2006).

According to these studies, glossing of difficult terms would seem like an essential tool for language learners' vocabulary acquisition. Some programs have been designed for this purpose, for instance the *Glosser* system which involves advanced morphological analysis (Dokter et al., 1998; Dokter & Nerbonne, 1998; Nerbonne, Dokter, & Smit, 1998). However, these tools tend to be stand-alone platforms and many – like *Glosser* – have been discontinued. In today's educational institutions, the adoption of one Learning Management System (LMS) for the whole institution is often recommended. The concurrent use of another learning environment is difficult to manage for both teachers and learners. The best solution is therefore to adapt existing LMSs and/or create tools that are portable to other platforms. Preference should be given to well-disseminated open source platforms such as *Moodle* for at least two main reasons. First, they can be run with limited resources and support and can therefore contribute to reducing the digital divide globally. Second, these platforms have a very large user base and being part of a lively community of users worldwide is a real boost for both teachers and learners.[6]

Despite their usefulness, glossaries are rarely present in Learning Management Systems. Botturi's (2004) survey of nine LMSs shows that only five of those tested have a glossary. In addition, existing glossaries tend to be quite rudimentary and user unfriendly. *Moodle*, the top LMS today and arguably the best (cf. Graf and List, 2005), is an exception. Its glossary is more sophisticated, as it includes an auto-linking functionality. As soon as a word or phrase is entered in the glossary, it will automatically show up in each new text where the word or phrase appears. This is clearly an improvement which allows for "economies of scale" for the teacher. However, the glossary has two major flaws. First, it is linguistically crude, as it relies on simplistic pattern-matching techniques: to be recognised, a word needs to have

[6] *Moodle* has over 400,000 registered users in 193 countries and several discussion groups, including a special "*Moodle* for Language Teaching" forum. More information can be found on the *Moodle* website: http://moodle.org/.

exactly the same form as the word entered in the glossary.[7] For instance, the forms *went* and *go* are not recognised as forms of one and the same lemma GO. Even if the basic form *go* is already in the glossary, the form *went* will not be automatically linked to the glossary entry. The glossary is not "intelligent", that is, it does not rest on any linguistic analysis. Second, the interface makes it difficult for teachers to correct any erroneous link. As part of the Kaleidoscope project, we have remedied these two flaws by (1) integrating a POS-tagger into the *Moodle* glossary tool and (2) improving *Moodle*'s text view interface.

For the first operation, we opted for the *TreeTagger*, an open source POS-tagger developed by the University of Stuttgart[8] which has the advantage of being available for several languages. We integrated the English version of the tagger into *Moodle*. The entire text goes through the tagger, which outputs the grammatical categories and basic word forms of each word. As a result, the form *provides*, for example, is analysed as an inflected form of *provide* and automatically linked to the glossary entry *provide*.

The second stage, namely the improvement of the teacher interface, is all the more necessary as the POS-tagging is not 100% error-free. For instance, depending on the context, *leaves* can be considered as the plural of the noun *leaf* or as the third person singular of the verb *to leave*. This is not straightforward for a computer program, which often generates the wrong analysis. Therefore, we needed to be able to provide teachers with ways to correct these mistakes, as it is not acceptable to provide learners with resources that contain errors. It was therefore necessary to give teachers quick and easy control over the glossary links. In the new interface, when a teacher is logged in and enters a new text, all of the words in the text are clickable and open a pop-up window, in which there is either the glossary entry for this word if it is already in the glossary or an empty entry if it is not. A box was added in the pop-up window that could be ticked if the teacher wanted to remove a link and another box if the teacher wanted to correct an erroneous link (e.g. if *leaves*, plural of *leaf*, is in the text but it is automatically linked to the verb *leave*). Providing user-friendly interfaces is essential for all technology-enhanced tools, as it can boost acceptance among teachers who often – and at times quite rightly – view them as disruptive rather than sustaining innovations.

6.5.2 Error Interface

As part of the Kaleidoscope project, we have designed a web-based error interface, called *Exxelant*[9] (Granger, Kraif, Ponton, Antoniadis, & Zampa, 2007), that can give researchers, teachers and learners easy and versatile access to authentic learner

---

[7] It is possible to add variants to the glossary but this is cumbersome for teachers, especially in the case of languages with extended morphology.

[8] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html.

[9] *Exxelant* stands for EXample eXtractor Engine for LANguage Teaching.

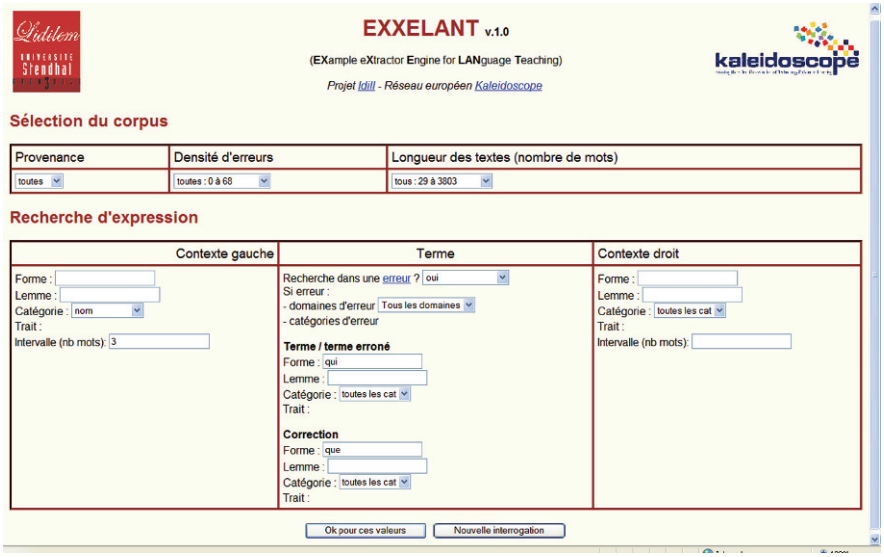

Fig. 6.1 Search for errors concerning the confusion between "qui" and "que" as a relative pronoun

errors and their corrections. Taking as input an XML formatted corpus, which contains error annotations and morphosyntactic tags, this tool allows extraction of examples using a query system that combines various kinds of criteria: error category, part-of-speech, corrected forms, error-prone forms, learners' mother tongue and level. As part of the project, the tool has been tested on a POS-tagged version of a corpus of learner French, the FRIDA corpus.[10]

To illustrate how *Exxelant* works, we take the example of teachers wanting to investigate learners' errors affecting relative pronouns, and more particularly cases where the subject pronoun *qui* is used instead of the object pronoun *que* in environments where the pronoun has a noun as a left-hand context. As shown in Fig. 6.1, the interface is divided into two main parts. The first (*sélection du corpus*) allows users to select the corpus: source (whole corpus or only part of it), error density (numbers of errors per 100 words) and text length. The second part (*Recherche d'expression*) allows users to specify their query on the basis of the left-hand context, the term (errors and/or correction) and the right-hand context. In our example, we are searching for an erroneous term (i.e. "forme=qui" and "erreur=oui") for which the corrected form is "que" (i.e. "forme=que"). This term must be preceded by a noun ("catégorie=nom"). Such a query outputs sentences such as "*Les étudiants **qui [que]** j'ai rencontré pendant le cours m'ont aidé à m'intégrer sans problème*". Users can access the complete learner production for each sentence.

[10] The FRIDA learner corpus (FRench Interlanguage DAtabase) is a corpus of French as a Foreign Language compiled within the framework of the EU-funded FreeText project (Granger, Vandeventer, & Hamel, 2001, Granger, 2003).

Although *Exxelant* was initially designed for teachers, it has many features in common with Hegelheimer and Fisher's (2006) *iWRITE* system which was designed to be used directly by learners in activities of noticing and collaborative error solving. As pointed out by the authors, the tool "can be used to raise learners' grammatical awareness, encourage learner autonomy, and help learners prepare for editing or peer editing" (p. 270). *Exxelant* could easily be adapted to perform similar functions.

The expansion of the Internet makes it possible to share and disseminate these resources and systems, which could greatly contribute to the expansion of corpus use in language learning. Several CALL systems now use and exploit raw or annotated corpora; the care taken in compiling and annotating these corpora contributes greatly to the overall quality of the programs.

6.6 Conclusion: From TELL to TEL

This study has pleaded for greater integration of natural language processing and corpus insights into TELL. Things are clearly moving as regards corpora, as evidenced by the fact that one of the latest issues of ReCALL journal is entirely devoted to "Integrating corpora in language learning and teaching" (Chambers, 2007), but as pointed out by the editor, the articles in the volume "represent only part of the potential of this developing area" (ibid: 250). In particular, learner corpora deserve more attention than they have received so far. As for NLP, one of the main factors that account for the current lack of integration was pointed out by Holland over 10 years ago and is still valid today:

> The most important reason for this failure is that NLP (Natural Language Processing) programs which underlie the development of ICALL cannot account for the full complexity of natural human languages (Holland, 1995, p. viii).

However, we claim that there is no need to wait until NLP can account for the "full complexity" of language to bring NLP and TELL closer together. The research carried out within the Kaleidoscope network has demonstrated that it is possible and indeed desirable to integrate NLP technologies, provided certain conditions are met: (1) only technologies that have a high degree of reliability are used; (2) the techniques are used in carefully selected contexts; and (3) teachers are given full control over the output to facilitate correction in case of error. In other words, what we need is a judicious combination of audacity and caution. Combined use of NLP and CL techniques can lead to a great leap forward in automatic error feedback and automatic rating, two fields where Milton (2002) suggests "it is particularly worth investing in research" (p. 24).

In this project, we have focused on web-based environments, and more particularly on Learning Management Systems. Our study confirms that LMSs need to be adapted to meet the needs of the different fields as suggested by Graf and List (2005) and Kukulska-Hulme and Shield (2004). Future research should focus on fuller adaptation of LMSs to the discipline of language learning, and the components of the ideal LLMS, that is, Language Learning Management System, should be

identified and implemented. At this stage, it is still debatable whether a totally new type of platform should be built or whether existing platforms such as *Moodle* can be expanded with discipline-specific interoperable modules. Another avenue for future research lies in the rapid development of mobile language learning environments (Chinnery, 2006; Gilgen, 2005; Kiernan & Aizawa, 2004; Kukulska-Hulme, 2007). The migration of NLP and corpus technologies to these new environments is one of the major challenges for the TELL agenda.

But integration should go further than that. Natural language is ubiquitous in technology-enhanced learning (TEL): it is present in both the input (texts, instructions, scripts) and the output (answers to exercises, collaborative writing, etc.) of the learning process and is the main channel of interactive communication between the tutor and the learner and between the learners. Sophisticated automatic analysis should therefore be a major feature of all TEL applications, in both hard and soft sciences, not only in language learning. It can help develop new types of scaffolding tools which will foster independent inquiry by learners. Intelligent glossaries, for example, have a role to play in all disciplines. Medical TEL applications, for example, would clearly benefit from an intelligent glossary of medical terms automatically linked to multimedia files and hyperlinked to domain-specific corpora for additional examples. On the other hand, learner output that consists of language – be it in the form of answers to questions or interactions via email, forum, blog or chat – is a particularly rich type of "trail" left behind by learners in TEL environments (cf. Chapter 12). These language trails can be submitted to a wide range of linguistic analyses, some of which, such as automatic discourse analysis (cf. Hilbert, Lobin, Bärenfänger, Lüngen, & Puskás, 2006), are particularly relevant. The applications seem limitless and constitute a near virgin territory waiting to be explored.

References

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning and Technology, 5*, 202–232.

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé, T., Loiseau, M., & Ponton, C. (2004). NLP-based scripting for CALL activities. In L. Lemnitzer, D. Meurers & E. Hinrichs (Eds.), *Proceedings of Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning* (pp. 18–25). Retrieved May, 28, 2008 from http://acl.ldc.upenn.edu/coling2004/W6/index.html.

Aston, G., & Burnard, L. (1998). *The BNC handbook. Exploring the British national corpus with SARA*. Edinburgh, United Kingdom: Edinburgh University Press.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15–36). Amsterdam: Benjamins.

Botturi, L. (2004). *Functional assessment of some open-source LMS*. Unpublished manuscript. Retrieved May 31, 2008, from http://www.elearninglab.org/docs/risorse/report/OS_review_Nov2004.pdf

Brun, C., Parmentier, T., Sandor, A., & Segond, F. (2002). Les outils de TAL au service de la e-formation en langues [NLP tools for e-language learning]. In F. Segond (Ed.), *Multilinguisme et traitement de l'information* (pp. 223–250). Paris: Hermès Science Publications.

Chambers, A. (Ed.). (2007). Integrating corpora in language learning and teaching [Special issue]. *ReCALL, 19*(3).

Chinnery, G. M. (2006). Emerging technologies. Going to the MALL: Mobile Assisted Language Learning. *Language Learning and Technology, 10*, 9–16.

Chuang, F. -U., & Nesi, H. (2006). An analysis of formal errors in a corpus of Chinese student writing. *Corpora, 1*, 251–271.

Chun, D. M. (2007). Come ride the wave: But where is it taking us? *CALICO Journal, 24*, 239–252.

de Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures, New Series 2*, 225–246.

Constantinescu, A. I. (2007). Using technology to assist in vocabulary acquisition and reading comprehension. *The Internet TESL Journal, 13*(2).

Dokter, D., & Nerbonne, J. (1998). A session with Glosser-RuG. In S. Jager, J. Nerbonne & A. van Essen (Eds.), *Language teaching and language technology* (pp. 88–94). Lisse, The Netherlands: Swets & Zeitlinger.

Dokter, D., Nerbonne, J., Schurcks-Grozeva, L., & Smit, P. (1998). Glosser-RuG: A user study. In S. Jager, J. Nerbonne & A. van Essen (Eds.), *Language teaching and language technology* (pp. 167–176). Lisse, The Netherlands: Swets & Zeitlinger.

Gilgen, R. (2005). Holding the world in your hand: Creating a mobile language learning environment. *EDUCAUSE Quarterly, 28*, 30–39.

Graf, S., & List, B. (2005). An evaluation of open source e-learning platforms stressing adaptation issues. In P. Goodyear, D. Sampson, D. Jin-Tan Yang, Kinshuk, T. Okamoto et al. (Eds.), *Proceedings of the International Conference on Advanced Learning Technologies* (pp. 163–165). Washington, DC: IEEE Computer Society Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO, 20*, 465–480.

Granger, S. (2008a). Learner corpora in foreign language education. In N. Hornberger (Ed.), *Encyclopedia of language and education* (Vol. 4) (pp. 337–351). New York: Springer.

Granger, S. (2008b). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*.(Vol. 1) (pp. 259–275). Berlin, Germany: Walter de Gruyter.

Granger, S., Dagneaux, E., & Meunier, F. (Eds.). (2002). *The International Corpus of Learner English: Handbook and CD-ROM*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL, 19*, 252–268.

Hegelheimer, V., & Fisher, D. (2006). Grammar, writing, and technology: A sample technology-supported approach to teaching grammar and improving writing for ESL learners. *CALICO Journal, 23*, 257–279.

Heift, T., & Nicholson, D. (2001). Web delivery of adaptive and interactive language tutoring [Electronic version]. *International Journal of Artificial Intelligence in Education, 12*, 310–325.

Hilbert, M., Lobin, H., Bärenfänger, M., Lüngen, H., & Puskás, C. (2006). A text-technological approach to automatic discourse analysis of complex texts. In M. Butt (Ed.), *Proceedings of KONVENS 2006* (pp. 52–55). Konstanz, Germany: Universität Konstanz.

Holland, V. M. (1995). The case for intelligent CALL. In V. M. Holland, J. D. Kaplan & M. R. Sams (Eds.), *Mobile learning: Intelligent language tutors: Theory shaping technology* (pp. 7–16). Mahwah, NJ: Lawrence Erlbaum.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus. Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management, 12*, 119–125.

Kiernan, P. J., & Aizawa, K. (2004). Cell phones in task based learning: Are cell phones useful language learning tools? *ReCALL, 16*, 71–84.

Kraif, O., & Ponton, C. (2007). *Du bruit, du silence et des ambiguïtés: Que faire du TAL pour l'apprentissage des langues?* [Noise, silence and ambiguities: What can NLP do for language learning?]. Unpublished manuscript. Retrieved May, 28, 2008 from http://w3.u-grenoble3.fr/ponton/perso/docs/TALN07.pdf

Kraif, O., & Tutin, A. (in press). Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. In Natalie Kübler (Ed.), *Proceedings of TaLC2006, 7ème Conférence Teaching and Language Corpora*. Coll. Etudes contrastives, Peter Lang, Bruxelles.

Kukulska-Hulme, A. (2007). Mobile usability in educational contexts – what have we learnt? *International Review of Research in Open and Distance Learning, 8*, 1–16.

Kukulska-Hulme, A., & Shield, L. (2004). Usability and pedagogical design: Are language learning websites special? In L. Cantoni & C. McLoughlin (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004* (pp. 4235–4242). Chesapeake, VA: AACE.

L'haire, S. (2004). *Vers un feedback plus intelligent. Les enseignements du projet Freetext* [Towards more intelligent feedback: What have we learned from the Freetext Project]. Unpublished manuscript. Retrieved May, 28, 2008, from http://w3.u-grenoble3.fr/lidilem/talal/actes/JourneeTALAL-041022-lhaire.pdf

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing, 8*, 275–281.

Lomicka, L. (1998). To gloss or not to gloss: An investigation of reading comprehension online. *Language Learning and Technology*, *1*, 41–50.

Metcalf, V., & Meurers, D. (2006). *Towards a treatment of word order errors: When to use deep processing and when not to*. Unpublished manuscript. Retrieved May, 28, 2008, from http://www.ling.ohio-state.edu/icall/handouts/calico06-metcalf-meurers.pdf

Milton, J. (2002). *Literature review in languages, technology and learning.* Swansea, United Kingdom: University of Wales Swansea, Centre for Applied Language Studies.

Mitkov, R. (Ed.). (2003). *Handbook of Computational Linguistics.* Oxford: United Kingdom: Oxford University Press.

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor & T. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 239–250). Amsterdam: Rodopi.

Nerbonne, J. (2003). Computer-assisted language learning and natural language processing. In R. Mitkov (Ed.), *Handbook of computational linguistics* (pp. 670–698). Oxford, United Kingdom: Oxford University Press.

Nerbonne, J., Dokter, D., & Smit, P. (1998). Morphological processing and computer-assisted language learning. *Computer-Assisted Language Learning, 11*, 421–437.

Römer, U. (2008), Corpora and language teaching. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. (Vol. 1) (pp. 112–130). Berlin, Germany: Walter de Gruyter.

Scott, M. (2004). *WordSmith tools 4*. Oxford, United Kingdom: Oxford University Press.

Selva, T. (2002). Génération automatique d'exercices contextuels de vocabulaire [Automatic generation of contextual vocabulary exercises]. In *Proceedings of TALN 2002* (pp. 185–194). Nancy, France: Université de Nancy. Retrieved May, 28, 2008, from http://www.loria.fr/projets/JEP-TALN/actes/TALN/articles/TALN17.pdf

Tribble, C., & Barlow, M. (Eds.). (2001). Using corpora in language teaching and learning [Special issue]. *Language Learning and Technology, 5*(3).

Vandeventer, A. (2001). Creating a grammar checker for CALL by constraint relaxation: A feasibility study. *ReCALL, 13*, 110–120.

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning and Technology, 10*, 85–101.