ReCALL **19** (3): 252-268 © 2007 Cambridge University Press
doi:10.1017/S0958344007000237 Printed in the United Kingdom

Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness[1]

SYLVIANE GRANGER
Université catholique de Louvain, Centre for English Corpus Linguistics, Place Blaise Pascal 1, B-1348 Louvain-la-Neuve, Belgium
(*email: sylviane.granger@uclouvain.be*)

OLIVIER KRAIF, CLAUDE PONTON, GEORGES ANTONIADIS AND VIRGINIE ZAMPA
LIDILEM, Université Stendhal Grenoble3, BP-25 – 38040 Grenoble Cedex 9 France
(*email: {Olivier.Kraif, Claude.Ponton, Georges.Antoniadis, Virgine.Zampa}@u-grenoble3.fr*)

Abstract

Learner corpora, electronic collections of spoken or written data from foreign language learners, offer unparalleled access to many hitherto uncovered aspects of learner language, particularly in their error-tagged format. This article aims to demonstrate the role that the learner corpus can play in CALL, particularly when used in conjunction with web-based interfaces which provide flexible access to error-tagged corpora that have been enhanced with simple NLP techniques such as POS-tagging or lemmatization and linked to a wide range of learner and task variables such as mother tongue background or activity type. This new resource is of interest to three main types of users: teachers wishing to prepare pedagogical materials that target learners' attested difficulties; learners themselves for editing or language awareness purposes and NLP researchers, for whom it serves as a benchmark for testing automatic error detection systems.

Keywords: Learner corpora, error tagging, error detection, NLP, web interface, French

1 Introduction

Right from the earliest days of corpus linguistics, researchers have highlighted the relevance and importance of corpus-based research for foreign language learning and

1. The research reported in this article is part of a wider project on Integrated Digital Language Learning (IDILL) carried out within the framework of the EU-funded network of excellence Kaleidoscope dedicated to research in the field of technology-enhanced learning: http://www.noe-kaleidoscope.org /pub/

teaching. As far back as 1967, Francis wrote an article analyzing the implications of the Brown corpus for the teaching of English to speakers of other languages. The link between corpus linguistics and teaching has grown steadily ever since and corpora are now a major component of the pedagogical scene. The major contribution of corpus-based analysis is that it gives researchers access to detailed qualitative and quantitative information on native speakers' typical patterns of use, making it possible to produce more effective pedagogical resources.

However, native corpora give no indication of what is difficult for learners. As pointed out by Nesselhauf (2004: 125) "For language teaching [...] it is not only essential to know what native speakers typically say, but also what the typical difficulties of the learners of a certain language, or rather of certain groups of learners of this language, are". For this, we need learner corpora, electronic collections of texts produced by foreign language learners. Numerous studies[2] have shown that by applying corpus linguistic techniques to learner corpora, it is possible to identify the features that characterize learner language, distinguishing between features shared by several learner populations and those that are typical of one particular learner group, thereby gaining a clearer picture of the distance that still separates the learner's language from the targeted 'norm'. While a wide range of L2-specific patterns of overuse, underuse and misuse can be uncovered from raw, i.e. unannotated, learner corpora, the value of learner corpora increases exponentially when they are enriched with error annotations. Language practitioners can use the resource to produce pedagogical tools – dictionaries, grammars, textbooks, CALL programs – that address learners' attested difficulties. In actual fact, learner-corpus-informed pedagogical tools are still relatively scarce, but there is clearly enormous potential. One particularly promising avenue for future progress in the field comes from the field of Natural Language Processing (NLP). Combining use of NLP techniques such as POS-tagging, parsing, lemmatization and other processing and error-tagged corpora, makes it possible to design a wide range of resources, ranging from highly sophisticated automatic error detection and correction systems to simpler, but potentially very useful web-based interfaces to authentic errors and their corrections.

The aim of this article is to highlight the role that learner corpora can play in CALL. We take stock of research in the field and advocate a realistic approach that reconciles the current capabilities of NLP tools and the realities of foreign language teaching. In section 2 we introduce learner corpora and highlight their contribution to foreign language learning and teaching. Section 3 focuses on error annotation and the contribution of error-tagged learner corpora to CALL. In section 4 we highlight the contribution of NLP techniques to the exploitation of learner corpora in two types of application: NLP-based error detection and feedback and NLP-based error analysis interface. Section 5 features a web-based error interface for French as a Foreign Language, the Example extractor engine for language teaching. In the concluding section, we outline some avenues for future research.

2. An extended learner corpus bibliography is stored on the following website: http://cecl.fltr.ucl.ac.be/learner%20corpus%20bibliography.html

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

2 Learner corpora

Learner corpora are electronic collections of foreign or second language learner texts collected on the basis of strict design criteria. The word 'texts' underlines an important distinguishing feature of learner corpus data, i.e. the fact that they contain continuous discourse and not a set of decontextualized sentences. The term 'error corpus', which is sometimes used to refer to learner corpora, is therefore clearly a misnomer as learner corpora, like all corpora, contain productions in their entirety, including both correct and incorrect uses. An equally important part of the definition is the requirement for strict design criteria. A haphazard collection of learner data of unknown provenance cannot claim to be a learner corpus. As learner language is extremely heterogeneous, a learner corpus will only be useful – be it for theoretical or applied purposes – if it is properly coded for a wide range of variables, both learner variables such as age, gender, mother tongue background – and text variables such as medium, task type or time limitation.

Launched simultaneously but completely independently in academic and commercial circles in the late 1980s, learner corpus collection has resulted in a wide range of corpora differing in content and size. At first limited to English (for a survey of learner corpora of English, see Pravec, 2002), it is now beginning to include a much wider range of languages: French (Myles & Mitchell, 2004), Swedish (Hammarberg, 1999), Norwegian (Tenfjord, Meurer & Hofland, 2004) and German (Lüdeling, Maik, Kroymann & Adolphs, 2005), to name but a few.

Historically, one of the factors that has acted as a brake to learner corpus research is the difficulty of collecting and computerizing the data. This difficulty is reflected, for example, in the 12-year gap between the start of the collection of the *International Corpus of Learner English* and the release of the data in CD-ROM format (Granger, Dagneaux & Meunier, 2002). The growing use of Computer-Mediated Communication (CMC) and Learning Management Systems (LMS) in foreign language teaching should in principle allow researchers to collect learner data in electronic format as part of their normal teaching activities, thereby greatly speeding up learner corpus collection. In practice however, this approach is still more the exception than the rule, for a number of reasons. For one thing, many teacher-training programmes do not contain a corpus-oriented (let alone a learner-corpus-oriented) module and teachers are therefore largely unaware of the benefit they can derive from collecting and using corpora. In addition, many technology-enhanced tools, in particular LMSs, do not provide teachers with easy ways of retrieving learners' production data. This said, there are signs that this is rapidly changing. Belz and Vyatkina (2005) and Belz (forthcoming) report on a bilingual combined native/learner corpus of English and German collected within the framework of a telecollaborative programme involving email and synchronous chat. Suzuki, Jung, Watanabe, Min and Yoshihara (2004) saved data posted on a bulletin board system by Japanese and Korean learners of English as a corpus in the web server and carried out quantitative and qualitative analyses of the data. Wible, Kuo, Chien, Liu & Tsao (2001) describe a language learning environment called *IWiLL*,[3] which has been designed to serve the needs of a second language writing course. With the students' permission, the essays sent to the teacher over the Internet are stored together with the teachers' error

3. IWiLL stands for Intelligent Web-based Interactive Language Learning.

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

annotations in a database, thereby producing, as a by-product of the system, a large learner corpus, the *Taiwan Learners' Corpus*, which continues to grow as the writing environment is used by the students and their teachers. However, learner corpus analysis provides a much more precise picture of learner language than has ever been available before. Contrastive interlanguage analysis, a method which consists in carrying out quantitative and qualitative comparisons of different learner populations and/or comparisons of learner language and the targeted language, is a powerful heuristic that has revealed patterns of overuse, underuse and misuse in learner language which are considerably extending our understanding of interlanguage.

Learner corpus research is still a very young field and it should therefore come as no surprise, especially in view of the high level of analytical work involved, that only a small number of learner-corpus-informed pedagogical tools are currently on the market. The field where learner corpus data have had most impact so far, at least for English, is that of pedagogical lexicography. Three learners' dictionaries – the *Longman Dictionary of Contemporary English* (2005), the *Cambridge Advanced Learner's Dictionary* (2003) and the *Macmillan English Dictionary for Advanced Learners* (2007) – have drawn on large bodies of learner corpus data and on the basis of their findings, incorporated very useful warnings in the form of 'common learner error' sections or 'get it right' boxes to draw learners' attention to common mistakes. Other fields, such as that of EFL writing, also stand to gain much from learner corpus findings, as evidenced by recent research in the field of English for Academic Purposes (Gilquin, Granger & Paquot, forthcoming). As regards courseware, the impact of learner corpus studies appears to have been greater on electronic tools than traditional textbooks (cf. section 3.2). This may well be due to the high flexibility of the electronic medium, which enables academics to bypass publishers and produce their own tools directly.

3 Learner corpora and CALL

3.1 Error tagging

While improving learners' accuracy may not be the only or the most important teaching objective, it is a feature – albeit to varying degrees – of practically all language courses. If the researcher has access to a learner corpus, s/he can analyze learner errors using one of the following two methods:

- a corpus-based method, involving looking for words, phrases or structures that s/he knows or suspects to be problematic. In this case, the corpus is used to verify intuitions and acquire a more accurate understanding of the error-prone items using concordancing techniques which highlight the specific patterns of misuse. Lenko-Szymanska's (2004) study of demonstratives is a good example of this approach.
- a corpus-driven method, which consists in carrying out automatic comparisons of learner and native corpora to identify learner-specific features. In this case, the researcher has no a priori idea of the error-prone items; corpus comparison is used as a heuristic to uncover them. Granger and Rayson's (1998) automatic profiling of learner texts uncovers many instances of significant over- or

underuse, many of which turn out to be due to erroneous or stylistically inappropriate use.

Although undeniably useful, these two methods have severe limitations, the first because intuition only gives access to the errors that are most salient in teachers' minds and the second because misuse does not always result in over- or underuse. As a result, error tagging, which consists in annotating learner errors with a standardized system of error tags, is growing increasingly popular. Although it is a highly time consuming process, beset with many difficulties, it has proven to be the only reliable way of gaining access to an exhaustive and consistently coded catalogue of the errors produced by a given learner population.

Error tagging systems differ in the types of error taxonomy used, the granularity of the tagsets, the formats in which they are encoded and the possibility of introducing one or more levels of analysis (cf. Lüdeling, Maik, Kroymann & Adolphs, 2005 for a description of a multi-layer annotation system). Milton and Chowdhury's (1994) system was designed to tag a corpus of English writing produced by Hong Kong learners, while Izumi, Uchimoto and Isahara's (2004) system was designed to error tag a corpus of speech from Japanese learners of English. The system described by Nicholls (2003), on the other hand, was used to annotate the Cambridge Learners' Corpus, a large EFL corpus containing data from a wide range of learner populations. Granger (2003) describes an error system initially designed for annotating the *International Corpus of Learner English* (Dagneaux, Denness & Granger, 1998) and later adapted for French within the framework of the *FreeText* project.[4] The three-tiered system, illustrated in Figure 1, specifies the error domain (form, grammar, lexis, etc.), error category (gender, number, tense, etc.) and word category (adjective, noun, verb, etc.) of each erroneous item. Corrections are also included (for a more detailed description, see Granger, 2003). In Figure 1 two grammatical errors <G> have been tagged. Both have to do with the error category of gender <GEN> but the first affects an indefinite determiner <DEI> (masculine *tous* corrected as feminine *toutes*) and the second a past participle verbal form <VSP> (masculine *décidés* corrected as feminine *décidées*).

3.2 Remedial CALL programs

Learner corpus data – usually error-tagged – have served as the basis for a number of remedial CALL programs. The pioneer of learner-corpus-informed CALL programs is Milton (1998), who developed a writing kit called *WordPilot*. This program combines remedial exercises targeting Hong Kong learners' attested difficulties and a writing aid

<G><GEN><DEI> #Toutes$ Tous </DEI></GEN></G> les choses concernant les employés seront <G><GEN><VSP> #décidées$ décidés </VSP></GEN></G> ici

Fig. 1. FRIDA: Sample of error-tagged text.

4. FreeText is an EU-funded project carried out within the framework of the Fifth Framework Programme (IST-1999-13093).

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

tool which helps learners to select appropriate wording by accessing native corpora of specific text types. Cowan, Choi and Kim's (2003) *ESL Tutor* program is an error correction courseware tool that contains units targeting persistent grammatical errors produced by Korean ESL students. Like *WordPilot*, the *ESL Tutor* is L1-specific, addressing errors that are clearly transfer-related. Unlike *WordPilot*, however, the learner corpus data used were not error-tagged and the authors have had to rely on the SLA literature to identify typical Korean learners' errors supplemented with incremental searches in the raw learner corpus. The program, in which structured negative feedback and the promotion of noticing play a major role, addresses persistent errors like the overpassivization of unaccusative verbs (**the accident was occurred early in the morning*) and includes error detection and correction exercises. One of the most promising but also most challenging features of the program is the variable feedback that learners receive throughout the session. Chuang and Nesi (2004, in press) have carried out a corpus-based error analysis of academic texts written by Chinese students studying in the medium of English and have designed on that basis a remedial online self-study package called *GrammarTalk* which targets high frequency errors such as article errors.

4 Learner corpora and NLP-based CALL

The CALL programs described above are innovative in that they are informed by error-tagged learner corpus data but they are otherwise fairly traditional. Once the errors have been identified, they are used as a basis for standard CALL exercises, such as multiple-choice, jumbled-sentence, matching/ordering or gap-fill. The link between the error-tagged material and the pedagogical resources is thus indirect. In this section we describe two new sets of applications – automatic error detection and correction systems and web-based error interfaces – which make a much more direct use of error annotations and involve a range of Natural Language Processing (NLP) tools.

4.1 NLP-based error detection and feedback

A variety of NLP-based techniques involving parsing with relaxed constraints (Vandeventer 2001), graded constraint parsing (Menzel & Shröder, 1999) or the implementation of 'mal rules' (i.e. an error grammar as in Foster & Vogel, 2004), have been designed to process learner language in the hope of "having a machine serve as a tireless language model, interlocutor and error correction or grading authority" (Dodigovic, 2005: 97). Implementing Intelligent CALL (ICALL) systems into CALL would make it possible to incorporate free production exercises alongside the traditional tightly controlled exercises such as fill-in-the-blank. In an ideal situation, the learners would produce free answers and the system would detect the errors and provide appropriate feedback. However, as pointed out by Antoniadis, Echinard, Kraif, Lebarbé, Loiseau and Ponton (2004) "the existing models are both silent and noisy at the same time: some errors are not detected, and correct expressions are wrongly pointed out as errors". Indeed, there are several studies that support this observation. Izumi, Uchimoto and Isahara (2004: 125) report a recall rate of c. 50% and a precision rate of c. 60% for the detection of article errors produced by Japanese learners of English. Looking back at

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

the *FreeText* project, L'haire (2004) expresses his disappointment at the relatively weak performance of the error tagging system and pinpoints overflagging as one of the major problems. This lack of reliability affects most classical NLP applications – machine translation, spell checking and grammar correction, dialogue generation, speech synthesis and recognition – and makes their integration into Foreign Language Learning tools highly problematic. As current parsers often fail to analyze correct utterances, it is hardly surprising that they should have difficulty handling learner productions which contain orthographic, lexical and grammatical errors. As Tschichold (2003) points out, "there are today no viable and comprehensive computational grammars capable of dealing with the error types found in learner language".

In view of the limitations of current-day technology, one may wonder whether the integration of NLP techniques into CALL is not premature. At any rate, researchers are divided on the issue. While recognizing the limitations of the *FreeText* error diagnostic system, Vandeventer (2001) considers that "the present, imperfect state of the technology should not be a hindrance". She suggests using the tools in their present state and warning learners of their limitations. Other researchers, however, advocate a more cautious attitude. For Armaral and Meurers (2006), "processing completely free production input, allowing any number and type of errors is not tractable". The evaluation of *FreeText*'s error diagnostic system by language teachers has demonstrated that, while they were in principle very much in favour of integrating such tools into CALL programs, they were nevertheless very reluctant to use a program that overlooked mistakes in students' writing and – even worse – that marked correct text as incorrect (Cosme, Delghust, Gouverneur, Granger & Husquet, 2003).

One realistic and promising way forward is to choose the processing method in function of the targeted error and the activity design. Not all errors require deep processing techniques. As demonstrated convincingly by Metcalf and Meurers (2006), different types of word order errors call for different processing: those involving phrasal verbs (*they give up it*) can be handled successfully by means of instance-based regular expression matching while errors involving adverbs (*it brings rarely such connotations*) require more sophisticated parsing algorithms. To determine which errors fall within the scope of a specific technique, a corpus containing errors may be used to carry out a precise evaluation of the error detection system (Foster, 2004). In order to then adapt the detection system to take into account the targeted errors, it is essential to arrive at a precise description of the errors. A 'positive' view of the error may lead to a systematisation, an error grammar which "has the advantage of providing a linguistic model of ungrammaticality - this means that ill-formed sentences can be diagnosed as such, instead of being viewed merely as sentences occurring with low frequency" (Foster & Vogel, 2004:269). But, even a 'negative' conception of error, as in constraint retraction systems, can benefit from a corpus-based error description. Menzel and Schröder (1999) present an interesting architecture where grammatical constraints are graded, in order to "determine how serious one considers a constraint violation, thus yielding constraints of different strength" (ibid:23). Only a learner corpus can help to determine to what extent a constraint violation is expectable, common, or exceptional.

The conclusion that can be drawn at this stage is that automatic error detection and feedback of fully unconstrained learner output is over-ambitious in view of the current state of NLP technology and researchers should target more constrained environments.

Whatever the technique used, a thorough knowledge of errors and their contexts of use is essential. In the following section, we describe web-based tools that can give researchers, teachers and learners easy and versatile access to authentic errors and their corrections.

4.2 NLP-based error analysis interface

Error-tagged learner corpora, though undeniably useful, are not particularly easy to use. By converting them into proper databases and adding a web-based interface, they become user-friendly browsing and exploration tools. In this section, we give a brief description of two such web-based interfaces, one aimed at Second Language Acquisition (SLA) researchers, the other aimed directly at learners. In the next section, we provide a detailed description of *eXXelant*, a web-based error interface aimed at language practitioners and researchers alike. The three applications have a lot in common, notably the fact that simple NLP techniques, such as tokenization, POS-tagging and lemmatization, are used to supplement error tags with other types of linguistic annotation.

The *ASK* corpus is a corpus of Norwegian as a Second Language that has been compiled in order to facilitate empirical studies on the acquisition of Norwegian and SLA studies in general. It aims to allow researchers to "test hypotheses generated by previous studies in Norwegian as a second language" and "generate new hypotheses of lexical, grammatical and textual features of SLA, as well as hypotheses on individual and external factors influencing the language acquisition process in more general terms" (Tenfjord, Hagen & Johansen, in press). All the texts have been error-tagged and POS-tagged and are supplied with a range of personal data (mother tongue, age, duration of residence in Norway, etc.). Beside the learner texts, the database also contains grammatically tagged, fully-corrected versions of the texts as a parallel corpus. A very flexible query system coupled with a web search interface allows searches for combinations of words, error categories, grammatical annotation and personal data and displays the results in a variety of formats: "as KWIC concordances, as pairs of matching sentences from the original and the corrected corpus [...] and as sentences visualized using user-definable (XSLT) style sheets that highlight different aspects of the text" (Tenfjord, Meurer & Hofland, 2004).

Unlike the *ASK* corpus, the *iWRITE* system designed by Hegelheimer and Fisher (2006) directly targets the learners. It aims to "raise learners' grammatical awareness, encourage learner autonomy, and help learners prepare for editing or peer editing." The system is designed to prepare activities based on interactions in the classroom, like peer editing, where linguistic awareness is reinforced through learners' collaboration and teacher questions. Learners' written essays, encoded in XML files including error annotation (error category, description, and corrected form) can be accessed through a web interface and the corpora are directly browsed by the learners themselves. They can have access to all the errors in a selected category (and the contexts in which they appear); they can also display a single essay (selected on the basis of native country, essay topic, and TOEFL scores), in unmarked form, or with a particular error category highlighted. In the practice section of the system, they can download worksheets in Word format and try to correct the highlighted errors. The system is consistent with an

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

interactionist approach (Chapelle, 1998) where software use may become a vector for learner/teacher/form interactions: "Noticing linguistic input is viewed as a prerequisite for acquisition […], and noticing is more likely to occur during interaction. Hence, software features that enhance noticing in general and that help the learner to focus on form […] are viewed as beneficial" (Hegelheimer & Fisher, 2006: 261). In a subsequent study, Hegelheimer (2006) reports the results of a preliminary empirical study aimed at assessing the utility of the iWRITE system. The study shows that "overall positive findings were obtained concerning learners' attitudes and grammatical awareness and there was some indication of improvement of grammatical accuracy" (*ibid*: 9). The author concludes that iWRITE holds promise for intermediate-level learners but calls for additional studies with larger populations to corroborate the results.

5 EXample eXtractor Engine for LAnguage Teaching (eXXelant)

eXXelant is a web-based error interface which results from the collaboration between the University of Louvain (Centre for English Corpus Linguistics) and the University of Grenoble LIDILEM Laboratory. The system is based on two types of resources:

- A large error-annotated corpus of French as a Foreign Language, the French *Interlanguage Dababase (FRIDA)*, collected and error-tagged at Louvain within the framework of the *FreeText* project;
- An NLP-enhanced version of the corpus, *FRIDA-bis*;
- A web interface to extract significant examples and statistics.

The resulting tool provides answers to the following types of questions: Does this kind of error occur in the corpus? How frequently? In what contexts? Produced by which learner population?

5.1 From FRIDA to FRIDA-bis

The *FRIDA* corpus contains 500,000 words of texts written by learners of French as a Foreign Language. It is made up of three subcorpora of similar size which contain data from English-speaking learners, Dutch-speaking learners and learners from mixed mother tongue backgrounds. Two thirds of the corpus, some 300,000 words, have been fully error-tagged with the three-tiered system described above and a wide range of error statistics have been extracted from the corpus. By way of illustration, Table 1 gives the breakdown of the corpus in terms of the nine error domains: grammar (G), Form (F), Lexis (L), Syntax (X), Punctuation (Q), Register (R), Style (Y), Morphology (M) and Typo (Z).[5] It appears from the table that two domains, grammar and form, account for 50% of all errors in the corpus (for more information on the corpus and how it was used

5. The system distinguishes between sentence grammar errors <G> and morphological errors <M>, which are further subdivided into inflectional and derivational errors. In some error tagging systems the two categories are combined under the general umbrella term of grammar. Note that in our system the category of form <F> only includes spelling errors and formal errors resulting from homonymy.

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

Table 1 *Breakdown of error domains in FRIDA*

Tag	N. of occurrences	%
G	11 779	25.28
F	11 452	24.64
L	7 198	15.51
X	7 061	15.21
Q	5 707	12.29
R	1 402	3.02
Y	862	1.85
M	784	1.68
Z	155	0.33

in the *FreeText* project, see Granger, Vandeventer and Hamel, 2001 and Granger, 2003).

FRIDA-bis is an NLP-enhanced XML version of the *FRIDA* corpus which has undergone the following three types of processing:

- Data clean-up and conversion into XML format;
- Automatic meta-tagging of each learner text (error density, text length);
- Automatic POS-tagging of all items in the database (text, errors and corrections) with the TreeTagger, an open-source POS-tagger.[6]

The resulting corpus is made up of 764 texts and contains 9,466 sentences, 20,474 errors and 179,642 words. The difference between the two formats – *FRIDA* and *FRIDA-bis* – appears clearly from a comparison of Figures 1 and 2. In Figure 2 each token (word, punctuation mark, number, etc.) contained in the markup <tok> bears additional information such as lemma (*base* attribute), part-of-speech (*ctag*) and morphosyntactic features (*msd*). The <ERR> markup indicates an error, where <INI> contains the original text and <COR> the corrected version. Error type is encoded in the <ERR> attributes where DOM indicates the error domain, CER the error category, and CGR the POS category of the error-prone structure.

5.2 Search interface

If researchers and teachers are to be able to explore the corpus extensively and effectively, the corpus needs to be equipped with an efficient query system which enables users to search on the basis of a wide range of criteria. To this end, the corpus was indexed in a relational database and the following search criteria implemented:

- Learners' mother tongue

6. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

```
<ERR IDE="7" DOM1="G" CER1="GEN" CGR1="DEI">
<INI><tok id="t30" orth="Tous" base="tout" ctag="pro" msd="indef">Tous</tok></INI>
<COR><tok id="t31" orth="Toutes" base="unknown" ctag="nom" msd="">Toutes</tok></COR>
</ERR>
<tok id="t32" orth="les" base="le" ctag="det" msd="artic">les</tok>
<tok id="t33" orth="choses" base="chose" ctag="nom" msd="femin">choses</tok>
<tok id="t34" orth="concernant" base="concerner" ctag="ver" msd="ppres">concernant</tok>
<tok id="t35" orth="les" base="le" ctag="det" msd="artic">les</tok>
<tok id="t36" orth="employés" base="employé" ctag="nom" msd="mascu">employés</tok>
<tok id="t37" orth="seront" base="être" ctag="ver" msd="ifutu pluri">seront</tok>
<ERR IDE="8" DOM1="G" CER1="GEN" CGR1="VSP">
<INI><tok id="t38" orth="décidés" base="décider" ctag="ver" msd="ppass mascu">décidés</tok></INI>
<COR><tok id="t39" orth="décidées" base="décider" ctag="ver" msd="ppass">décidées</tok></COR>
</ERR>
<tok id="t40" orth="ici" base="ici" ctag="adv" msd="">ici</tok>
```

Fig. 2. Sample from the FRIDA-bis corpus.

- Error density (error count for 100 words)
- Text length (number of words)
- Search item (word form, lemma, grammatical category, error domain/category, etc.)
- Correction (word form, lemma, grammatical category, etc.)
- Right- and left-hand context (word form, lemma, grammatical category, etc.)

Figure 3 illustrates the search for all past participle (i.e. "catégorie=verbe, trait=participle passé") errors (i.e. "erreur=oui") encountered in the corpus with the verb *avoir* (i.e.

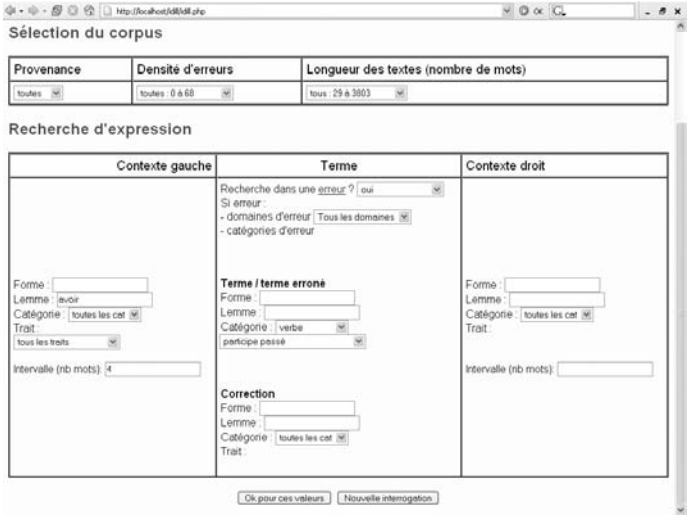
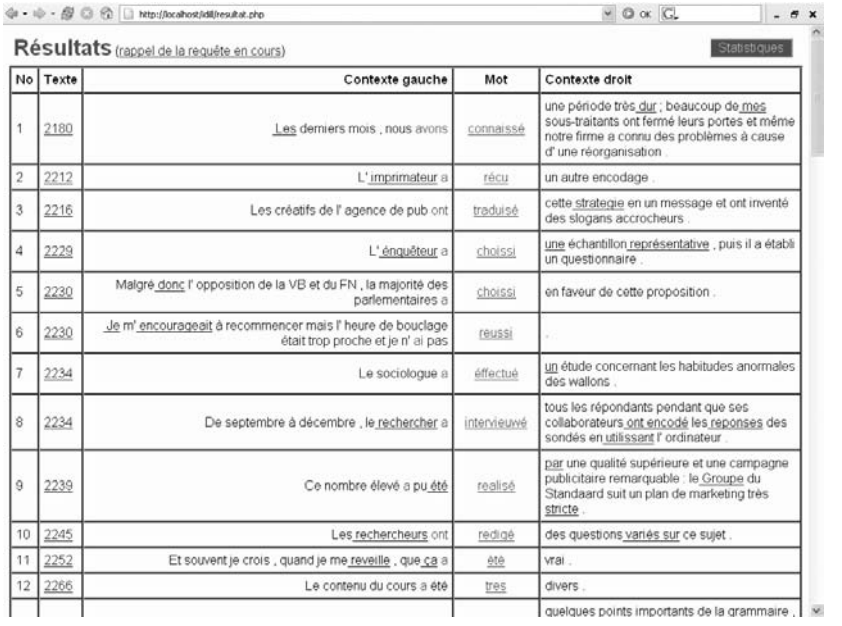


Fig. 3. The search for the past participle.

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press



Résultats (rappel de la requête en cours) Statistiques

No	Texte	Contexte gauche	Mot	Contexte droit
1	2180	Les derniers mois , nous avons	connaissé	une période très dur ; beaucoup de mes sous-traitants ont fermé leurs portes et même notre firme a connu des problèmes à cause d' une réorganisation .
2	2212	L' imprimateur a	récu	un autre encodage .
3	2216	Les créatifs de l' agence de pub ont	traduisé	cette strategie en un message et ont inventé des slogans accrocheurs .
4	2229	L' énquêteur a	choissi	une échantillon représentative , puis il a établi un questionnaire .
5	2230	Malgré donc l' opposition de la VB et du FN , la majorité des parlementaires a	choissi	en faveur de cette proposition .
6	2230	Je m' encourageait à recommencer mais l' heure de bouclage était trop proche et je n' ai pas	reussi	.
7	2234	Le sociologue a	éffectué	un étude concernant les habitudes anormales des wallons .
8	2234	De septembre à décembre , le rechercher a	interviewé	tous les répondants pendant que ses collaborateurs ont encodé les reponses des sondés en utilisant l' ordinateur .
9	2239	Ce nombre élevé a pu été	realisé	par une qualité supérieure et une campagne publicitaire remarquable : le Groupe du Standaard suit un plan de marketing très stricte .
10	2245	Les rechercheurs ont	redigé	des questions variés sur ce sujet .
11	2252	Et souvent je crois , quand je me reveille , que ça a	été	vrai .
12	2266	Le contenu du cours a été	tres	divers .
				quelques points importants de la grammaire ,

Fig. 4. The results of the search from Figure 3.

"lemme=avoir") as left-hand context. The results of the search are displayed in Figure 4.

eXXelant has been especially designed with the following two categories of users in mind: language practitioners (teachers and materials designers) and researchers working in the fields of SLA or NLP. Using the interface, language practitioners will be able to gain a better understanding of learners' errors and design tailor-made pedagogical materials that target those difficulties. SLA researchers will have a solid empirical basis on which to test existing SLA theories and elaborate new ones. NLP researchers will have an ideal benchmark for fine-tuning and assessing their error detection and diagnosis systems. By carefully parametrizing errors and their contexts, they will be able to identify which phenomena are worth processing for which categories of learners and which are beyond the capabilities of current-day technology. Although *eXXelant* was not initially designed for direct use by learners, it has many features in common with Hegelheimer and Fisher's *iWRITE* system and could easily be adapted to perform similar activities.

5.3 eXXelant searches illustrated

The interface allows for a wide range of searches. Users can start from an error domain or category and extract all the examples that display the targeted errors. They can also refine their search, for example by restricting it to a particular lemma or context. To derive maximum benefit from the error interface, users need to get used to the error tagging system, which they can easily do by reading the detailed error tagging manual (Cassart, Dagneaux, Granger, Husquet, Verhulst & Watrin, 2002). However, users who do not want to invest time in coming to grips with the error typology can still derive

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

Table 2 *Errors involving the possessive determiner leur(s)*

Case number	Query criteria		Number of instances found in each quantile				Total
	Erroneous form	Corrected form	Low (0-12)	Medium (12-17)	Medium + (17-24)	High (24-68)	
1	lemma *leur*	lemma *son*	2	4	0	4	10
2	lemma *son*	lemma *leur*	10	10	8	16	44
3	form *leur*	form *leurs*	3	3	5	3	14
4	form *leurs*	form *leur*	2	2	1	0	5
	Total		17	19	14	23	73

great benefit from using the interface. For example, a search on the lemma *avoir* followed by a past participle will give access to all uses of this verbal form, both correct and incorrect. As error tagging is a highly time-consuming, hence costly enterprise, error-tagged learner corpora are few and far between. It is therefore useful to supplement the error-tagged data with raw data or lightly annotated data, for example with only the corrected forms inserted.

One of the great advantages of *eXXelant* is that it gives access to both the erroneous form and its correction. A search for all erroneous instances of the form *qui* generates 66 occurrences (out of 1341 occurrences of the word). A quick look at the erroneous forms suggests that a number of the errors result from a confusion between *qu'il* and *qui*, as illustrated in example (1).[6]

1. Un autre problème ***qui*** ne faudrait pas sous-estimer est le sous-effectif du fisc. (correct form: *qu'il*)
2. Je sais ***qui*** parler anglais est necessaire à notre époque (correct form: *que*)
3. Le langue ***qui*** je vais utiliser dans mon vie proffesionel (correct form: *que*)

To confirm this hunch, the user does not need to go through the 66 concordance lines. S/he can simply restrict the search by searching for the lemma *il* in the correction. For cases like (2) where the error is due to a confusion between the relative pronoun *qui* and the conjunction of subordination *que,* the search can be restricted to cases where the correction includes the conjunction *que*. To extract instances such as (3) which display a confusion between the relative pronouns *qui* and *que*, it is possible to restrict the search to erroneous instances of *qui* corrected as *que*, which have a noun in a 3-word window to the left of *qui*.

As the interface also contains learner and task variables, this leaves room for interesting comparisons between learners and learner groups. As regards error density, the learner texts have been divided into four groups of equal importance (quantiles)

6. Learner texts have not been normalized and may therefore contain several errors.

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

Table 3 *Breakdown of son/leur confusions according to country of origin*

Learner country	son/leur confusions		% learner texts (words)
	N.	%	
UK	3	7%	9%
NL	12	27%	38%
BE	11	25%	37%
BG	1	2%	7%
CH	1	2%	2%
JP	16	36%	2%

according to the number of errors per 100 words. This measure can serve as a rough indicator of proficiency. Table 2 shows the distribution of the errors involving the possessive determiner *leur*(s).

As the figures are quite low, they should be interpreted with caution. However, it is interesting to note that instead of the expected increase in the number of errors from low to high error density, each quantile displays approximately the same number of errors, which indicates that this form remains problematic even at more advanced proficiency levels.[7] A qualitative analysis of the examples shows that this is indeed the case but that the errors are of a different nature: they tend to be clear-cut violations of possessor number or lack of agreement with the head noun in the high error density texts (cf. examples 4 and 5) and more complex cases where the possessor is semantically plural, as in (6), in the low-density texts.

4. La Gran Place, avec ***leur*** Hotel de Ville (…)(correct form: *son*)
5. j'ai vaux a Gent, une ville tres antique, avec leur eglisses, e ***leur*** canals (correct form: *ses*)
6. Il est vrai que chaque pays, dans un futur union d' Europe perdra certains aspects de ***leur*** souveraineté, mais ils se dirigeront toujours (possessor: *chaque pays*; correct form: *sa*)

Figures for case number 2 in Table 2 show that the semantic link with the possessor number is problematic for most students and that the singular form *son* plays a canonical role compared to *leur* (44 *son*/*leur* confusions vs. 10 *leur*/*son* confusions). Using the learners' country of origin as search criterion, it is possible to show the breakdown of this error across the different learner populations. As appears from Table 3, the breakdown shows that this error is overrepresented in the Japanese subcorpus which makes up a mere 2% of the learner texts and yet contains 36% of the errors.

7. Further research is needed, however, as low error density texts are usually longer than high error density texts and this has an influence on the total number of errors.

https://doi.org/10.1017/S0958344007000237 Published online by Cambridge University Press

6 Conclusion and avenues for future research

Error-tagged learner corpora hold great potential for CALL. Enriched with the help of simple NLP techniques such as tokenization, POS-tagging and lemmatization and a flexible web interface, they can be directly used by learners for editing or language awareness purposes or tapped by teachers to design activities that target learners' attested difficulties, in particular language awareness and focus-on-form activities (cf. Amaral, Metcalf & Meurers, 2006). For researchers developing error diagnostic systems, such annotated corpora are an ideal benchmark for assessing which phenomena are worth pursuing and which are beyond the scope of the current state-of-the-art. They can use them to determine what are typical errors and identify their contexts of occurrence, and on the basis of this, build efficient 'error grammars' capable both of detection and diagnostics, taking account of criteria such as the learner's mother tongue. These data are crucial in explaining why errors have been made and how to correct them. The descriptive observations made on the basis of learner corpora are the cornerstone of such systems. To illustrate this approach, we are now testing a set of rules, extracted from observations made on *Frida-bis* through *eXXelant*, to diagnose errors occurring in the agreement of the past participle in the French 'passé composé' tense. The results of this diagnostic system can then be evaluated using the annotated corpus itself, with a part of the corpus used for fine tuning the system and the rest for evaluation. As in Foster (2004) the annotated corpus is used to quantify the silence / noise rates of error detection. The overall success rate of the system will enable us to determine if it is reliable enough to handle free productions or if it should only be integrated into constrained environments within CALL programs. The conclusion that Gamper & Knapp (2002) draw for Intelligent CALL systems in general applies to error diagnostic systems: "While some systems are rather promising, additional research efforts are required in order to tackle the above mentioned problems and to develop authentic learning systems". The main condition for success consists in identifying the appropriate type of didactic integration, making what Bar-Hillel (1964) calls "a judicious and modest use of mechanical aids".

References

Amaral, L., Metcalf, V. and Meurers, D. (2006) Language awareness through re-use of NLP technology. *Paper presented at the pre-conference workshop on 'NLP in CALL - computational and linguistic challenges', CALICO 2006*. Abstract downloadable from http://ccat.uwaterloo.ca/~mschulze/icali_workshop06.html

Amaral, L. and Meurers, D. (2006) Where Does ICALL Fit into Foreign Language Teaching. *CALICO Conference 2006*. University of Hawaii. May 19, 2006. http://www.ling.ohio-state.edu/icall/handouts/calico06-amaral-meurers.pdf

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé, T., Loiseau, M. and Ponton, C. (2004) NLP-based scripting for CALL activities. In: *Proceedings of Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning (COLING 2004)*, August 2004, Geneva, 18–25.

Bar-Hillel, Y. (1964) The future of machine translation. In: Bar-Hillel, Y. (ed.), *Language and Information: Selected Essays on their Theory and Application*. Reading, MA: Addison-Wesley, 180–184.

Belz, J. A. (forthcoming) Telecollaboration, contrastive learner corpus analysis, and data-driven learning: Implications for language program direction. In: Belz, J. A. and Thorne, S. L. (eds.),

Internet-mediated Intercultural Foreign Language Education. Boston: Heinle & Heinle.

Belz, J. A. and Vyatkina, N. (2005) Learner Corpus Research and the Development of L2 Pragmatic Competence in Networked Intercultural Language Study: The Case of German Modal Particles. *Canadian Modern Language Review/Revue canadienne des langues vivantes*, **62** (1): 17–48.

Cambridge Advanced Learner's Dictionary (2003) (Gillard, P. [ed.]) Cambridge University Press: Cambridge.

Cassart, A., Dagneaux, E., Granger, S., Hustquet, C., Verhulst, N. and Watrin, P. (2002) *Final Error Typology*. Deliverable 14, *FreeText* project. Centre for English Corpus Linguistics, Louvain-la-Neuve: Université catholique de Louvain.

Chapelle, C. (1998) Multimedia CALL: Lessons to be learnt from research on instructed SLA. *Language Learning & Technology,* **2**(1): 22–34.

Chuang, F.-Y. and Nesi, H. (2004) Accuracy and the Chinese learner: an EAP approach to teaching grammar. In: Coverdale-Jones, T. (ed.), *Responding to the Needs of the Chinese Learner*. University of Portsmouth Working Papers, 26–32.

Chuang, F.-Y. and Nesi, H. (forthcoming) An analysis of formal errors in a corpus of L2 English produced by Chinese students. To appear in *Corpora*.

Cowan, R., Choi, H. E. and Kim, D. H. (2003) Four Questions for Error Diagnosis and Correction in CALL. *CALICO Journal*, **20** (3): 451–463.

Cosme, C., Delghust J.-L., Gouverneur, C., Granger, S. and Husquet, C. (2003) *FreeText. Deliverable 19, A report on Software Testing*. Centre for English Corpus Linguistics, Louvain-la-Neuve: Université catholique de Louvain.

Dagneaux, E., Denness, S. and Granger, S. (1998) Computer-aided Error Analysis. *System*, **26** (2): 163–174.

Dodigovic, M. (2005) *Artificial Intelligence in Second Language Learning. Raising Error Awareness*. Clevendon, Buffalo & Toronto: Multilingual Matters.

Foster J. (2004) Parsing Ungrammatical Input: An Evaluation Procedure. *Proceedings 4th International Conference on Language Resources and Evaluation (LREC-04)* Lisbon, Portugal, Vol. 6, 2039–2042.

Foster, J. and Vogel, C. (2004) Parsing Ill-Formed Text Using an Error Grammar. *Artificial Intelligence Review*, **21** (3–4): 269–291.

Francis, W. N. (1967) The Brown University Standard Corpus of English: Some implications for TESOL. In: Robinett, B.W. (ed.), *On Teaching English to Speakers of other Languages*. Washington, D.C.: TESOL, 131–135.

Gamper, J. and Knapp, J. (2002) A Review of ICALL systems. *Computer Assisted Language Learning (CALL)*, **15** (4): 329–342. (Extended version downloadable from http://www.eurac.edu/NR/rdonlyres/3CA3C4BB-664D-4B41-BF55-98A5ED0BEEF8/0/icallExtended.pdf)

Gilquin, G., Granger, S. and Paquot, M. (forthcoming) Learner Corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*.

Granger, S. (2003) Error-tagged learner corpora and CALL: a promising synergy. *CALICO Journal* (Special issue on Error Analysis and Error Correction in Computer-Assisted Language Learning), **20** (3): 465-480.

Granger, S., Vandeventer, A. and Hamel, M. J. (2001) Analyse de corpus d'apprenants pour l'ELAO basé sur le TAL. *TAL,* **42**(2): 609–621.

Granger, S. and Rayson, P. (1998) Automatic profiling of learner texts. In: Granger, S. (ed.), *Learner English on Computer. London* & New York: Addison Wesley Longman, 119–131.

Hammarberg, B. (1999) *Manual of the ASU Corpus, a longitudinal text corpus of adult learner Swedish with a corresponding part from native Swedes*. Stockholms universitet: Institutionen för lingvistik.

Hegelheimer, V. (2006) Helping ESL writers through multimodal, corpus-based, online grammar resources. *CALICO Journal*, **24**(1): 1–28.

Hegelheimer, V. and Fisher, D. (2006) Grammar, writing, and technology: A sample technology-

supported approach to teaching grammar and improving writing for ESL learners. *CALICO Journal*, **23**(2): 257–279.

Izumi, E., Uchimoto, K. and 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** (2): 119–125.

Lenko-Szymanska, A. (2004) Demonstratives as anaphora markers in advanced learners' English. In: Aston, G., Bernardini, S. and Stewart, D. (eds.), *Corpora and Language Learners*. Amsterdam & Philadelphia: Benjamins, 89–107.

L'haire, S. (2004) Vers un feedback plus intelligent. Les enseignements du projet *Freetext*. *Actes de la Journée d'étude de l'ATALA*, Grenoble, 22 octobre 2004. http://w3.u-grenoble3.fr/lidilem/talal/

Longman Dictionary of Contemporary English (2005) (Summers, D. [ed.]) Pearson Education: Harlow.

Lüdeling, A., Maik, W., Kroymann, E. and Adolphs, P. (2005) Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005,* Birmingham. http://www.corpus.bham.ac.uk/PCLC/

Macmillan English Dictionary for Advanced Learners (2007) (Rundell, M. [ed.]) Macmillan: Oxford.

Menzel, W. and Schröder, L. (1999) Error diagnosis for language learning systems. *ReCALL*, Special online edition, 20–30. http://www.eurocall-languages.org/ recall/index.html

Metcalf V. and Meurers D. (2006) When to Use Deep Processing and When Not To – The Example of Word Order Errors. Pre-conference Workshop on NLP in CALL – Computational and Linguistic Challenges. *CALICO 2006*. May 17, 2006. University of Hawaii. http://www.ling.ohio-state.edu/icall/handouts/calico06-metcalf-meurers.pdf

Milton, J. (1998) Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In: Granger, S. (ed.) *Learner English on Computer*. London & New York: Addison Wesley Longman, 186–198.

Milton, J. and Chowdhury, N. (1994) Tagging the interlanguage of Chinese learners of English. In: Flowerdew, L. and Tong, A.K. (eds.) *Entering Text*. Hong Kong: Hong Kong University of Science and Technology, 127–143.

Myles, F. and Mitchell, R.(2004) Using information technology to support empirical SLA research. *Journal of Applied Linguistics*, **1** (2): 169–196.

Nesselhauf, N. (2004) Learner corpora and their potential for language teaching. In: Sinclair, J. (ed.) *How to use corpora in language teaching*. Amsterdam/Philadelphia, Benjamins, 125–152.

Nicholls, D. (2003) The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In: Archer *et al*. (eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (CL 2003), 572-581.

Pravec N. (2002) Survey of learner corpora. In: *ICAME Journal*, **26**: 81–114.

Suzuki, C., Jung, K., Watanabe, Y., Min, S. and Yoshihara, S. (2004) An analysis of Japanese and Korean Students online. Discussion based on movie reports. *Asian EFL Journa*l, **6**(2). http://www.asian-efl-journal.com

Tenfjord, K., Hagen, J. E. and Johansen, H. (in press) The Hows and Whys of coding categories in a learner corpus (or 'How and why an error tagged learner corpus is not ipso facto one big comparative fallacy'). *Rivista di Psicolinguistica Applicata (RiPLA),* **VI**(3): 93–108.

Tenfjord, K., Meurer, P. and Hofland, K. (2004) The ASK corpus – a language learner corpus of Norwegian as a second language. *Proceedings of the Sixth Teaching and Language Corpora Conference (TALC 2006)*. http://www.ugr.es/~talc6/talc_search/proceedings/60.html

Tschichold, C. (2003) Lexically Driven Error Detection and Correction. *CALICO Journal*, **20**(3): 549–559.

Vandeventer, A. (2001). Creating a grammar checker for CALL by constraint relaxation: A feasibility study. *ReCALL* **13**(1):110–120.

Wible, D., Kuo, C-H., Chien, F-Y., Liu, A. and Tsao, N-L. (2001) A web-based EFL writing environment: integrating information for learners, teachers, and researchers. *Computers and Education*, **37**: 297–315.