Learner corpus evidence fed to a learning management system

Leanne V. Bartley

Departamento de Filologías Inglesas y Alemana
Facultad de Filosofía y Letras
Universidad de Granada, 18071, Granada, España

and

Salvador Valera

Departamento de Filología Inglesa
Facultad de Humanidades y Ciencias de la Educación
Universidad de Jaén, 23071, Jaén, España

ABSTRACT

Learner corpora are claimed to be powerful resources for the diagnosis of language learner difficulties. As such they should ideally be used, among others purposes, to inform teaching. This paper shows how a learner corpus compiled at the Universities of Granada and Jaén, Spain, is used for such purposes. It describes the application of the evidence obtained from a learner corpus of English by Spanish university students to the development of teaching materials that are then made available to students through e-learning resources. While still an in-house development, the teaching resources described in the paper are to follow the NOn-native Spanish corpus of English (NOSE) corpus as regards its release and public access.

Keywords: e-learning, error annotation, POS annotation, Foreign Language Teaching (FLT).

1. INTRODUCTION

Learner corpora have been the subject of much research since the first publications on the issue (e.g. Granger 1998). Two decades later, much of the research in the field focuses on design and processing, and on interlanguage studies, i.e. studies on the language learners' progress towards the target language.

The pedagogical applications of learner corpora have also been discussed at length and, although the use of local learner corpora in the classroom have been claimed to have a great potential (Seidlhofer 2002), corpus-based classroom work still predominantly uses native language corpora.

One of the main features and assets of learner corpora is that they can adopt the descriptive methods and can be submitted to the analytic tools used for native language corpora. Therefore, learner corpora are likely to be annotated, either with error tags or POS (part-of-speech) tags, and can be submitted to quantification. As a result, objective observations about the aspects where students show greater difficulties are easily gathered, and consequently represent an invaluable source of evidence and material for the design of activities relevant to the students sampled in the corpus or of a similar profile.

This paper intends to contribute to the exploitation of local learner corpora and its integration within the teaching vehicles available to students at the universities where the corpus was collected. It therefore reports on the combined use of the NOSE corpus, the EARS annotation system (Díaz-Negrillo 2009), and the ILIAS learning management system.

2. INFORMATICS FOR ERROR ANALYSIS AND E-LEARNING

The applications of the learner corpus NOSE to foreign language teaching (FLT) build on former research in two complementary fields: i) applications and benefits of e-learning resources currently in use at the University of Jaén and according to the results obtained in pilot studies (e.g. Díaz-Negrillo & Valera 2006); and ii) error annotation and error retrieval in databases of English by non-native speakers (e.g. Gamon et al. 2009).

ILIAS (*Integriertes Lern-, Informations- und Arbeitskooperations-System*)[1] is a well-known learning management system. It has been in use at the University of Jaén for several years now and is also the vehicle chosen by the University of Jaén project PID23B to help high school students through to their first year of tertiary education in the course *Introduction to English Linguistics*.

In that project, as in the rest of the applications of the course, the materials made available by lecturers to students have been, for a long time, based on the lecturers' past experience of what students may need at different stages in their progress rather than on experimental evidence of their needs. In order to avoid that and to provide materials based on the actual data of what students need, the NOSE corpus is used here.

The NOSE corpus is a development of the original NOCE corpus (Díaz-Negrillo 2007), a multi-layered error-annotated corpus of English by Spanish university students learning English as a foreign language. The c. 300,000-word corpus bears annotation at six different levels according to the field where the deviation from the target language belongs: spelling, punctuation, word grammar, clause grammar, phrase grammar or lexis.

The annotation scheme is flexible in that it allows to classify deviations merely as belonging to one of the above major areas, or to subclassify it within such areas to varying degrees of detail according to what each of the six descriptive levels above allow for. Thus, one and the same deviation may be described merely as involving word grammar or as involving the undue selection of a plural mark where it should not occur (e.g. *difficults* for *difficult*). The range of error tags used for the annotation ranges between 6 at its more general to 612 at its most detailed (cf. Díaz-Negrillo 2009 on the annotation system). This allows one to retrieve and arrange error types by a number of criteria, e.g. the area of interest of each researcher.

The NOSE corpus is the web-based version of the NOCE corpus. It is also different from its predecessor in that the NOSE improves on the former version in a number of respects (cf. Díaz-Negrillo 2011):

i) including POS annotation by three automatic taggers (Tnt, Stanford and Brill tagger) and therefore allows comparison of different POS scheme,
ii) using a revised version of the error tagset and, therefore, of the annotation,
iii) allowing interrater reliability studies by virtue of including error annotation by several different human taggers of different profiles (e.g. native vs. non-native speakers of English), and
iv) using a different search engine: ANNIS (Zeldes et al. 2009).

[1] ILIAS is freely available from http://www.ilias.de/docu/

3. NOSE EVIDENCE AND WHAT IT ENTAILS

Computerized corpora take a long time and effort to design, compile, computerize and bring online. Still, when they do, they offer a wealth of data. Learner corpus evidence has been exploited in a number of respects (cf. Díaz-Negrillo 2011).

When, as in this case, the corpus is annotated and the annotation is exhaustive, the possibility to retrieve data on specific learner difficulties, in this case the six levels at which the corpus is annotated, is easily at hand.

The first analyses conducted on corpus data reveal clear patterns as regards the most frequent and widespread deviations contained in the corpus (e.g. Díaz-Negrillo and Valera 2010). Thus, for example, as regards vocabulary, the corpus shows 303 errors distributed over 52 types.

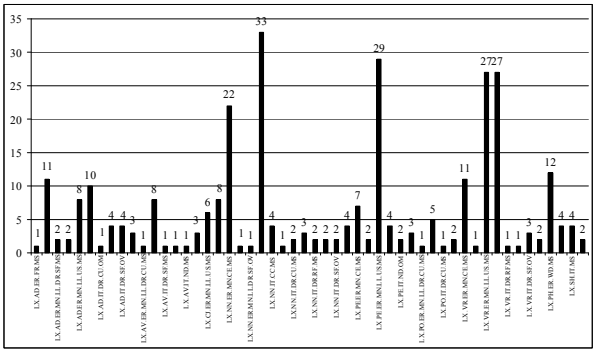

Figure 1. Frequency distribution of Lexis (LX) errors in NOSE (by number of occurrences)

The distribution of the errors is unequal, and five error types amount to 45.54% of the total number of errors. The error dominance observed varies over the six levels of annotation contemplated by EARS and used for the NOSE corpus. This means that the corpus may provide clear and an accurate picture can be obtained of the students' major weaknesses. Again, in the case of vocabulary, two error types prevail: one relates to external errors in nouns (LX.NN.ER.MN) (18.15%), and the other to errors of a variety of types in verbs (LX.VR) (17.82%), as in the following examples:

(1) It depends on the <LX.NN.ER.MN.LL.US.MS>carreer</LX.NN.

ER.MN.LL.US.MS> that you have studied […].[2]

(2) food is very bad for healthy and it can <LX.VR.IT.CC.MS> produce us serious problems</LX.VR.IT.CC.MS>[3]

As the corpus relies on samples of learner language collected at three points of the academic year (beginning, mid-way and end), it can also be used for longitudinal studies within the timeframe of one and the same academic year. According to research on the evolution of students (Bartley and Benítez Castro 2010), hardly any improvement can be noticed from the first to the last sampling throughout the six levels of error description:

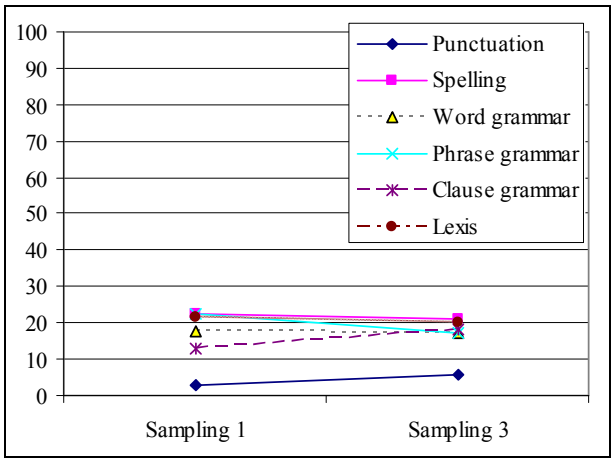

Figure 2. Percentages of incorrect tokens by error type (Bartley and Benítez-Castro 2010)

When courses in English proficiency are taught based on textbooks, it is often difficult to adjust the textbooks' contents to the group's needs. By contrast, corpus evidence like the above shows statistically significant deviations which tend to occur in learner language. Differences between groups may exist, but the error distribution remains in the same range year after year.

Based on the corpus evidence, lecturers can prepare and administer teaching materials for self-study targeting precisely the areas where learning difficulties exist or are likely to arise. As certain associations have also been observed between errors types, these materials can also be selected according to what each of the levels needs more work on. In our example, vocabulary, omission and overinclusion play a minor role in contrast to what happens in other levels, e.g. punctuation and phrase grammar. As the corpus also contains POS annotation, these associations can focus on prepositions (LX.PE) and pay comparatively less attention to auxiliaries (LX.AX) or pronouns (LX.PO).

[2] LeXis, NouN, ExteRnal [existent word], MeaNing, LexicaL, USage, MiSselection.

[3] LeXis, VeRb, InTernal [work combination does not exists], ColloCation, MiSselection

4. CORPUS-BASED E-LEARNING

ILIAS is a useful learning management system that caters successfully for a number of needs, although it also has some limitations, like lack of flexibility as regards activity type or the language used by the system (only Spanish). In its present version, it allows a range of activities, even if not all of them are equally suitable for foreign language teaching or, at least, for the purposes described here.

Figure 3. ILIAS menu of activities

Thus, *Pregunta flash* (an activity based on a visual cue) or *Pregunta mapa de imagen* (an activity based on a picture where the student selects an area) are better suited to other types of studies, because they test the ability to identify, e.g. unhealthy tissue in a scan. Of the options listed in Figure 3, following are examples of activities targeting the areas of error dominance described above according to what students may have access to via ILIAS.

The paper therefore presents only some of the possible activities that are possible. They are intended to meet language difficulties, i.e. the deviation patterns are the subject of a number of activities designed for online use via ILIAS. Different activities are designed according to whether they are used as proactive or as remedial teaching. The activities are made available to students of the same profile as those whose written output was sampled for the compilation of the NOSE corpus, according to their specific needs.

5. ACTIVITIES

The activities presented here are of four types:

i) Spot the mistakes (Figure 4),
ii) Multiple choice (Figure 5),
iii) Matching pairs (Figure 6), and
iv) Cloze text (Figure 7).

Other activities that are possible are not included here for brevity. The activities are based on the evidence retrieved from the NOSE corpus in the sense that they meet the students' needs evidenced in the NOSE corpus, and in some cases the activities actually use the learner language contained in the NOSE.

This means that the students have a second chance to revise their writing and learn from their errors. This does not breach anonymity, as the samples are coded and no names are associated to the samples in the corpus.

In the first activity used here as an example, a text has been selected which contains vocabulary verb errors. To avoid distracting the student from the focus of the activity, other mistakes were corrected and the non-nativeness effect of the text was improved as much as possible.

Figure 4. An example of an activity based on NOSE language (I)

On other occasions, the BNC can be used to avoid excessive exposure to non-native language and give prominence to language authenticity. This is the case of the activities presented in Figures 5 to 7:

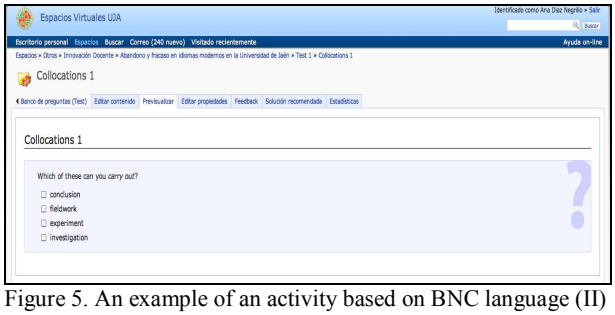

Figure 5. An example of an activity based on BNC language (II)

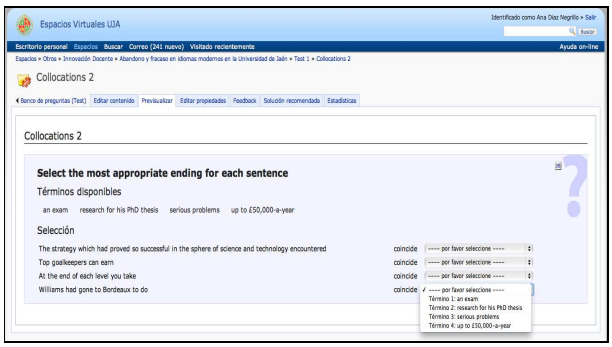

Figure 6. An example of an activity based on BNC language (III)

In some exercises, the learner encounters a range of options among which there is a learner's selection in a similar context in NOCE, the item that occurs in that context in the BNC and, in some cases, other similar choices, although incorrect, selected by the researcher as distractors. An example of this type of activity is in Figure 7:

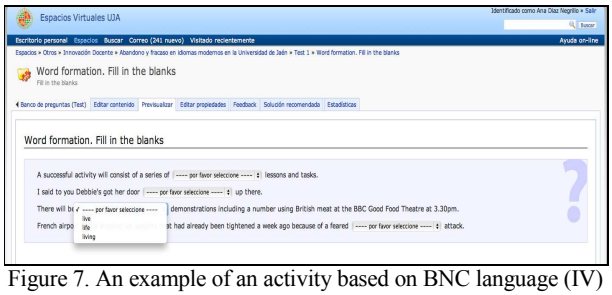

Figure 7. An example of an activity based on BNC language (IV)

6. CONCLUSIONS

This is an application of computerized learner corpora to language learning. It has described how evidence from a learner corpus can be used to develop remedial activities in a learning management system. It has also shown that the combined use of a computerized learner corpus, of a fine-grained error annotation system and of e-learning resources allows to optimize the student's e-learning time and the teaching resources.

7. ACKNOWLEDGEMENTS

The work presented here has been supported by funds of the Spanish Ministry of Science and Innovation (*Ministerio de Ciencia e Innovación*) and FEDER funds under project HUM2007-60107/FILO. It is also part of the University of Jaén project PID23B.

8. REFERENCES

[1] L. Bartley and M.Á. Benítez-Castro, "Lexical competence in English learner writing", paper presented at the **31st Annual NY's TESOL Applied Linguistics Conference**. New York, 2010.

[2] A. Díaz-Negrillo, **A fine-grained error tagger for learner corpora**. Unpublished Ph.D. thesis, University of Jaén, 2007.

[3] A. Díaz-Negrillo, **EARS: A user's manual**. Munich: LINCOM Academic Reference Books, 2009.

[4] A. Díaz-Negrillo, "A web-based access to a Spanish language learner corpus", paper presented at the **Second International Conference on Education, Training and Informatics: ICETI 2011**, March 27th-30th, Orlando, 2011.

[5] A. Díaz-Negrillo and J. Fernández-Domínguez, "Error tagging systems for learner corpora", **RESLA**, vol. 19, 2006, pp. 83-102.

[6] A. Díaz-Negrillo and S. Valera, Propuestas docentes para la enseñanza en línea de la Lengua Inglesa I. In **Actas de las Jornadas de Trabajo sobre Experiencias Piloto de Implantación del Crédito Europeo en las Universidades Andaluzas**. Cádiz: Universidad de Cádiz, 2006.

[7] A. Díaz-Negrillo and S. Valera, "A learner corpus-based study on error associations", **Procedia Social and Behavioral Sciences,** vol. 3, 2010, pp. 72-82.

[8] M. Gamon, C. Leacock, C. Brockett, W.B. Dolan, J. Gao, D. Belenko and A. Klementiev, "Using statistical techniques and web search to correct ESL errors", **CALICO**, vol. 26, issue 3, 2009, pp. 491-511.

[9] S. Granger, **Learner English on Computer**. London: Longman, 1998.

[10] B. Seidlhofer, "Pedagogy and local learner corpora: working with learning-driven data", in S. Granger, J. Hung and S. Petch-Tyson (eds.), **Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching**, Amsterdam and Philadelphia: John Benjamins, pp. 213-234, 2002.

[11] A. Zeldes, J. Ritz, A. Lüdeling and C. Chiarcos, "ANNIS: A Search Tool for Multi-Layer Annotated Corpora", in **Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009**.