# Learner corpora: the case of the NOSE corpus

**Ana Díaz-Negrillo**
**Departamento de Filologías Inglesa y Alemana**
**Facultad de Filosofía y Letras**
**Universidad de Granada, 18071, Granada, España**

## ABSTRACT[1]

This paper provides a brief overview of the scope of learner corpus research and describes a learner corpus by Spanish university students of English, the NOn-native Spanish corpus of English (NOSE). It presents the corpus data, its annotation and how it can be retrieved and exploited for research purposes in the areas of interlanguage studies and automatic recognition of learner-specific features. It also reviews the various research topics that have been investigated in the corpus.

**Keywords:** learner corpus research, error annotation, POS annotation, SLA, FLT.

## 1. INTRODUCTION

It has taken native corpora half a century to evolve from the first computerized databases of raw data (Francis and Kučera 1964) to annotated corpora that can currently be accessed online (e.g. the British National Corpus, hereafter BNC). Based on the experience of their native counterparts, learner corpora are gradually evolving and reaching similar degrees of development. This can be seen, for example, in the number and variety of tools that are used for quantification purposes and in the diversification of learner corpora types.

Another measure of one such degree is online accessibility to learner corpora. Similar stages are being covered as with the BNC, one of the first 100 million-word corpora that became widely accessible. The BNC was originally available only on purchase, then became available online with access to 50 results per query selected at random with a basic search engine, and today it is accessible via websites using a range of interfaces and query engines (BNCweb, BYU-BNC[2]).

Although the situation has changed over the past years, learner corpora are not as readily available as native language corpora. This is one of the limitations most usually referred to in the description of learner corpora. The obvious precedent of many learner corpora, the International Corpus of Learner English (ICLE), is available commercially and in CD-ROM format. The corpus described here, the NOSE corpus[3], is among the first to be soon released and made freely accessible online to the research community.

This paper describes the web-based access to NOSE. It is a report on the type of information that is available in the corpus, how it can be retrieved and how it can be put to use for research purposes in the areas of interlanguage studies, language pedagogy and automatic recognition of learner-specific features.

## 2. LEARNER CORPORA: MAIN FEATURES AND APPLICATIONS

Learner corpora are systematic computerised collections of learners' language productions that are used mainly for investigation in second language acquisition and foreign language teaching (Nesselhauf 2004: 125; 136).

The field of learner corpus research emerged in the 90s when the pioneer ICLE mentioned above started to be collected. Since then, learner corpus projects have been set up at universities and research centres to collect learner corpora of various sizes and languages (Pravec 2002). The main rationale behind the collection of learner corpora was that systematically collected large amounts of learner data that could be submitted to quantification, just as is common practice with native corpora, would provide objective insights into learner language actual use and needs.

[1] This article is a revised extended version of Díaz Negrillo (2011).

[2] BNCweb: http://www.natcorp.ox.ac.uk/; BYC-BNC: http://corpus.byu.edu/bnc/

[3] In its first version the corpus was known as NOCE but has become NOSE in its second version currently under way.

Learner corpora have incorporated techniques from corpus linguistics, like annotation, and have been exploited in a number of directions, which only shows the multidisciplinary nature of this type of specialised language corpora. Large learner corpora have been used for[4]:

- SLA research into a wide range of aspects, for example, syntax (Myles 2005), vocabulary (Lenko-Szymanska 2002), collocations Nesselhauf (2005), overall learner language features (Asención-Delaney and Collentine 2011), etc.
- corpus-informed learner dictionaries, for example the *Cambridge Advanced Learner's Dictionary* (2003) or the *Macmillan English Dictionary for Advanced Learners* (2007), and
- corpus-informed foreign language learning remedial books, for example the series *Common Mistakes at … and how to Avoid Them* (Powell, 2004; Tayfoor, 2004; Driscoll 2005; Cullen, 2007; Moore, forthcoming).

Small learner corpora have been used for:

- immediate remedial work, for example Ragan, (2001), Seildhofer (2002) and Mukherjee and Rohrbach (2006), and
- the development of *ad hoc* annotation, for example, Díaz-Negrillo (2009).

Finally, NLP research has also been undertaken using learner corpora in the field of Computer Assisted Language Learning (CALL), where, for instance, WordPilot (Milton 1998) and ESL Tutor (Cowen, Choi and Kim, 2003) programs to assist language learning stand out.

Despite the progress made after around 25 years of learner corpus research there are areas that call for development. For example, the need of greater research into POS annotated corpora has been pointed out (Meunier 2010), among one of the areas that need development, as well as the investigation into the specific grammatical categories that are needed to describe learner language (Díaz-Negrillo et al. 2010). There is also a need for stronger dialogue between NLP and learner corpus research so that improvement and widening of the applications of learner corpora is possible. Finally, this paper is based on the assumption that if such a dialogue is to be attained, making the corpora complied and annotated freely available to the research community is a first crucial step.

[4] See http://www.uclouvain.be/en-cecl-lcBiblio.html for an extended list of references on learner corpus research.

## 3. THE NOSE CORPUS

The NOSE corpus was collected with two main aims in mind. First to represent the difficulties of the students sampled in the corpus and, second, to investigate into error tagging categories and error description.

After some years of development, NOSE can claim to have three outstanding features:

- It has been used internally to diagnose the needs of students from the degrees of English at the two universities previously mentioned and therefore propose remedial work to counteract students' difficulties.
- The interest of the research group in learner corpus annotation has developed into the design and application to the corpus of an error tagset that offers detailed descriptions of learner errors.
- By the development of a web-based access to the corpus, NOSE will be searchable online and therefore will be at the disposal of the research community.

While the first area, that is, application of the corpus to the design of classroom activities, is described also in this volume (Bartley, Díaz and Valera 2011), and the second has also been described elsewhere (Díaz-Negrillo 2009), this paper describes the corpus and the web-based access to it.

**Data**

At slightly over 300,000 words, NOSE consists in approximately 1000 samples, in the range of 250-300 words. The samples are of various argumentative and descriptive topics and are written in English by approximately 500 Spanish students of English from the universities of Granada and Jaén, Spain.

The samples were collected during four academic years, 2003-2005 and 2007-2009 at the University of Granada, and during two academic years at the University of Jaén, 2007-2009. The text collection took place at three different stages of each academic year (October, February and June). At each stage of the academic year, students were required to write a composition about either one of three argumentative topics or about a subject of their choice, classified as *free writing*. The range of topics offered to students was selected based on what was thought to be of interest to the students at that particular time. The topics differed at each point of data collection, although the option of free writing was constant through the three different stages. This procedure was repeated with each new intake of students at both universities.

**Annotation**

The corpus is annotated with the EARS annotation system (Díaz-Negrillo 2009), a flexible annotation scheme that classifies errors as belonging to six linguistic levels (spelling, punctuation, word grammar, clause grammar, phrase grammar or lexis) and four layers of further error description in terms of:

- the unit involved in the error (word-class, phrase or clause type, etc.),
- the category associated with it (number, complementation, derivation, etc.),
- a distinction between usage and realization errors, and
- a surface structure modification classification (misselection, overinclusion, ordering and omission).

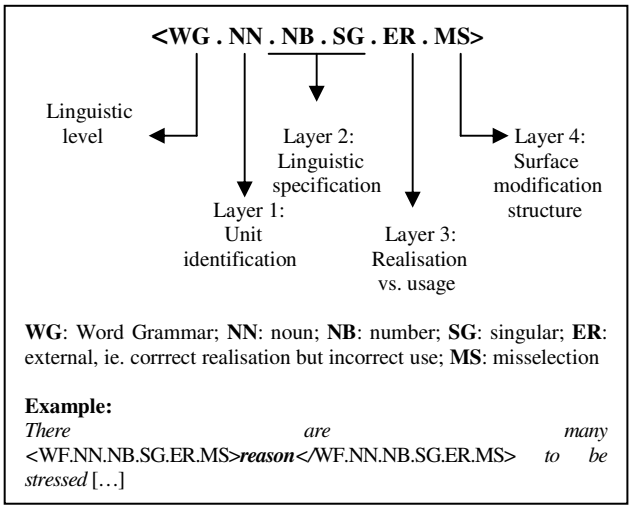


Figure 1. An example of a tag from EARS

In terms of design, this tagset stands out for its fine-grained description of errors and comprehensive scope of description. As shown in Figure 1, the tagset describes errors according to different classifications, four layers, and with varying degrees of specification in some of the layers, see for instance layer 2 in Figure 1. This allows users to annotate according to the layer or degree of specification that best suits his or her research purposes. This also means that a corpus, like NOSE, which is annotated with full tags can also be searched by individual codes. In its first version the tagset contains over 612 possible tags (Díaz-Negrillo 2009). However, the number is likely to change after revision of its design, which is currently under way, is accomplished.

In addition to error tags, NOCE also contains POS annotation at an experimental stage. Currently, the corpus is tagged with three POS automatic taggers, the TnT tagger (Brants 2000)[5], the Standford tagger (Toutanova 2003)[6] and the TreeTagger (Schmid 1994)[7], and research is being undertaken towards adaptation of native linguistic categories to learner-specific categories (Díaz-Negrillo, Meurers, Valera & Wunsch 2010). The latter is a feature that up to date has not been reported in related literature.

**ANNIS2: The corpus search tool**

For retrieval of corpus data, ANNIS2 (Zeldes et al. 2010) is used[8]. ANNIS2 is a search tool that allows searches in the corpus data by error tags, POS tags or learner tokens. Figure 2 shows some of the hits retrieved after searching verb errors in the corpus. ANNIS shows different layers of annotation, which in this case includes error annotation, unfolded for the first and third hits, and POS annotation, unfolded for the second hit:

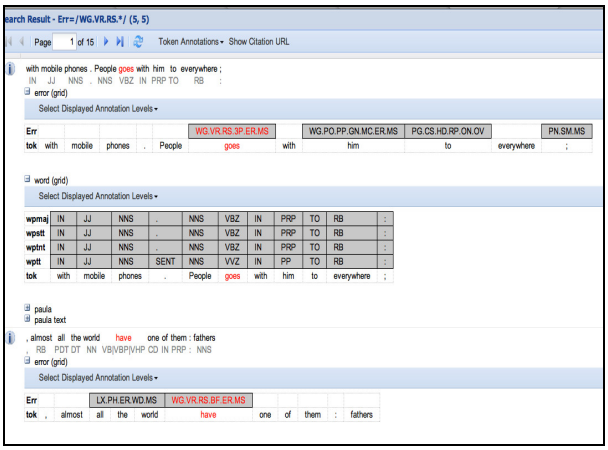

Figure 2. A screenshot of a query of NOSE with ANNIS

**NOSE web-access**

The query engine in preparation allows retrieval of the corpus texts. The interface is user-friendly and allows to search by the following:

- informant profile, which contemplates a number of variables (for example, sex, age, previous knowledge of English, language experience, provenance, etc.),

[5] The TnT tagger is freely available at http://www.coli.uni-saarland.de/~thorsten/tnt/

[6] The Stanford tagger is freely available at http://nlp.stanford.edu/software/tagger.shtml

[7] The TreeTagger is tagger is freely available at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[8] ANNIS is freely available at http://www.sfb632.uni-potsdam.de/d1/annis/

- topic (nine all in all and a free writing option), and
- text type (descriptive, usually associated with the free writing option vs. argumentative, which is the case of the other 9 options).

The corpus can also be searched by sample collection date (early in the academic year, halfway through the academic year, at the end of the academic year), so longitudinal studies within one academic year are also possible. The above variables can also be combined for specific searches therefore allowing a wider range of research possibilities. This means that researchers interested in any of the applications of learner corpus research suited for such a corpus can retrieve the corpus or subcorpora that can be later on submitted to data analysis by using specific tools such as Wordsmith tools or AntConc[9].

## 4. RESEARCH

Work on NOSE was carried out within the framework of a research project which envisaged two major actions:

i) Technical: Design, compilation, computerization and processing of a corpus of English by Spanish learners of the universities of Granada and Jaén, Spain.
ii) Descriptive: Analysis and exploitation of the corpus data for the design of more effective, experimental data-driven teaching strategies.

The two types of actions resulted in two different types of dissemination actions, both based on the separation established above. The project developed from the background knowledge of the issue gained as a result of previous research (Díaz-Negrillo 2007), which allowed faster progress. Specifically, the project relied on a theoretical and experimental approach where several annotation schemes and corpus design approaches were compared (Díaz-Negrillo & Fernández-Domínguez 2006) and then formalized as an exhaustive but also flexible and user-friendly tagging tool (Díaz-Negrillo & García Cumbreras 2007).

The first results appeared as a user's manual to the error annotation and retrieval system proposed for the project (Díaz-Negrillo 2009). The annotation scheme, described elsewhere, was then put to use on corpus evidence and, as a result, two posters were presented at the CALICO 2008 and 2009 conferences (Díaz-Negrillo & Valera 2008 and 2009), where the difficulties and advantages of the annotation scheme used were discussed. The project also proposed strategies for the annotation of particularly difficult cases, and preliminary conclusions for automatic annotation in learner corpora in a paper published two years later as a result of the cooperation of the project team with the team based in the University of Tübingen (Díaz-Negrillo, Meurers, Valera & Wunsch 2010).

The last paper to be added here, Bartley & Valera (2011) and the present paper, discuss two issues on the application of the corpus once completed and available online. The former discusses how the corpus evidence can be put to use as a set of learning activities available through a learning management system (here, ILIAS), as experienced at the University of Jaén. The latter casts a glance backwards (that is, reviews the corpus capabilities) and also forward in that it describes an online application and offers the corpus for general use online.

The descriptive line of research resulted in papers of two types: general and specific. Among the former stands out Díaz-Negrillo & Valera (2010), a review of the statistically significant error distribution in learner English according to the evidence of the corpus, and also of error associations, i.e. co-occurring errors across different descriptive levels or types of errors (internal, i.e. of formation vs. external, i.e. of use). Another relevant paper is Bartley & Díaz-Negrillo (2010) which is based on the MA dissertation Bartley (2010). Here an often neglected topic, non-nativeness in learner language, is studied based on the corpus evidence. The focus here is on the distinction between error and non-native formulation, where the latter does not violate any of the principles of the target language but is communicatively inefficient.

Among the specific papers in the description of learner language, the project has contributed a study on the evolution of lexical competence as regards variety and accuracy (Bartley & Benítez Castro 2010) and three further studies on textual analysis (Bartley & Hidalgo Tenorio 2009 and 2010; Benítez Castro and Fernández Domínguez 2011), where the focus is on the influence of sex on language choice, language modality and transitivity, respectively.

## 5. CONCLUSION

Learner corpus research has been very active since the first initiatives in the 1990s. However, it still has a long way ahead in terms of corpus data exploitation and applications. For this, greater dialogue between the different disciplines involved is necessary as well as greater availability of the resources developed.

[9] AntConc is freely available at http://www.antlab.sci.waseda.ac.jp/software.html

This paper intends to contribute to the later by describing a corpus of learner English by Spanish university speakers that will be available to the research community. It has described the contents of the corpus, its annotation and how the corpus has been exploited so far.

## 6. 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.

## 7. REFERENCES

[1] Y. Asención-Delaney and J. Collentine, "A Multidimensional Analysis of a Written L2 Spanish Corpus", **Applied Linguistics** Advance Access Published January 17, 2011.

[2] L. V. Bartley**, Identification and description of non-nativeness in a corpus of learner English**. Unpublished MA dissertation, University of Jaén, 2010.

[3] L. Bartley, & E. Hidalgo Tenorio 2009. "People is happy, or on the way learners of English construe identity through text and talk". Paper presented at **IGALA6 (Sixth International Gender and Language Association Conference)**, Tokyo, September 18-20, 2009.

[4] L. Bartley & M.Á. Benítez-Castro 2010. "Lexical competence in English learner writing". Paper presented at the **31st Annual NY's TESOL Applied Linguistics Conference**, Teachers College Columbia University, New York City, NY, April 17, 2010.

[5] L. V. Bartley and A. Díaz-Negrillo, "Identificación, Descripción y Análisis de Aspectos No Nativos en un Corpus de Estudiantes Hispano-hablantes de Inglés", in **Memorias de La Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2010**, Orlando, FL, June 29-July 2, 2010.

[6] L. Bartley & E. Hidalgo Tenorio 2010. "Modality in a learner corpus, or on how identity is articulated in narratives". Paper presented at the **Fourth International Conference on Modality in English**, Complutense University, Madrid, September 9-11, 2009.

[7] L. Bartley and S. Valera, "Learner corpus evidence fed to a learning management system". **Proceedings of the ICETI Conference 2011**, Orlando, FL, March 27-30, 2011.

[8] M. Á. Benítez Castro and J. Fernández Domínguez."Transitivity in a learner corpus, or on how students' experiences are shaped by their semantic choices". Paper presented at **Corpus Linguistics 2011: Discourse and Corpus Linguistics**, Birmingham, July 20-22, 2011.

[9] T. Brants, "TnT - A Statistical Part-of-Speech Tagger", in **Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000**, Seattle, WA. URL http://www.coli.uni-saarland.de/~thorsten/publications/Brants-ANLP00.pdf

[10] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees", in **Proceedings of International Conference on New Methods in Language Processing**, Manchester, UK, 1994. URL http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

[11] R. Cowan, H.E. Choi and D.H. Kim, "Four Questions for Error Diagnosis and Correction in CALL", **CALICO Journal,** 2003, Vol. 20, No. 3, pp. 451-463.

[12] P. Cullen, **Common Mistakes at IELTS Intermediate and how to Avoid them**, Cambridge: Cambridge University Press, 2007.

[13] A. Díaz-Negrillo, **A Fine-grained Error Tagger for Learner Corpora.** Unpublished Ph.D. Thesis, University of Jaén, 2007.

[14] A. Díaz-Negrillo, **EARS: a User's Manual**, Munich: LINCOM Academic Reference Books, 2009.

[15] A. Díaz-Negrillo "A web-based access to a Spanish language learner corpus", in **Proceedings of the 2nd International Conference on Society and Information Technologies ICSIT 2011**, Orlando, FL. Internacional Institute of Informatics and Systemics (IIIS), 2011, 174-178.

[16] A. Díaz-Negrillo and J. Fernández-Domínguez. "Error Tagging Systems for Learner Corpora", **RESLA**, 2006, Vol. 19, pp. 83-102.

[17] A. Díaz-Negrillo and M.Á. García-Cumbreras, "A Tagging Tool for Error Analysis on Learner Corpora". **ICAME Journal Computers in English Linguistics**, 2007, Vol. 31, pp. 197-203.

[18] A. Díaz-Negrillo, D. Meurers, H. Wunsch and S. Valera. "Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT", **Language Forum**, 2010, Vol. 36, No. 1-2, pp. 139-154.

[19] A. Díaz-Negrillo and S. Valera. "On Error Associations in Learner Corpora", **Procedia Social and Behavioural Sciences**, 2010, Vol. 3, pp. 72-82.

[20] L. Driscoll, **Common Mistakes at PET and how to Avoid them**, Cambridge: Cambridge University Press, 2005.

[21] W. N. Francis and H. Kučera (eds.), **A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers**. Brown University. Providence, RI, 1964.
[22] P. Gillard (ed.), **Cambridge Advanced Learner's Dictionary**, Cambridge: Cambridge University Press, 2003.
[23] A. Lenko-Szymanska, "Passive and active vocabulary knowledge in advanced learners of English", in B. Lewandowska-Tomaszczyk and P.J. Melia (eds.), **PALC'99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Łódź, April 15-18, 1999**. Frankfurt am Main: Peter Lang, 2000, pp. 287-302.
[24] F. Meunier, "Learner Corpora and English Language Teaching: Checkup Time", **Anglistik: International Journal of English Studies**, 2010, Vol. 21, No.1, 209-220.
[25] J. Milton, "WORDPILOT: enabling learners to navigate lexical universes", in S. Granger and J. Hung (eds.), **Proceedings of the International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching**, December 14-16, 1998, The Chinese University of Hong Kong, 1998, pp. 97-98.
[26] J. Moore, **Common Mistakes at Proficiency and how to Avoid them**, Cambridge: Cambridge University Press, Forthcoming.
[27] J. Mukherjee and J.-M. Rohrbach, "Rethinking applied corpus linguistics from a language-pedagogical perspective: new departures in learner corpus research", in B. Kettemann and G. Marko (eds.), **Planning, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop**, Frankfurt am Main: Peter Lang, 2006, pp. 205-232.
[28] F. Myles, "The emergence of morpho-syntax in French L2", in J-M. Dewaele (ed.), **Focus on French as a foreign language: multidisciplinary approaches**. Clevedon: Multilingual Matters, 2005, pp. 88-113.
[29] N. Nesselhauf, "Learner corpora and their potential for language teaching" in J. Sinclair (ed.), **How to Use Corpora in Language Teaching,** Amsterdam and Philadelphia: John Benjamins, 2004, pp. 125-152.
[30] N. Nesselhauf, **Collocations in a Learner Corpus**, Amsterdam and Philadelphia: John Benjamins, 2005.
[31] D. Powell, **Common Mistakes at CAE and how to Avoid them**, Cambridge: Cambridge University Press, 2004.
[32] N.A. Pravec, "Survey of Learner Corpora", **ICAME Journal**, 2002, Vol. 26, pp. 81-113.
[33] P. H. Ragan, "Classroom Use of a Systemic FunctionaI Small Learner Corpus", in A. Ghadessy, A. Henry and R.L. Roseberry (eds.), **Small Corpus Studies and ELT. Theory and Practice**, Amsterdam and Philadelphia: John Benjamins, 2001, pp. 207-236.
[34] M. Rundell (ed.), **Macmillan English Dictionary for Advanced Learners. Second Edition**, Oxford: Macmillan Education, 2007.
[35] 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, 2002, pp. 213-234.
[36] S. Tayfoor, **Common Mistakes at First Certificate and how to Avoid them**, Cambridge: Cambridge University Press, 2004.
[37] K. Toutanova, D. Klein, C. Manning and Y. Singer. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network", in **Proceedings of HLT-NAACL 2003**, pp. 252-259. http://nlp.stanford.edu/~manning/papers/tagging.pdf
[38] 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.