# Shorter notice

## A tagging tool for error analysis on learner corpora

***Ana Díaz-Negrillo and Miguel Ángel García-Cumbreras***
***University of Jaén***

### *1 Introduction*

A decade of research on errors as contained in learner corpora, usually known as Computer-aided Error Analysis (CEA), has proved not just that increasing attention is being paid to the field (Leech 1997: 15; Dagneaux et al. 1998: 163 et passim; Granger 2003b: 542; Tono 2003: 804–805), but also shown what elements are basic for this purpose: a learner corpus and an error tagset for annotation of errors in the corpus (cf. Granger 1999; Tono 2000).

Interestingly, while a good number of corpora of learners of different target languages and of diverse L1 backgrounds have been collected or are still under collection since the early 90s, tools for annotation of learner corpora seem to be scarce. It also seems that, unlike tools for morphological, syntactic or semantic annotation in corpus linguistics research, the error tagging systems that exist have in general not accomplished stages of public availability (see for instance Milton and Chowdhury 1994; Weinberger 2002; Granger 2003a; Nicholls 2003; Fitzpatrick and Seegmiller 2004; Izumi et al. 2004; cf., however, Hutchinson 1996).

Funded by the Andalusian Regional Council (*Consejería de Educación y Ciencia de la Junta de Andalucía*, Spain), a four-year project based in the English Department of the University of Jaén was launched in 2003 for research on error annotation of learner corpora, specifically on an error tagging system for use on English written material by Spanish learners. The system consists of an error taxonomy and software tools derived from the former.

## *2 The error taxonomy*

The taxonomy provides a language-specific, fine-grained classification of errors building on existing error tagsets which, as the one at Louvain-la-Neuve (Hutchinson 1996; also, Dagneaux et al. 1998) or the one used on the Cambridge Learner Corpus (Nicholls 2003), are intended for a range of languages and may not cover a number of possibilities in the diagnosis of errors specific to language learners of particular L1 backgrounds.

The tagging system is aimed at detail of error description and specificity to use on English written material by Spanish learners. Based on the assumption that "[…] the more refined the tagset the more refined the analysis" (Meunier 1998: 20, re POS taggers), use of detailed error tagging tools is also expected to provide fine-grained analysis of error-tagged data. Descriptive detail is reached in this taxonomy by the incorporation of several sets of information in tags, consisting of:

- identification of the units under description. Alongside error information, the unit where the error is found is identified in the tag with, for example, a punctuation mark for punctuation errors, POS information for grammatical and lexical errors, syntactic functions for syntactic errors, etc.,
- distinction between internal and external errors under the major category of word grammar, where the former refer to errors involving flawed construction of a unit, hence inexistent, e.g. **childs*, and the former to errors involving incorrect use of an existent realization or item, e.g. **everybody in the world have access to it*, and
- narrow linguistic subcategorization of errors. In each of the cases, linguistic information is provided alongside a linguistic level definition (see section 3), relating surface structure modifications (omission, substitution, misordering and misselection), and/or a subcategorization of the linguistic level.

The resulting description is illustrated in Table 1:

*Table 1:* Error description as in the error coding system at Jaén[1]

| | |
|---|---|
| Punctuation:<br>Full stop, End sentence, Omission | *the knowledge of a foreign will always be something useful ↓ |
| Spelling:<br>Word boundary, Merging / Splitting | *our **weakpoint /** *in **bottle necks** |
| Lexis:<br>Self-coinage, Adjective | *not all the language are identically "**cotizides**" |

| | |
|---|---|
| Word Grammar:<br>Tense, Present, External, Verb | *Yesterday I **get up** at 7:40 |
| Clause Grammar:<br>Negation, Assertiveness, Pronoun | *With ETA we can not make **something** |
| Discourse:<br>Co-reference, Personal, Pronoun | *So this is a very special book **who** marked a generation |

## 3 The error editor

The software tools are essentially an editor for computer-assisted insertion of tags in the error-annotation process. Tag options are arranged hierarchically on a menu-driven interface patterned on the usual software menus of word-processors; that is, users can move from general to specific levels of description throughout a chain of linguistic categories and subcategories to arrive at a suitable definition of the error in question. Unlike other taxonomies organised around grammatical errors or errors associated with the POS system, the present taxonomy comprises seven main levels of linguistic description, namely punctuation, spelling, word grammar, phrase grammar, clause grammar, lexis and discourse. Additionally, the taxonomy incorporates information about the superficial modification compared to the target version (omission, substitution, misordering and misselection), thus taking account of the two descriptive approaches recommended by Tono (2003: 804) for the construction of error taxonomies. In practice, linguistic terminal categories lead to error-type categories and the resulting selection accounts for an error description by linguistic and target modification typologies. The error and the error definition thus obtained are automatically bracketed by an opening and a closing XML tag, which at a later stage may be retrieved for CEA research with the aid of data retrieval software, for example WordSmith Tools (Scott 1996). Alongside the main function of tag insertion, the editor includes further editing functions for eventual annotation revision.

The taxonomy used for the error tagging system relies on the evidence found in a written learner 41,421-word corpus collected from 99 first-year Spanish university students doing a degree in English Studies at the University of Granada (see Table 2). At a first stage, a section of the corpus including 17,695 words from the 28 participants present in the three annual samplings (November, March and June) was selected and analysed for errors and a pilot taxonomy and tagset were built (see Table 2). This section was then annotated with a preliminary rudimentary version of the tagset so as to have a closer look at the errors and gain insights into practical aspects involved in the process of error annotation.[2]

*Table 2:* Learner corpus data

| | Participants | Samples | Number of Words |
|---|---|---|---|
| Corpus stage 1 | 28 | 84 | 17,695 |
| Corpus stage 2 | 99 | 188 | 41,421 |

Following revision of the taxonomy and of the error tagset based on the preliminary annotation of the sample, at a second stage, the whole corpus of 41,421 words was analysed for a sounder version of the tool. At the moment, the tagset is being incorporated in the editor designed for assisted annotation of this corpus and of any others which may find this resource useful.

Like the taxonomy, the tagset has been made specific to Spanish learners of English, thus allowing direct access to difficulties of such a learner community during data analysis. In preliminary stages of corpus revision, a number of particular errors demonstrated a salient incidence with respect to others from one and the same category, thus calling for further categorization. Indeed, through further subcategorization, the error classification becomes more detailed but also more specific to such a learner community. This is shown in Table 3 in respect of syntactic errors:

*Table 3:* Syntactic errors as described at Jaén

| | |
|---|---|
| Phrase Grammar:<br>Postmodifier, Noun Phrase | *there are relations **of business** |
| Clause Grammar:<br>Omission, Subject | *In the enterprises ↓ is very important too. |
| Clause Grammar:<br>Misordering, Adjunct | *we just take **from our pocket** the mobile phone |
| Clause Grammar:<br>Structure, Extraposition, Subject | *it’s a good experience **that you can speak with people** |
| Clause Grammar:<br>Structure, Adverbial | *they are the key **for have a job in the future** |

Nevertheless, the tagset is open to modification for specific users' needs. Deletion or insertion of current and new error categories in menus is possible to allow different subcategorization of data and use by a wider research community.

## *4 Conclusion*

Overall, the project is intended to provide a useful tool for SLA researchers and language teachers' approach to errors produced by Spanish learners of English. At present, the error annotation system is close to completion, pending:

- compiling a tagging manual,
- refinement of the tagset based on the feedback gathered from its application on the SPICLE,[3] and
- dissemination of the tool.

## *Notes*

1. Please note that there might be more than one error in the stretches of learner language provided. For clarification, the error under description is highlighted in bold and, in omission errors, ↓ is used.
2. More information about this stage of the research can be found in Díaz-Negrillo (forthcoming).
3. Spanish component of the International Corpus of Learner English.

## *References*

Archer, Dawn, Paul Rayson, Andrew Wilson and Tony McEnery (eds.). 2003. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University: University Centre for Computer Corpus Research on Language.

Dagneaux, Estelle, Sharon Denness and Sylviane Granger. 1998. Computer-aided error analysis. *System* 26: 163–174.

Díaz-Negrillo, Ana. Forthcoming. An error tagging system for the analysis of Spanish corpora of learner English. To appear in *The Grove*. *Working Papers in English Studies.*

Fitzpatrick, Eileen and Steve Seegmiller. 2004. The Montclair electronic language database project. In U. Connor and T. A. Upton (eds.). *Applied corpus linguistics. A multidimensional perspective*, 223–237. Amsterdam: Rodopi.

Granger, Sylviane. 1999. Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In H. Hasselgård and S. Oksefjell (eds.). *Out of corpora. Studies in honour of Stig Johansson*, 191–202. Amsterdam: Rodopi.

Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (eds.). *Computer learner corpora, second language acquisition and foreign language teaching*, 3–33. Amsterdam: John Benjamins.

Granger, Sylviane. 2003a. Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20(3): 465–480. [Special issue on error analysis and error correction in computer-assisted language learning.]

Granger, Sylviane. 2003b. The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37(3): 538–546.

Hutchinson, John. 1996. *UCL Error Editor.* Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Izumi, Emi, Kiyotaka Uchimoto and Hitoshi Isahara. 2004. SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *ICAME Journal* 28: 31–48. Available online at http://nora.hd.uib.no/icame/ij28/Izumi.pdf.

Izumi, Emi, Kiyotaka Uchimoto and Hitoshi Isahara. 2005. Error anotation for corpus of Japanese learner English. *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005),* Jeju Island (Korea), 15 October 2005, 71–80.
Available online at http://acl.ldc.upenn.edu/I/I05/I05-6009.pdf.

Leech, Geoffrey. 1997. Introducing corpus annotation. In R. Garside, G. Leech and T. McEnery (eds.). *Corpus annotation. Linguistic information from computer text corpora*, 1–18. London: Longman.

Leńko-Szymańska, Agnieszka. 2003. Lexical problem areas in the advanced learner corpus of written data. In B. Lewandowska-Tomaszczyk (ed.). *PALC' 2001. Practical Applications in Language Corpora. Papers from the International Conference at the University of Łodz, 15–18 April 1999,* 505–520. Frankfurt am Main: Peter Lang.

Lewandowska-Tomaszczyk, Barbara and Patrick James Melia (eds.). 2000. *PALC' 99: Practical Applications in Language Corpora.* Frankfurt am Main: Peter Lang.

Lewandowska-Tomaszczyk, Barbara, Agnieszka Leńko-Szymańska and Anthony McEnery. 2000. Lexical problem areas in the PELCRA learner corpus of English. In B. Lewandowska-Tomaszczyk and P. J. Melia (eds.), 303–312.

Lüdeling, Anke, Maik Walter, Emil Kroymann and Peter Adolphs. 2005. Multilevel error annotation in learner corpora. Paper presented at the Corpus Linguistics Conference 2005, Birmingham, England, 14–17 July 2005. Available online at http://www.linguistik.hu-berlin.designato.de/korpuslinguistik/projekte/falko/FALKO-CL2005.pdf.

Mason, Oliver and Rafal Uzar. 2000. NLP meets TEFL: Tracing the zero article. In B. Lewandowska-Tomaszczyk and P. J. Melia (eds.), 105–115.

Meunier, Fanny. 1998. Computer tools for learner corpora. In S. Granger (ed.). *Learner English on computer*, 19–37. London: Longman.

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

Nicholls, Diane. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.), 572–581.

Tono, Yukio. 2000. A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk and P. J. Melia (eds.), 323–343.

Tono, Yukio. 2003. Learner corpora: Design, development and applications. In D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.), 800–809.

Weinberger, Ursula. 2002. Error analysis with computer learner corpora. Unpublished M.A. dissertation. Lancaster University.