RESLA 19 (2006), 83-102

ERROR TAGGING SYSTEMS FOR LEARNER CORPORA

ANA DÍAZ-NEGRILLO
JESÚS FERNÁNDEZ-DOMÍNGUEZ
Universidad de Jaén

ABSTRACT. *Learner corpora are used to investigate computerised learner language so as to gain insights into foreign language learning. One of the methodologies that can be applied to this type of research is computer-aided error analysis (CEA), which, in general terms, consists in the study of learner errors as contained in a learner corpus. Surveys of current learner corpora and of issues of learner corpus research have been published in the last few years (Granger 1998, 2002, 2004a; Meunier 1998; Pravec 2002; Tono 2003; Nesselhauf 2004; Myles 2005), where information on CEA research can be found, although usually limited. This article is centred on CEA research and is intended as a review of error tagging systems, including error categorizations, dimensions and levels of description.*

KEYWORDS. *Second language acquisition, learner corpus research, computer-aided error analysis.*

RESUMEN. *Los corpus de estudiantes se utilizan para la investigación de la lengua de estudiantes en formato electrónico con el fin de arrojar luz al proceso de adquisición de lenguas extranjeras. Una de las metodologías que se utilizan en este campo es el análisis informatizado de errores que, en términos generales, consiste en estudiar los errores recogidos en un corpus de estudiantes. Revisiones de los corpus de estudiantes existentes y de cuestiones relacionadas con el campo de la investigación en corpus de estudiantes han sido publicadas en los últimos años (Granger 1998, 2002, 2004a; Meunier 1998; Pravec 2002; Tono 2003; Nesselhauf 2004; Myles 2005), donde se proporciona información sobre la investigación en análisis informatizado de errores, aunque ésta es normalmente limitada. Este artículo se centra en el campo de análisis informatizado de errores y trata de proporcionar una revisión de los sistemas existentes de etiquetado de errores, sus categorizaciones, dimensiones y niveles de descripción.*

PALABRAS CLAVE. *Adquisición de segundas lenguas, investigación en corpus de estudiantes, análisis informatizado de errores.*

1. Introduction

Second Language Acquisition (henceforth SLA) research and computer corpora have come together from the early 90s into what is known as *learner corpus research* (Tan 2005: 126; see also Leech 1998: xvi). This recent enterprise is based on the analysis and interpretation of *computer learner corpora* (henceforth CLC), which have been defined as '[...] systematic computerized collections of texts produced by language learners' (Nesselhauf 2004: 125; 2005: 40; similarly Granger 2002: 7; 2003a: 538; cf. also Leech 1998: xiv).[1] For around fifteen years, learner corpus research has relied on a methodological framework largely inherited from findings of corpus linguistics and from previous methodological approaches to SLA: from corpus linguistics, learner corpus research has adopted the quantificational approach to data and mechanisms of analysis like annotation; from SLA, it has incorporated the methodologies of Contrastive Analysis (henceforth CA) and Error Analysis (henceforth EA) which, improved and complemented with the advantages of computerised corpus research, have given rise to a powerful apparatus for the quantitative and qualitative study of foreign language learning.

Indeed, work on a learner corpus is usually intended to disclose areas where learners tend to show underuse or overuse of linguistic features as opposed to native-language, usually through the methodology of comparisons, or to gain insights into the errors that learners of a given language tend to make, through the methodology of error analysis (Leech 1998: xv; Granger 2002: 12). A major difference between both approaches is that, while work on raw corpora is possible, a specific type of annotation is used in the latter, in particular error tagging. Specifically designed '[...] to cater for the anomalous nature of learner language' (Granger 2002: 18), error tagging is indeed inherent to learner corpora and has become a central part of a methodology of learner corpus analysis known as *computer-aided error analysis* (henceforth CEA) (Dagneaux et al. 1998: 163).

This article is intended as a survey of error tagging systems in use nowadays. To that end, publications on the subject in specialized journals have been examined as well as specific online forums and proceedings of conferences which may have hosted research in the field. It should be noted, however, that not all the projects designing error tagging systems have provided documentation of their systems.[2] For illustration of particular tagsets, this review includes examples of annotated material taken from publications on the tagsets in question.

2. The Computer-Aided Error Analysis (CEA) Methodology

CEA finds its origins in the methodology of EA, which enjoyed both popularity and severe attack during the 70s. Such criticism was mainly due to the data used: the material explored was broadly associated with lists of errors gathered at elicited practices where little attention was paid to task, learner or language variables, hence usually materializing

in non-natural and heterogeneous collections of data (Ellis 1994: 49-50). A further weakness attributed to EA is that errors were explained according to inadequate taxonomies often grounded on non-observable, subjective characteristics and including overlapping categories (Abbott 1980: 122 et passim; Dulay et al. 1982: 143 et passim).

CEA claims to surmount the first flaw, the type of data used in EA by focusing on a different type of material, mainly extensive computerized collections of learner material. The main features of learner corpora improve on SLA data collections as follows:

i) Learner corpora provide a comprehensive picture of learner language performance. Previous methodological approaches to data capture relied on various forms of elicitation, which was justified by the difficulty to obtain features which are rare in natural or uncontrolled data collections. Learner corpora allow infrequent features of learner data to appear, which makes them qualify as comprehensive collections of naturally occurring data (Fitzpatrick & Seegmiller 2004: 223).[3]

ii) Learner corpora are computerized, which is essential in the use of extensive collections of data (Granger 2004a: 126). Before present-day computerised learner corpora, technology could not aim at a relevant sample size and number of subjects, since a bigger corpus would also be unmanageable and inaccessible for want of storage and data retrieval software. Today, computerization makes manageability of data possible and means that corpora can be submitted to a whole range of software tools for corpus analysis, among them error annotation, widening the possibilities and systematization of analysis (Granger 1998: 6).

iii) Results obtained from learner corpus research are considered more reliable than those obtained in previous SLA practices. In the definition of learner corpora above, *systematic* means that data are collected according to a number of criteria, hence allowing for diversified material, and that this collection is to be representative and balanced for the validity of generalizations (van Rooy & Schäfer 2002: 325; Granger 2003b: 465; Nesselhauf 2004: 127; 2005: 40).

iv) Learner corpora contain data in context, which leads to a better understanding of learner material, contributes to the production of refined results, while it also answers to criticism levelled against the restricted scope of EA to errors (Hammarberg 1974: 188). Since errors can be studied alongside non-erroneous learner language, it can be agreed that "[...] learner corpora give us access not only to errors but to learners' total interlanguage" (Granger 1998: 6; cf. also Leech 1998: xvii).

As for the second weakness, CEA researchers have attempted to overcome inadequacy of error classifications by reformulating error taxonomies (see Leńko-Szymańska (2003) for a different view). Nowadays, it is generally claimed that taxonomies should be grounded on the description of observable data and include well-defined linguistic categories to minimize subjectivity in the process of error diagnosis

and categorization (Dulay et al. 1982: 144; Dagneaux et al. 1998: 166; James 1998: 102). Yet, one aspect that current EA is said to be in need of further work is standardisation of error typologies. Unlike other areas where more standardisation might exist, such as learner corpus design, corpus researchers have yet to agree on a general scheme of error annotation (Tono 2003: 801). Shared tendencies may be observed but, in general, research groups often appear to design their own error tagging systems and explore different tagging models and error typologies. Indeed, the diversity of error tagging systems seems to be evidence of the constant questioning of emerging approaches to error annotation, and also of the need for a benchmark for the analysis of computerized learner errors.

Therefore, it can be said that, although other approaches have developed since the heyday of EA, the study of learner errors is still recognised as a fertile undertaking (cf. Ellis 1994: 20, 67-68; Leech 1997: 15; Dagneaux et al. 1998: 163 et passim; Granger 2003a: 542; Tono 2003: 804-805; Ellis & Barkhuizen 2005: 70). Although language instructors may be aware of areas of difficulty in their students, the only way of uncovering error frequencies and, in general, of becoming more aware of their performance is through the analysis of error-tagged material (Milton & Chowdhury 1994: 130; Granger 2002: 14). Besides, as it has been claimed (Granger 2002: 14; see also Izumi et al. 2004: 35), the analysis of errors provides actual evidence of the areas which learners still need to master, therefore disclosing their pedagogical needs.

3. Error Tagging Systems

The construction of an error annotation system includes the design of a taxonomy of errors alongside its pertinent tags, which are to be inserted in the learner corpus manually. Automatization of error annotation has been attempted (cf. Mason & Uzar 2000; Tono 2000; Izumi et al. 2004), although it still seems to be far from automatic annotation practices involved, for instance, in grammatical tagging. The common practise is, however, to use an editor which aids in computer-assisted insertion of tags. Some documented tools of this type are the *Université Catholique de Louvain Error Editor* (*UCLEE*) (Hutchinson 1996), the *TagEditor* (Izumi et al. 2003) and, more recently, a pilot editor under construction at the University of Jaén (Díaz-Negrillo & García-Cumbreras forthcoming). They include tag-associated error categories arranged on a menu-driven interface which the user can select and insert in the text as he/she revises the learner material, and decides on the nature of the error confronted. In addition to error tags, corrections or reconstructions of the target version are often inserted in the tagging process. Once errors are fully tagged, error tags can be retrieved with the aid of software retrieval tools and analysed quantitatively and qualitatively according to the researcher's interest.

CEA has often been undertaken for research on specific problem areas (for example, Granger (1999) on verb tenses, Lewandowska-Tomaszczyk et al. (2000) and Leńko-Szymańska (2003) on the lexical learner problems, Mason & Uzar (2000) on the

article system or Tono (2000) on the acquisition of grammatical morphemes also on lexical issues). Where error tagging has been used in these studies, tags have been specifically created for the study of the aspect under consideration (Tono 2000; Leńko-Szymańska 2003) or part of a larger tagset has been utilized according to the aspect of research (Granger 1999; Izumi & Isahara 2004; Tanimura et al. 2004). Beside specific error tagsets and taxonomies, large-scale error tagging systems exist aimed at covering a wider scope of errors of varied types, which are the type of annotation systems reviewed here. In most cases, each is associated with a learner corpus for which the error system has been constructed, or which, conversely, has served for the construction of the error system. This is important because the corpus may determine the annotation provided. No specific name is usually given to the actual error annotation systems, so they are usually referred to by the name of their associated project or university. Table 1 includes information of the learner corpora associated with error annotation systems:

Learner Corpus	**Approximate Size (in words)**	**L2 Level**	**L1**	**Purpose**	**Mode**	**Learner Language**
CBACLE	1,000,000	Various	Chinese	Academic	Written	English
CLC	20,000,000	Various	Various	Commercial	Written	English
C–LEG	28,000	Advanced	English	Academic	Written	German
FALKO	36,000	Advanced	Unspecified	Academic	Written	German
FRIDA	200,000	Advanced	Various	Academic	Written	French
HKUST	25,000,000	Upper secondary education	Chinese	Academic	Written	English
ICLE	2,000,000	Advanced	Various	Academic	Written	English
JEFLL	700,000	Various	Japanese	Academic	Spoken and written	English
LLC	10,000,000	Various	Various	Commercial	Written	English
MELD	100,000	Advanced	Unspecified	Academic	Written	English
NICT JLE	2,000,000	Various	Japanese	Academic	Spoken	English
PELCRA	500,000	Various	Polish	Academic	Written	English

Table 1. *Learner corpora associated with error tagging systems.*

Information about all the error systems associated with the learner corpora in Table 1 is not always accessible. For this survey a selection of the best documented and representative error tagging systems has been made, which are those associated with:

i) the *Cambridge Learner Corpus* project (henceforth *CLC*),

ii) the FreeText project,

iii) the *Université Catholique de Louvain* (henceforth *Louvain*), and

iv) the *National Institute of Information and Communications Technology Japanese Learner of English* (henceforth *NICT JLE*) (previously known as *Standard Speaker Text* (*SST*) corpus).

Among the four more extensively documented error tagging systems recently mentioned, a most comprehensive account of the tagset is provided in the Louvain *Error Tagging Manual Version 1.1* (Dagneaux et al. 1996), which is part of the error tagging system package available commercially (Hutchinson 1996). An overview of the latter annotation scheme is also provided in Dagneaux et al. (1998). This system was developed at the *Centre of English Corpus Linguistics*, *Université Catholique de Louvain*, Belgium (http://cecl.fltr.ucl.ac.be/, Granger 2004b) and it has been used for research purposes (cf. for example Granger 1999; Díez Prados et al. 2006). The *CLC* error tagset is associated with a learner corpus used for commercial purposes, namely in-house design of ELT material (http://www.cambridge.org/elt/corpus/learner_corpus.htm, Cambridge University Press 2006). An account of the tagset is provided in Nicholls (2003). The FreeText error system, also partly developed at the *Centre of English Corpus Linguistics*, *Université Catholique de Louvain*, Belgium, was used as a tool for learner error diagnosis and the construction of relevant exercises in the framework of a CALL program (http://www.latl.unige.ch/freetext/index.html, FreeText 2004a).[4] The error tagset was presented in Granger (2003b), while a publication on the results of the project is, for instance, L'haire & Vandeventer Faltin (2003) (see also http://www.latl.unige.ch/freetext/en/publications.html (FreeText 2004b) for further publications of the project). Similarly, the *NICT JLE* error system has been used for the development of automatic detection of learner errors, although the main aim of the project was claimed to be the construction of an interlanguage model of Japanese learner English (http://leo.meikai.ac.jp/~tono/, Tono 2004). This error tagset is presented in Izumi et al. (2004), while other publications of the projects are for instance Tono et al. (2001), Izumi et al. (2003) and Izumi et al. (2004; 2005).

4. Features of Error Tagging Systems

According to Granger (2003b: 467), for an error annotation system to be fully effective, it should be consistent, informative, flexible and reusable. The value of including alternative descriptions of individual errors has been stressed elsewhere too (Milton & Chowdhury 1994; Lüdeling et al. 2005). In this section, the principles mentioned above are discussed and illustrated with examples of particular error systems.

A first requirement concerns consistency of the error tagging system. The system should allow consistent and systematic tagging for annotations to be reliable (Fitzpatrick

& Seegmiller 2004: 227). Particularly in the case of large learner corpora as the ones which exist today, it is not always easy to keep track of the decisions that the annotator has to make about error classifications throughout the error tagging process. For this purpose, Granger (2003b: 467) recommends the elaboration of tagging guidelines including detailed information on error categories and annotation principles, as has been done for the Louvain system (Dagneaux et al. 1996).[5] The elaboration of an error manual is also necessary to help use the tagset as well as to keep subjectivity low, which is one of the thorny questions of error-annotation (de Mönnink 2000: 82; Granger 2003b: 474). Indeed, since error tagging is generally manual, the annotator is freer to decide on what and how to tag. As Milton & Chowdhury (1994: 129) point out, annotation "[...] rests on native-speaker's intuition", which may entail some degree of subjectivity in the annotation process. To avoid this, the elaboration of an error manual becomes a compulsory complement of any error tagger.

According to Granger (2003b: 467), error systems should also be informative enough so that the annotation accounts for well-defined error descriptions, while at the same time being *manageable* for annotators, the actual researchers who conduct a manual task. This requirement addresses directly the issue of *granularity* or *delicacy* of error tagsets, i.e. how much detail is encoded. One guiding principle regarding delicacy was claimed by Meunier (1998: 20), for whom the more detailed the system the more refined the analysis and, therefore, the greater disclosure of research possibilities (see also Ellis and Barkhuizen 2005: 60; Díaz-Negrillo forthcoming).

Granularity can be recognized, for example, in the number of tags or codes and in the variety of aspects covered by each tag. The error systems reviewed here range from 31 codes of the *CLC* error system to 100 tags of the FreeText system. However, due to the different application of tagsets during the annotation process, the number of tags cannot always be compared across tagsets to measure granularity. Whereas an annotator using the Louvain or the *NICT JLE* system has to choose one among the 40 or 45 tags of each of them to annotate an error, a researcher annotating with the three-tiered FreeText tagger[6] has to select 3 tags from three different groups of tags out of a total of 100 tags to annotate the same error, and one tagging with the *CLC* error system may need to select an individual code to build a tag or combine various codes to arrive at a final coding of an error.

Tagsets commonly reflect the structure of error taxonomies, therefore, the number and type of dimensions in taxonomies, as well as the aspects covered by each of them, can also be taken into consideration to assess granularity of error tagsets. Section 5.1 reviews the dimensions covered by each of the systems and gives evidence of the predominant weight of the linguistic nature of error taxonomies. The linguistic levels more commonly covered by the error taxonomies are spelling, grammar and lexis. On the other hand, classifications of phonetic, pragmatic or discoursal errors do not seem to be always present in error tagging systems, and when present, their error categories are rather limited. The following table presents the different levels of linguistic analysis of the revised systems:

	CLC	**FreeText**	**Louvain**	***NICT JLE***
Phonology	–	–	–	–
Punctuation	√	√	√	–
Spelling	√	√	√	–
Grammar	√	√	√	√
Lexis	√	√	√	√
Pragmatics	√ (register)	√ (register)	√ (register)	√ (non-verbal cues)
Discourse	√ (pronoun reference)	√ (cohesion)	√ (unnatural discourse)	√ (self-corrections)

Table 2. *Levels of linguistic categorization.*

Finally, Part-Of-Speech (henceforth POS) identification being a common feature of error tagging systems, it may be worth mentioning how this aspect is encoded in the systems reviewed. In this respect, the FreeText system seems to stand out among the rest. It considers 11 major word-class categories and 54 subcategories, while the *CLC* system considers 9 word-classes, the Louvain system 6 and the *NICT JLE* system 11. In addition, the FreeText system includes 12 punctuation categories for identification of punctuation errors, which are specified only in the Louvain system in 3 punctuation categories.

Flexibility in error tagsets, a further principle, implies that categories can be adapted for specific research purposes, for example, particular linguistic levels of description (morphology, lexis, syntax, etc.), or certain linguistic aspects (use of accents, the article system, complementation patterns, etc.). For Granger (2003b: 467), flexibility should be inherent to two stages in the use of error tagsets: annotation and retrieval. As Nicholls (2003: 572) remarks, error tags "[...] are not an end in themselves, but rather act as bookmarks [for queries]". According to this, tagsets should provide the information that one wishes to search for. At the stage of annotation, this means that different levels of specificity of description should be possible according to research purposes, for example, by means of the addition or deletion of tags or codes in tags. Similarly, at a later stage of material retrieval, the system should allow versatile searches at various levels of specificity according to what has been tagged. Hierarchical tagsets are particularly suitable in this respect. Tags in the Louvain and FreeText systems contain a hierarchical structure according to, in this order, linguistic levels, word-classes and error categories. This allows to annotate at different levels of specificity and to search across these three levels of error description in isolation or in combinations of two or three. These searches seem easier with the FreeText system, due to its three-tiered structure. In

addition, the FreeText error system allows combination of the three levels in all tags, whereas in the Louvain system word-classes are not always identified (see 5.2).

A further requirement concerns reusability of the tagset. According to Granger (2003b: 467) error tagsets should allow their use on learner corpora of a variety of languages, for which general error categories become necessary (see also Tono 2003: 801; Tono 2004). Except for the Louvain tagset, which is commercially available, error tagsets appear to be of internal use, so it is hard to tell about the reusability of the tagsets without testing them on different corpora. In general, the error taxonomies reviewed include error categories that may be used for the description of various learner languages, for example *spelling error* (*CLC*, FreeText and the Louvain systems) or *word order* (the *CLC*, FreeText, Louvain and *NICT JLE* systems). In particular, the FreeText and the Louvain system classify errors at different levels of specificity, which facilitates application of their most general categories, for example those referring to linguistic levels, and adaptation of the tagset according to the language needs. By contrast, the *CLC* and the *NICT JLE* systems do not seem to allow for identification of errors at different levels of specificity. Rather, they classify errors by error types, which in some cases may include L2-biased categories, such as *American spelling* (*CLC*) or case of *relative pronoun* (*NICT JLE*) in tagsets used for the description of learner English.

This principle is intended to ensure not only reusability of tagsets but also standardisation of annotation. Certain variables such as learner language, L1 background or size of the corpus differ across learner corpora and, therefore, they may have an effect on the tagset being built and, in turn, on its reusability. The tagset builders may want to make their tagset specific for a given corpus, which is valid and very much depends on the researcher's interests, but they also seem to be aware of the restrictions of reusability of their tagset.

Another guiding principle is discussed by Milton and Chowdhury (1994: 129) and Lüdeling et al. (2005), and concerns the necessity of including alternative descriptions of errors to cater for different possible interpretations and avoid high degrees of subjectivity. As pointed out by Lüdeling et al. "[…] it is impossible **not** to interpret [learner errors]" (2005, bold as in the original), so accounting for various possible error descriptions seems more reliable than providing only one.

Lüdeling et al.'s (2005) annotation system, i.e. the *FALKO* system, allows inclusion of several alternative descriptions favoured by the multi-level model of annotation which it uses. In the model exhibited by the *FALKO* tagging system, error annotations are inserted outside the main body text which allows different types of annotations to be made at different levels as well as addition and deletion of layers according to the researcher's interests. They claim that flat annotation models of the type of the *CLC* or the Louvain system are unsuitable for inclusion of additional interpretation of errors since, once annotations are added alongside the errors, additional layers of annotation cannot be inserted. However, the *HKUST* system, although being flat in nature, claims to include more than one interpretation of errors in their annotations (Milton and Chowdhury 1994: 129). It should however be noted that, although structurally possible,

as has been maintained, this approach may not necessarily ensure inclusion of all possible interpretations: since it is based on a subjective activity, it would be not only laborious but unattainable to compile all possible interpretations of errors (cf. Milton and Chowdhury 1994: 129; see also Tono 2003: 804).

5. Error Taxonomies

Error annotation systems rely on error taxonomies which contain categories for error classification. Categories are associated with error tags used in the process of error annotation to classify the actual errors in the learner corpus to the categories in the taxonomy. Therefore, the viewpoint of the description of errors in tags very much depends on the structure of the taxonomy used.

5.1. *Dimensions in Error Taxonomies*

Current error coding systems usually combine various dimensions of description, hence favouring varied exploitation of the material. Following James (1998: 104-113), Tono (2003: 804) agrees on the inclusion of two dimensions in error taxonomies:

i) a linguistic category classification, accounting for linguistic features of learner errors (lexis, tense, complementation, etc.), and
ii) a target modification taxonomy, accounting for superficial alterations of learner errors as compared to the target native version (omission, addition, order, etc.).

Most of the systems reviewed ground their error taxonomies on linguistic classifications. In some cases, the taxonomies include only linguistic categories, as is the case of the *NICT JLE* system.[7] In others, the systems consist of a major number of linguistic error categories, although they may also include tags for target modification descriptions, which do not combine with categories of linguistic description. This is the case for the Louvain taxonomy which contains tags to annotate order, omission or redundancy errors. Example (1) illustrates an omission error (*WM*, i.e. *word missing*), tagged with the Louvain error system:

(1) […] big ruined walls stood (WM) 0 $rising$ towards the sky. (*Louvain*; Dagneaux et al. 1998: 166).

Finally, there are error systems which, largely based on linguistic classifications, allow a combination of both types of descriptions in the same error categories, as advised by James (1998). This is the case of the *CLC* and FreeText tagsets, although this is found only in a limited number of their tags. In the *CLC* tagset, this is said to occur in the majority of the annotations, as in (2), which includes a tag describing an error of missing punctuation, *MP*. There are however other cases where a combination of both dimensions does not exist, as in (3), which includes a tag describing an error of argument structure, *AS*:

(2) He died<#MP>|.We</#MP> buried him the next day. (*CLC*; Nicholls 2003: 574),

(3) <#AS>Hardly ever do you meet | You hardly ever meet </#AS> people […] (*CLC*; Nicholls 2003: 576).

The FreeText tagset allows a combination of target modification descriptions with the levels of syntax and punctuation. Example (4) shows a combination of the level of *syntax*, *X*, and a target modification classification, *ORD* for *order*, and example (5) illustrates a combination of the level of *punctuation*, *Q*, and a target modification classification, *CON* for *confusion*. By contrast, example (6) shows an error described only with regard to linguistic features identified with the level of *morphology*, *M*, and the error category of *derivation-suffixation*, *MDS*:

(4) <X><ORD> Je peux m'amuser bien (bien m'amuser). (FreeText; Granger 2003b: 478),[8]

(5) <Q><CON> La langue devient plus française – (:) on l'appelle maintenant le créole francisé. (FreeText; Granger 2003b: 478),

(6) <M><MDS> […] qui continue à évolutionner (évoluer). (FreeText; Granger 2003b: 478).

The *FALKO* error tagging system seems to offer a multidimensional approach to error description. Although at the time of writing this paper it is still under construction, it has already claimed to include linguistic and target modification information and, in addition, to incorporate explanation of errors, which seems to diverge with respect to the descriptive character of the rest of error systems.[9]

5.2. *The Structure of Error Taxonomies and Error Tags*

It has been mentioned that error taxonomies are most frequently grounded on the linguistic description of errors. However, the way the linguistic information is organised in taxonomies varies from system to system. The Louvain and FreeText taxonomies are arranged by linguistic levels of analysis, in which POS information is later specified. Conversely, the *CLC* and *NICT JLE* taxonomies are arranged by word-classes, which are further defined according to their morphological, lexical and syntactic features.[10] In what follows, the structure of the error taxonomies reviewed and their pertinent tags are presented.

The Louvain and FreeText systems share certain features, as will be seen from their taxonomy structure and encoding of error categories. The construction of the FreeText system was based on and developed from the structure of the Louvain system. However, a major difference between the two, which also makes the FreeText system different from the rest, is its three-tiered tagging structure and its three-level taxonomy. The Louvain system includes two levels of description:

i) *error categories*, i.e. linguistic level (*form*, *punctuation*, *grammar*, *lexico-grammar*, *register*, *word redundant / word missing / word order* and *style*), and

ii) *error subcategories*, i.e. POS information (only occasionally) and type of error (*spelling*, *tense*, *grade*, *countable / uncountable*, etc.).

Both levels are presented in one tag, as can be seen in the two errors tagged in (7):

(7) […] barons that (GVT) lived $had lived$ in those (FS) castels $castles$. (*Louvain*; Dagneaux et al. 1998: 16),

where the first tag describes an error associated with the level of *grammar*, *G*, the word-class *verb*, *V*, and the grammatical category *tense*, *T,* and the second error with the level of *form*, *F*, and the subcategory of *spelling*, *S*. The word-class of the unit tagged is not identified in the second case. As in (7), the error tag is inserted before the error, while the correction appears after the error between $ symbols.

By contrast, the FreeText system accounts for three levels:

i) *domain*, i.e. linguistic level of analysis (*form*, *morphology*, *grammar*, *lexis*, *syntax*, *register*, *style*, *punctuation*, *typo*),

ii) *category*, i.e. error categories (*homonymy*, *compounding*, *voice*, *prefab*, *word order*, etc.), and

iii) *word category,* i.e. POS classification and sub-classification (*adjective-comparative*, *article-indefinite*, *noun-proper*, etc.).

These three levels are represented in a sequence of three independent tags, one for each of the levels:

(8) […] qui ne sait pas garder le moindre <F><DIA><NOM> #secret$ secrèt </NOM></DIA></F>. (FreeText; Granger 2003b: 470),

where the error is associated with the domain or level of *form*, *F*, the error category of *diacritics*, *DIA*, and with the word category *noun*, *NOM*. As shown, the FreeText system uses Extensible Mark-up Language (henceforth XML) tags: a sequence of opening tags and the correction is inserted in front of the erroneous form, while the closing XML tag comes after the error. Nowadays understood as an inherent feature of language corpora, the use of mark-up languages is one of the strengths of the FreeText, and of the *CLC* and *NICT JLE* systems as will be shown below, since it reputedly ensures standardisation and compatibility with other systems of the same kind. (See http://www.w3.org/XML/ The World Wide Web Consortium 2006).

In contrast with the Louvain and the FreeText systems, the *CLC* and *NICT JLE* taxonomies are organised by word-classes. The *CLC* system relies on a bidimensional taxonomy which combines a target modification dimension, consisting of:

i) *wrong form*
ii) *missing*
iii) *word or phrase needs replacing*
iv) *word or phrase is unnecessary*
v) *word is wrongly derived*

and a linguistic classification dimension centred around POS categorization, consisting of:

i) *pronoun*
ii) *conjunction*
iii) *determiner*
iv) *adjective*
v) *noun*
vi) *quantifier*
vii) *preposition*
viii) *verb*
ix) *adverb*.

The resulting annotation is shown in (9):

(9) […] lawyers, doctors, etc, <#UA>they</#UA> hardly earn #50,000 a year. (*CLC*; Nicholls 2003: 576)

The *CLC* tag thus includes a first code referring to a target modification classification (in the example, *U* for *unnecessary word*) and then the word-class associated with the word in question (in the example, *A* for *pronoun*). In addition to the two classifications presented above, the *CLC* taxonomy also includes the following categories:

i) *punctuation*
ii) *countability*
iii) *false friends*
iv) *agreement errors*
v) *additional error codes,* including a miscellaneous group of further errors, such as *incorrect argument structure*, *incorrect verb inflection*, *spelling confusion error* or *incorrect formation of negative*, among others.

POS information is specified in the *CLC* error annotation in most cases except for punctuation errors, and for some error categories grouped under *additional error codes*. As can be noticed, for the construction of tags, the *CLC* system is based on the

combination of codes from different levels rather than on the selection of predefined tags. The *CLC* system also makes use of XML tags.

The *NICT JLE* system is related to a monodimensional taxonomy of errors comprising information about (primarily) linguistic aspects of errors at two levels, including:

i) major categories, or POS categories (*noun*, *verb*, *modal verb*, *adjective*, *adverb*, *preposition*, *article*, *pronoun*, *conjunction*, *relative pronoun*, *interrogative* and *others*[11]), and

ii) error categories (*noun case*, *verb lexis*, *number of adjective*, *adverb inflection*, *complement of preposition*, etc.).

The *NICT JLE* taxonomy is represented in its tags, as illustrated in (10):

(10) I belong to two baseball <n_num crr= "team"> team</n_num>. (*NICT JLE*; Izumi et al. 2005: 75),

where *n* and *num* indicate the word-class of the error, *noun*, and the error category, *number*. XML grammar can again be recognised from *NICT JLE* tags; in this case, the correction, in quotation marks, is part of the opening tag. Extension of the *NICT JLE* system is towards categorization of unnatural (although not erroneous) instances of English usage (Izumi et al. 2005), although no concluding results have been released yet.

Finally, the *FALKO* tagging system is different from the rest of the systems in the model of the annotation it provides and the type of information it annotates. While the above mentioned tagging systems are examples of flat annotation, i.e. error tags are attached alongside errors, the *FALKO* system relies on standoff annotation, i.e. annotations do not interrupt the text, but are inserted separately, allowing for multi-level annotation (Lüdeling et al. 2005). With this model, *FALKO* annotates at three levels:

i) POS,

ii) one or more corrections, and

iii) linguistic information at the levels *orthography*, *word formation*, *agreement*, *government*, *tense*, *mood*, *word order* and *expression*.

In turn, each of these linguistic levels breaks down further into three sub-levels associated with Corder's (1974) and Ellis' (1994) steps in EA:

i) identification of errors,

ii) description of errors, according to a target modification classification, and

iii) explanation of errors, i.e. source of errors,

which add a further dimension, explanations, rarely considered in current error tagging systems. From the above, it may be said that the *FALKO* system incorporates an annotation model and an EA methodological approach to current CEA methodology never experimented on learner corpus research.

6. Other Approaches to Error Tagging

One approach to error annotation stands out among the rest of the tagging systems mentioned in 3. The error tagging system proposed by Fitzpatrick and Seegmiller (2004) for the *MELD* differs from the rest in the viewpoint it takes regarding the annotation of errors. Fitzpatrick and Seegmiller (2004: 226) state that a list of errors "[...] limits the errors recognized to those in the tagset [and] introduces the possibility that annotators will misclassify those errors that do not fit neatly into one of the tags on the list." This is why the *MELD* error annotation system relies on reconstruction of the erroneous language uses rather than on description of errors by use of codes:

(11) school systems {is/are}
since children {0/are} usually inspired
becoming {a/0} good citizens (Fitzpatrick and Seegmiller 2004: 226).

According to Fitzpatrick and Seegmiller (2004: 226), the advantages that reconstruction has over classifying are, first, that the process is faster and, second, that a reconstruction facilitates tagging and parsing of learner corpora. However, it seems that this approach might be considered to hinder the possibilities of EA, since retrieval of data becomes limited to the reconstructions. In addition, providing reconstructions or corrections in isolation stresses the problematic issue of interpretation of learner material. This is apparent even if they explain that the issue of reliability and consistency of annotation can be handled with consistency experiments, including interrater reliability, precision and recall (Fitzpatrick and Seegmiller 2004: 227 et passim).

7. Conclusion

CEA stands as a learner corpus methodology which proves useful to disclose insights into how languages are learnt. However, probably due to the manual work involved in it, CEA still seems to be lagging behind in learner corpus research practices, and to be still in its early stages of development as some features of error tagsets seem to show: first, error tagsets for description of written language abound, while error tagging systems for the analysis of spoken learner data are less frequent. Second, error taxonomies tend to account for diverse dimensions of error classification, encoding conventions and annotation models, which only shows that there is no standard of error

annotation. Tools do not seem to be shared by the research community and, when commercialized (e.g. the Louvain error tagging system) they do not appear to be used as widely as it would be expected. Third, while taxonomies may be often grounded mostly on linguistic categorization of errors, the linguistic levels considered focus on grammatical and lexical categorizations, and other levels (phonetic, pragmatic, discoursal) are hardly represented in error taxonomies. In accordance with the grammatical and lexical character of error tagsets, most error tagging systems tend to be token-based, hence the widely spread use of POS definitions as a central feature of linguistic categorization. In practice this means that not only are certain linguistic units left undefined in error tagsets, but also that certain levels of analysis are left unexamined. It seems reasonable to claim that, once automatization bridges one of the biggest obstacles of CEA research, new possibilities of research may be disclosed and CEA may stand as a practical undertaking worth looking into in more detail.

NOTES

1. See Granger (2002) for a review of learner corpus research and Pravec (2002) for a review of existing learner corpora. See Tono (2003), Granger (2004a) and Nesselhauf (2004) for a review of the development of learner corpora, learner corpus research and the potential of learner corpora in language pedagogy. See more recently Myles (2005) for an assessment of recent progress in the field and description of the potential of up to date computerized methodologies in SLA research.
2. The scarce documentation available for certain error systems is due to the fact that error systems may still be under construction and no results have been made available yet, as is the case of the *Fehlerannotiertes Lernerkorpus* (henceforth *FALKO*) and the *Polish and English Language Corpora for Research and Applications* (henceforth *PELCRA*).
3. Reservations in this sense have been voiced, for example '[...] foreign language teaching context usually involves some degree of 'artificiality' [...]' (Granger 2002: 8).
4. Please note that despite its association with Louvain, the FreeText error tagging system should be distinguished from the Louvain system already mentioned.
5. The elaboration of an error tagging manual for the Louvain system is also justified by its commercialization.
6. The FreeText error tagging system makes use of three tags in the annotation of an error, where each of the tags responds to a level of analysis.
7. Please note that there is, however, one error category (*misordering of words*) which is associated with target modification information.
8. Please note that this is not an example of a tagged error but rather of a combination of two tags (not three, as is usual in the FreeText tagging system) and a sequence containing the error in the *FRIDA* corpus. An example of a tagged error would be most adequate, but it was not found in published sources.
9. Please note that the *CLC* system includes the error category *false friends,* which is also explanatory rather than descriptive (Dagneaux et al. 1998: 166).
10. Please note that, beside linguistic information, the *CLC* system also includes an important amount of target modification information. Therefore, when its taxonomy of errors is said to be word-class centred, reference is being made only to the linguistic information of the system.
11. The major category *others* groups error categories such as *collocation*, *misordering of words*, *unknown type of error*, etc., *misordering of words* being the only error category associated with a target modification taxonomy.

REFERENCES

Abbott, G. 1980. "Towards a more rigorous analysis of foreign language errors". *IRAL* 18 (2): 121-134.

Cambridge University Press. 2006. *Cambridge Learner Corpus*. [Available at http://www.cambridge.org/elt/corpus/learner_corpus.htm]

Corder, S. P. 1974. "Error analysis". *The Edinburgh Course in Applied Linguistics (Vol. 3).* Eds. J. Allen and S. P. Corder. London: Oxford University Press. 122-154.

Dagneaux, E. et al. 1996. *Error Tagging Manual Version 1.1*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Dagneaux, E. et al. 1998. "Computer-aided error analysis". *System* 26: 163-174.

de Mönnink, I. 2000. "Parsing a learner corpus". *Corpus Linguistics and Linguistic Theory*. Eds. C. Mair and M. Hundt. Amsterdam: Rodopi. 81-70.

Díaz-Negrillo, A. Forthcoming. "Fine- vs. coarse-grained in Spanish corpora of learner English". *Greta. Revista para Profesores de Inglés.*

Díaz-Negrillo, A. and M. A. García-Cumbreras. Forthcoming. "A tagging tool for error analysis on learner corpora". *ICAME Journal*.

Díez Prados, M. et al. 2006. "The ICLE error tagging project: analysis of Spanish EFL writers". Paper presented at the *Fourth International Contrastive Linguistics Conference*, Santiago de Compostela, Spain, 19-23 September.

Dulay, H. C. et al. 1982. *Language Two*. Oxford: Oxford University Press.

Ellis, R. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Ellis, R. and G. Barkhuizen. 2005. *Analyzing Learner Language*. Oxford: Oxford University Press.

Fitzpatrick, E. and M. S. Seegmiller. 2004. "The Montclair electronic language database project". *Applied Corpus Linguistics. A Multidimensional Perspective.* Eds. U. Connor and T. A. Upton. Amsterdam: Rodopi. 223-237.

FreeText. 2004a. *FreeText: French in Context: an Advanced Hypermedia CALL System Featuring NP Tools for a Smart Treatment of Authentic Documents and Free Production Exercises.* [Internet document available at http://www.latl.unige.ch/freetext/]

FreeText. 2004b. *FreeText Publications.* [Internet document available at http://www.latl.unige.ch/freetext/en/publications.html]

Granger, S. 1998. "The computer learner corpus: a versatile new source of data for SLA research". *Learner English on Computer*. Ed. S. Granger. London: Longman. 3-18.

Granger, S. 1999. "Use of tenses by advanced EFL learners: evidence from an error-tagged computer corpus". *Out of corpora. Studies in Honour of Stig Johansson*. Eds. H. Hasselgard and S. Oksefjell. Amsterdam: Rodopi. 191-202.

Granger, S. 2002. "A bird's-eye view of learner corpus research". *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Eds. S. Granger, J. Hung and S. Petch-Tyson. Amsterdam: John Benjamins. 3-33.

Granger S. 2003a. "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.

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

Granger, S. 2004a. "Computer learner corpus research: current status and future prospects". *Applied Corpus Linguistics. A Multidimensional Perspective*. Eds. U. Connor and T. A. Upton. Amsterdam: Rodopi. 123-145.

Granger, S. 2004b. *Centre for English Corpus Linguistics*. [Internet document available at http://cecl.fltr.ucl.ac.be/]

Hammarberg, B. 1974. "The insufficiency of error analysis". *IRAL* 12 (3): 185-192.

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

Izumi, E. et al. 2003. "The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques". *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, United Kingdom, 28-31 March. 359-366.

Izumi, E. and H. Isahara. 2004. "Investigation into language learners' acquisition order based on the error analysis of the learner corpus". *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*. Tokyo, Japan, 10 December. 63-71.

Izumi, E. et al. 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, E. et al. 2005. "Error annotation for corpus of Japanese Learner English". *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*. Jesu Island, Korea, 15 October. 71-80. [Internet document available at http://acl.ldc.upenn.edu/I/I05/I05-6009.pdf]

James, C. 1998. *Errors in Language Learning and Use. Exploring Error Analysis*. London: Longman.

Leech, G. 1997. "Introducing corpus annotation". *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Eds. R. Garside, G. Leech and T. McEnery. London: Longman. 1-18.

Leech, G. 1998. "Learner corpora: what they are and what can be done with them". *Learner English on Computer*. Ed. S. Granger. London: Longman. xiv-xx.

Leńko-Szymańska, A. 2003. "Lexical problem areas in the advanced learner corpus of written data". *PALC' 2001. Practical Applications in Language Corpora*. Ed. B. Lewandowska-Tomaszczyk. Frankfurt am Main: Peter Lang. 505-520.

Lewandowska-Tomaszczyk, B. et al. 2000. "Lexical problem areas in the PELCRA learner corpus of English". *PALC' 99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Łódź, 15-*

18 April 1999. Eds. B. Lewandowska-Tomaszczyk and P. J. Melia. Frankfurt am Main: Peter Lang. 303-312.

L'haire, S. and A. Vandeventer Faltin. 2003. "Error diagnosis in the FreeText project". *CALICO Journal* 20 (3) Special issue on error analysis and error correction in computer-assisted language learning: 481-495.

Lüdeling, A. et al. 2005. "Multi-level error annotation in learner corpora". *Proceedings of the Corpus Linguistics 2005 Conference.* Birmingham, United Kingdom, 14-17 July. [Internet document available at http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc]

Mason, O. and R. Uzar. 2000. "NLP meets TEFL: tracing the zero article". *PALC' 99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Łódź, 15-18 April 1999*. Eds. B. Lewandowska-Tomaszcyk and P. J. Melia. Frankfurt am Main: Peter Lang. 105-115.

Meunier, F. 1998. "Computer tools for learner corpora". *Learner English on Computer*. Ed. S. Granger. London: Longman. 19-37.

Milton, J. and N. Chowdhury. 1994. "Tagging the interlanguage of Chinese learners of English". *Entering Text*. Eds. L. Flowerdew and K. K. Tong. Hong Kong: The Hong Kong University of Science and Technology. 127-143.

Myles F. 2005. "Interlanguage corpora and SLA research". *Second Language Research* 21 (4): 373-391.

Nesselhauf, N. 2004. "Learner corpora and their potential for language teaching". *How to Use Corpora in Language Teaching*. Ed. J. M. Sinclair. Amsterdam: John Benjamins. 125-152.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Nicholls, D. 2003. "The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT". *Proceedings of the Corpus Linguistics 2003 Conference.* Lancaster, United Kingdom, 28-31 March. 572-581.

Pravec, N. 2002. "Survey of learner corpora". *ICAME Journal* 26: 81-114. [Internet document available at http://nora.hd.uib.no/icame/ij26/pravec.pdf]

Tan, M. 2005. "Authentic language or language errors? Lessons from a learner corpus". *ELT Journal* 59 (2): 126-134.

Tanimura, M. et al. 2004. "From learners' corpora to expert knowledge description: analyzing prepositions in the NICT JLE (Japanese Learner English) corpus". *Proceedings of IWLeL 2004: an Interactive Workshop on Language e-Learning*. Tokyo, Japan, 10 December. 139-147. [Internet document available at http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/1405/1/16.pdf]

The World Wide Web Consortium (W3C). 2006. [http://www.w3.org/XML/]

Tono, Y. 2000. "A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora". *PALC' 99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Łódź, 15-18 April 1999*. Eds. B. Lewandowska-Tomaszczyk & P. J. Melia. Frankfurt am Main: Peter Lang. 323-340.

Tono, Y. 2003. "Learner corpora: design, development and applications". *Proceedings of the Corpus Linguistics 2003 Conference.* Lancaster, United Kingdom, 28-31 March. 800-809.

Tono, Y. 2004. *Standard Speaking Test (SST) Corpus*. [http://leo.meikai.ac.jp/~tono/]

Tono, Y. et al. 2001. "The Standard Speaking Test (SST) Corpus: a 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography" *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary.* Seoul, Korea, 8-10 August. 257-262. [Internet document available at http://leo.meikai.ac.jp/~tono/sst/asialex2001.pdf]

van Rooy, B. and L. Schäfer. 2002. "The effect of learner errors on POS tag errors during automatic POS tagging". *Southern African Linguistics and Applied Language Studies* 20: 325-335.