Error-tagged Learner Corpora and CALL: A Promising Synergy

Sylviane Granger
Centre for English Corpus Linguistics
University of Louvain

ABSTRACT

Learner corpora—electronic collections of foreign or second language learner data—constitute a new resource for second language acquisition (SLA) and foreign language teaching (FLT) specialists. They are especially useful when they are error-tagged, that is, when all errors in the corpus have been annotated with the help of a standardized system of error tags. This article describes the three-tiered error annotation system designed to annotate the *French Interlanguage Database* (*FRIDA*) corpus. The research took place within the framework of the FreeText project which aims to produce a learner corpus-informed CALL program for French as a Foreign Language. Once annotated, the *FRIDA* corpus was put through standard text retrieval software to extract detailed error statistics and to carry out concordance-based analyses of specific error types. The results were used to focus the CALL exercises on learners' attested difficulties and to improve the error diagnosis system integrated in the CALL program.

KEYWORDS

Learner Corpus, Corpus Annotation, Error Tagging, Concordance, CALL

1. LEARNER CORPORA

Learner corpora, also called interlanguage (IL) or L2 corpora, are electronic collections of authentic foreign or second language data. They differ from the data types commonly used by second language acquisition (SLA) and foreign language teaching (FLT) researchers in two major respects: (a) they are *computerized* and can therefore be analyzed using a wide range of linguistic software tools which provide for quick and efficient manipulation of the data via their search, count, and sort functions and NLP programs which enrich the data with linguistic information (e.g., grammatical category and syntactic structure) and (b) they are *big* and therefore constitute a much more reliable basis to describe and model learner language than has ever been available before. Size is obviously a relative notion. A corpus of 200,000 words is big in the SLA field where researchers usually rely on much smaller samples but minute in the corpus linguistics field at large where recourse to mega-corpora of several hundred million words has become the norm rather than the exception.

CALICO Journal, 20 (3), p-p 465-480.

Learner language differs from native language both quantitatively and qualitatively. It displays very different frequencies of words, phrases and structures, with some items overused and others significantly underused. (For examples, see Granger, 1998.) It is also characterized by a high rate of misuse, i.e. orthographic, lexical, and grammatical errors.[1] While frequency differences can be retrieved automatically by submitting an unannotated learner corpus to a text retrieval software program such as *WordSmith Tools* (Scott, 1996), errors prove much more difficult to detect. Current spell- and grammar-checking programs can handle a large number of native speaker errors but detect only a minority of L2-specific errors. A recent comparison of three spell- and grammar-checking programs for French brought out a 60%-80% success rate for errors produced by native and nonnative speakers of French alike, but the rate fell to 25%-35% for L2-specific errors (see Granger, Meunier, & Watrin, forthcoming).

Before one can hope to produce highly efficient error detection and correction programs, it is therefore necessary to collect a large learner corpus and to analyze the errors contained in it. Within the framework of the EU-funded FreeText project[2], which seeks to produce a CALL program for French as a Foreign Language (FFL) that incorporates NLP tools, including an error diagnosis system, the Louvain team collected and error-tagged a large collection of intermediate to advanced L2 French writing. The corpus, called *French Interlanguage Database* (*FRIDA*), contains 450,000 words, two thirds of which have been fully error-tagged. The error analysis system used to annotate the corpus is described in the following section.

2. ERROR ANNOTATION

2.1. Computer-aided Error Analysis

Once a very popular enterprise, error analysis (EA) is now out of favor with most SLA/FLT circles. It has gone down in history as a fuzzy, unscientific, and unreliable way of approaching learner language. However, errors are an integral part of interlanguage and are just as worthy of analysis as any other IL aspect. As stated by Ringbom (1987, p. 69) “Although error analysis certainly has its limitations, it must be regarded as an important key to a better understanding of the process underlying L2-learning.” Similarly, Ellis (1994, p. 20) notes that although early EA studies were unreliable and difficult to interpret, “the study of learner errors can still serve as a useful tool and is still undertaken.” In particular, a detailed description of learner errors cannot but contribute to one essential FLT aim—that of helping learners to achieve a high level of accuracy in the language.

The EA methodology adopted in the FreeText project is based on the computer-aided error analysis system developed for English by Dagneaux, Denness, and Granger (1998). It consists of the following steps:

1. manual correction of L2 French corpus,
2. elaboration of an error tagging system for L2 French,

3. insertion of error tags and corrections in the text files,
4. retrieval of lists of specific error types and error statistics, and
5. concordance-based linguistic analysis of major error types.

In order to be fully effective, an error annotation system should be:

1. *informative* but *manageable*: it should be detailed enough to provide useful information on learner errors, but not so detailed that it becomes unmanageable for the annotator;
2. *reusable*: the categories should be general enough to be used for a variety of languages;
3. *flexible*: it should allow for addition or deletion of tags at the annotation stage and for quick and versatile retrieval at the postannotation stage; and
4. *consistent*: to ensure maximum consistency between the annotators, detailed descriptions of the error categories and error tagging principles should be included in an error tagging manual.

In devising an error tagging system for L2 French, we were careful to ensure that these requirements were met.

2.2. Error Tagging System

Dulay, Burt, and Krashen (1982, ch. 7) suggest two major descriptive error taxonomies: (a) one based on linguistic categories (general ones such as morphology, lexis, and grammar and more specific ones such as auxiliaries, passives, and prepositions) and (b) the other focusing on the way surface structures have been altered by learners (e.g., omission, addition, misformation, and misordering). They present these two approaches as alternative taxonomies. Like James (1998), however, we believe there is a great benefit to combining them into a single bidimensional taxonomy or even, provided that an additional layer of information on errors is added, into a three-dimensional taxonomy.

The error tagging system developed to annotate the *FRIDA* corpus consists of three levels of annotation: error domain, error category, and word category. These three levels are descriptive rather than interpretative. We have deliberately decided not to use distinctions such as ‘errors’ versus ‘mistakes’ or ‘interlingual’ versus ‘intralingual’ errors, which are difficult to assign and better left for a second stage in the analysis.

2.2.1. Error Domain and Category

The error domain is the most general level: it specifies whether the error is formal (i.e. orthographic), grammatical, lexical, and so forth. Each error domain is subdivided into a variable number of error categories. Table 1 gives the breakdown of the nine error domains.

Table 1
Error Domains and Categories

Error Domains		Error Categories	
<F>	Form	<AGL>	Agglutination
		<MAJ>	Upper/lower case
		<DIA>	Diacritics
		<HOM>	Homonymy
		<GRA>	Other spelling errors
<M>	Morphology	<MDP>	Derivation-prefixation
		<MDS>	Derivation-suffixation
		<MFL>	Inflection
		<MFC>	Inflection-confusion
		<MCO>	Compounding
<G>	Grammar	<CLA>	Class
		<AUX>	Auxiliary
		<GEN>	Gender
		<MOD>	Mode
		<NBR>	Number
		<PER>	Person
		<TPS>	Tense
		<VOI>	Voice
		<EUF>	Euphony
<L>	Lexis	<SIG>	Meaning
		<CPA>	Adjective complementation
		<CPD>	Adverb complementation
		<CPV>	Verb complementation
		<CPN>	Noun complementation
		<FIG>	Prefab
<X>	Syntax	<ORD>	Word order
		<MAN>	Word missing
		<RED>	Word redundant
		<COH>	Cohesion
<R>	Register	<RLE>	Lexis
		<RSY>	Syntax
<Y>	Style	<CLR>	Unclear
		<LOU>	Heavy
<Q>	Punctuation	<CON>	Punctuation confusion
		<TRO>	Punctuation redundant
		<OUB>	Punctuation missing
<Z>	Typo		

For reasons of space, I will give a brief description of only one domain, that of lexical errors, and simply provide one representative example for all the other categories (see additional examples in Appendix A).

The lexical domain <L> groups all lexical errors due to:

1. insufficient knowledge of the conceptual (i.e., denotative) meaning of words: <SIG> (see example 1);
2. violations of the co-occurrence patterns of words. This category covers a wide spectrum from restricted collocations to idioms: <FIG> (see example 2); and
3. violations of the grammatical complementation (i.e., valency) patterns of words. This category covers the valency of verbs <CPV>, nouns <CPN>, adjectives <CPA> and adverbs <CPD> (see examples 3 and 4).

(1) *Il est un peu plus* *__commis à__ [*engagé envers*] *l'idéal européen.* <SIG>
(2) *leurs élèves suivent des* *__cours religieux__ [*cours de religion*]. <FIG>
(3) *... bien qu'il soit libre* *__à__ [*de*] *choisir ce qu'il fait.* <CPA>
(4) *le Canada pourrait se transformer* *__dans__ [*en*] *un type de superstructure.* <CPV>

2.2.2. Word Category

The word category of the erroneous item is tagged using a part-of-speech tagging system comprising 11 major categories, subdivided into 54 subcategories. (See the full list of tags in Appendix B.) The addition of this third tier to the system makes it possible to sort errors by grammatical categories and to draw up a list of relevant error categories for each one. For instance, there are four different relevant error categories for preposition errors: two lexical categories (semantic errors and complementation errors) and two syntactic ones (missing and redundant prepositions). We have not used automatic tagging programs because we were unsure to what extent the errors would affect the success rate of the programs. More importantly, we did not want to tag all words in the texts but only the erroneous forms so as to avoid having to work on overly cluttered text files.

2.2.3. Correction

Correct forms were also inserted in the text files next to the erroneous forms (a) to facilitate subsequent interpretation of the error annotations; and (b) to allow for automatic sorting on the correct forms. A sort on the corrections, used in the FreeText project to improve the success rate of the automatic spellchecking system,[3] in the <F><GRA><NOM> category, for example, gives access to lists of words and all their misspelt variants. (The seven different erroneous spellings of the word *développement* are shown in Figure 1).

Figure 1
Concordance of <F><GRA><NOM>: Sort on Correct Forms

tourné vers le	*<F><GRA><NOM># **dé veloppement$***	**dé vé lopement** culturel
Les ré sultats du	*<F><GRA><NOM># **dé veloppement$***	**dé vé lopment** technique
On a intensifié le	*<F><GRA><NOM># **dé veloppement$***	**dé veloppement** des sports.
Les chantiers de	*<F><GRA><NOM># **dé veloppement$***	**Dé velopement**.
On est pour le	*<F><GRA><NOM># **dé veloppement$***	**dé velopment** de l'avenir.
Elle connaî t un bon	*<F><GRA><NOM># **dé veloppement$***	**developement**.
La consé quence de ce	*<F><GRA><NOM># **dé veloppement$***	**developpment**, limité

A word of caution is needed here. While in some cases (see examples 2-4 above), the correction is indisputable, in other cases (cf. example 1), it is simply one among several that the annotator could have provided. In using the database, it is therefore important to bear in mind that some of the corrections only have an indicative value.

2.2.4. Tag Insertion

To speed up the tag insertion process, we have developed a purpose-built menu-driven editor which allows the annotator to insert error tags at the relevant point in the text by clicking on the appropriate tag from the error tag menu. Using the correction box, the analyst can also insert the corrected form with the appropriate formatting symbols.

In the sample error-tagged text in Figure 2, three errors have been annotated: two grammatical errors—the gender agreement error on the adjective *fort* and the number agreement error on the verb *penser*—and one formal error, the incorrect diacritic on the noun *secret*.

Figure 2
Sample Error-tagged Text

L'héritage du passé est très <G><GEN><ADJ> #fort$ forte </ADJ></GEN></G> et le sexisme est toujours présent. Beaucoup de gens pensent que la femme est un être pas très intelligent, qui bavarde beaucoup et qui ne sait pas garder le moindre <F><DIA><NOM> #secret$ secrèt </NOM></DIA></F>. Ces gens <G><NBR><VSC> #pensent$ pense </VSC></NBR></G> aussi que les femmes ne sont pas aptes à prendre des responsabilités.

3. ERROR STATISTICS AND ANALYSIS

Within the framework of the FreeText project, a large proportion of the *FRIDA* corpus (300,000 words) has been fully error-tagged by the Louvain team. A total of 46,241 errors has been detected manually and the appropriate error tags and corrections inserted in the text files with the help of the error editor.

One of the main advantages of the three-tiered error tagging system is that it allows for a wide range of searches, from the most general to the most specific. It is possible to retrieve all errors in a particular domain (e.g., all <Q> punctuation errors) or in a particular category (e.g., all <GEN> gender errors). It is also possible to search on a specific word category (e.g., a search on <ADJ> provides all the errors that affect adjectives). The most specific query involves full trigrams. For instance, a search on <G><NBR><VSP> will produce a list of all participle forms of verbs containing a number error.

Using detailed statistics extracted from the corpus, we have been able to rank the error domains and categories in decreasing order of frequency. As shown in Table 2, two domains—grammar and form—account for 50% of all errors in the corpus.

Table 2
Breakdown of Error Domains

Tag	Number of Occurrences	Percent
G	11,779	25.38
F	11,452	24.67
L	7,198	15.51
X	7,061	15.21
Q	5,707	12.29
R	1,402	3.02
Y	862	1.85
M	784	1.68
Z	155	0.33

Table 3 lists the top 20 error trigrams and shows that the most frequent error in the entire corpus belongs to the punctuation domain <Q>. The 'missing comma' error type <OUB> <PUV> alone accounts for 8.8% of the total number of errors in the corpus.

Table 3
Top 20 Error Trigrams

	Error Trigram			Number of Occurrences	Cumulative Percent
1	Q	OUB	PUV	4,805	8.80
2	F	GRA	NOM	1,610	12.26
3	L	SIG	PES	1,592	15.69
4	F	DIA	NOM	1,485	18.88
5	G	GEN	ADJ	1,306	21.69
6	G	TPS	VSC	1,200	24.27
7	L	FIG	SEQ	914	26.23
8	X	MAN	ADE	812	27.97
9	F	MAJ	NOM	783	29.65
10	G	NBR	NOM	765	31.29
11	L	SIG	VSC	745	32.89
12	L	CPV	PES	710	34.41
13	L	SIG	NOM	708	35.93
14	F	DIA	ADJ	689	37.41
15	X	MAN	PES	619	38.74
16	F	GRA	ADJ	609	40.05
17	R	RSY	SEQ	584	41.03
18	G	NBR	ADJ	582	42.55
19	G	CLA	AIN	580	43.79
20	G	MOD	VSC	523	44.91

It is interesting to note that the top 10 trigrams account for around one third (31%) of the errors in the corpus, the top 20 for 45%, and the top 50 for 70% of the errors.[4]

Once inserted in the text files, error codes can be searched using a text retrieval tool such as *WordSmith Tools*. Concordances of error tags allow the analyst to visualize errors in context and to sort them in various ways. Figure 3 shows the concordance of adjective complementation errors <CPA> involving prepositions <PES>. The sorting facility in *WordSmith Tools* makes it possible to sort the errors alphabetically on the adjectives governing the erroneous prepositions. This process brings out some particularly error-prone adjectives such as *différent, difficile*, and *nécessaire.*

Figure 3
Concordance of <L><CPA><PES>

selon la nature, tout à fait *différent*	*<L><CPA><PES>#**de**$* à celui que
expériences personnelles *différentes*	*<L><CPA><PES>#**de**$* à celles de
beaucoup plus *différente*	*<L><CPA><PES>#**de**$* que celle
dont les priorités sont *différentes*	*<L><CPA><PES>#**de**$* que celle des autres
on peut dire qu'il est bien *difficile*	*<L><CPA><PES>#**de**$* à résister
il devient très *difficile*	*<L><CPA><PES>#**de**$* à combattre
Il était *difficile*	*<L><CPA><PES>#**de**$* à savoir
Il est *nécessaire*	*<L><CPA><PES>#**de**$* 0 connaître l'anglais[5]
Il est *nécessaire*	*<L><CPA><PES>#**de**$* 0 avoir une bonne santé

4. INTEGRATION OF THE RESULTS INTO THE CALL PROGRAM

The goal of the FreeText project is to produce a hypermedia CALL program for intermediate to advanced learners of French as a Foreign Language (FFL) that relies on natural language processing and communicative approaches to second language acquisition. The program will contain a variety of exercises, both traditional CALL exercises (e.g., multiple choice and cloze tests) and more open-ended exercises which rely on NLP tools, in particular an automatic error diagnosis system which checks learners' answers and provides meaningful feedback.

Within the project, error statistics and analyses were used (a) to select the linguistic areas to focus on in the CALL program and to adapt the exercises as a function of the attested error types and (b) to turn an existing spellchecker of French[6] and a parser of French[7] into an integrated error diagnosis system capable of catering for the most frequently attested error types.

The most error-prone categories have been given special attention in the CALL exercises. They include errors in tense and mood, agreement (number and gender), articles, complementation, and prepositions. The exercises have been designed to reflect the type of context in which learners proved to make mistakes. For instance, for past participle agreement, care has been taken to include a large proportion of feminine subjects since gender agreement proves to be particularly problematic (*Pour cela la chambre à coucher et aussi la literie doivent être soigneusement aérées*) as well as long subjects whose head is relatively far removed from the verb (*Les couvertures de laine et les édredons qu'on utilise pour recouvrir certains meubles doivent toujours être aérés chaque semaine*).[8] Formal errors, which account for a quarter of the errors in the corpus, are targeted through dictation activities (16 in all) and a series of exercises targeting specific difficulties: homonyms and often confused words and phrases (e.g., *peut-être* and *peut être*), nationality words (*parler l'anglais; les Anglais*), and so on. Finally, due recognition is given to punctuation, the poor relation in most CALL programs and, yet, as attested by our error statistics, a major difficulty for learners: the most frequent error trigram in the corpus is the omission of the comma (see Table 2 above).

Error statistics and analyses were also used to adapt existing NLP tools (spellchecker and parser) and to turn them into an efficient error diagnosis tool. The resulting system does not cater for all categories of error. It can handle orthographic errors and quite a few grammatical errors (number and gender agreement, euphony, voice, etc.), some syntactic errors (adjective and adverb order), and a few lexical errors (complementation), but it cannot detect categories like punctuation errors, semantic lexical errors, and tense errors, which are beyond the capabilities of current parsers. The diagnostic tool allows learners to receive orthographic and grammatical feedback on their answers in the more open-ended CALL exercises. For orthographic errors, the program provides learners with a list of possible alternatives from which to choose, and, for grammatical errors, it identifies the nature of the error and refers learners to the relevant section in the hypertext grammar. (For more information on the error diagnosis system, see L'haire & Vandeventer in this volume.)

5. CONCLUSION

Learner corpus error annotation is a highly time-consuming and painstaking task. However, once the corpus is error-tagged, the return on investment is huge. An error-tagged learner corpus represents an unparalleled resource that gives researchers immediate access to detailed error statistics and that lends itself to automated error analysis, both of which have been used to fine tune the FreeText CALL program.

The three-tiered error annotation system designed for the project has proved to be very effective. The system contains a limited number of categories per tier (9 for domains, 36 for categories and 54 for word categories), which facilitates the annotator's task, while, at the same, allowing for relatively fine-grained analyses because of the numerous ways in which they can be combined (567 attested error trigrams in the corpus). Another advantage of the system is that it allows for versatile automated manipulation of the data. Using a text retrieval software tool such as *WordSmith Tools*, it is possible to retrieve any of the three levels of annotation, separately or conjointly, and to sort the concordance lines in a variety of ways in order to bring out recurrent error patterns. Though at first sight somewhat controversial, the decision to insert corrections in the text files also proved to be positive by making it possible to focus on one specific linguistic item (e.g., *que*) and to retrieve all the erroneous forms to which it can give rise (zero form—*ils trouvent *0 l'école est une expérience difficile*, *qui—des tâches *qui ne font pas les Français*, and *comme—pas aussi dur *comme chez eux*).

By limiting the number of tags and providing annotators with a comprehensive error-tagging manual, it has been possible to systematize error annotation to a large extent. However, it is important to realize that error annotation will always contain an element of subjectivity as the very notion of error is far from clear cut. As rightly pointed out by Milton and Chowdhury (1994, p. 129), "Tagging a learner corpus allows us, at least and at most, to systematize our intuitions." Error annotation is therefore intrinsically different from other less fuzzy types of annotation such as part-of-speech tagging.

It is also important to bear in mind that error tagging, in spite of its numerous advantages, is only concerned with learner misuse. It fails to uncover other aspects of interlanguage such as the under- and overuse of words and phrases, which together with downright errors contribute to the nonnativeness of learner productions. To access these frequency differences, it is not necessary to have an error-tagged learner corpus. A comparison between a raw, unannotated learner corpus and a comparable native corpus will automatically bring out the words and phrases that are significantly over- or underused by learners. For lack of time, we have only been able to use this approach sporadically in the FreeText project, albeit sufficiently to convince ourselves that it is the perfect complement to error annotation. When analyzing pronoun use, for example, we used the 'compare lists' facility in *WordSmith Tools*[9] and found that the pronoun *y* was significantly overused by FFL learners. A concordance-based analysis of the occurrences of the pronoun in the two corpora showed a massive overuse of the impersonal structure *il y a* as in *il y a beaucoup de tensions* 'there are a lot of tensions' in the learner corpus coupled with a clear underuse of anaphoric uses of *y*, as in *la Belgique refuse d'y prendre part* 'Belgium refuses to take part in it.'

Used discriminately and in full awareness of the above-mentioned limitations, an error-tagged corpus proves to be an invaluable tool to improve our knowledge of learner interlanguage and to adapt pedagogical materials—notably CALL programs—accordingly. Detailed error statistics make it possible to identify the most frequent error categories, while concordances of specific error types allow analysts to view the errors in context and to produce reliable descriptions of learner interlanguage. Within the framework of the FreeText project, error statistics and analyses were used to target the CALL exercises on learner-attested difficulties and to develop a learner corpus-informed error diagnostic tool capable of providing learners with automatic feedback on some of their most frequent errors. In addition, if the learner data are sufficiently well documented and the corpus is organized as a database, it is possible to customize the exercises in accordance with the learners' proficiency level and/or mother tongue background.[10]

While error-tagging may not be the be-all and end-all of interlanguage research, it provides highly valuable quantitative and qualitative insights into learner difficulties, which cannot fail to benefit all pedagogical foreign language learning tools, especially CALL programs.

NOTES

[1] The error rate is obviously in keeping with the learners' proficiency level. In our corpora of French as a Foreign Language, it ranges from 1 error every 5.2 words to 1 error every 9.5 words, with an average of 6.89.

[2] The project's full title is "French in Context: An advanced hypermedia CALL system featuring NLP tools for a smart treatment of authentic documents and free production exercises". The project receives financial support from the European Commission in the IST programme of the Fifth Framework Programme, contract IST-1999-13093. The project partners are the University of Geneva (Department of Linguistics), the University of Louvain (Centre for English Corpus Linguistics), the University of Manchester Institute of Science and Technology (Department of Language Engineering), and the French company Softissimo. More information on the FreeText project can be found on the project's web site (www.latl.unige.ch/freetext/index.html).

[3] Corrections are inserted between a hash and a dollar sign.

[4] Equally worthy of note is the fact that these 50 trigrams represent but 9% of the total number of different error trigrams attested in the corpus (N = 567).

[5] The symbol '0' is used to represent missing elements.

[6] The spellchecker is the spellchecking component of Softissimo's *Hugo 2000* (see www.softissimo.com/products/hugo.htm).

[7] The parser is the University of Geneva's *FIPS* parser (see Wehrli, 1997).

[8] For a more detailed analysis of participle errors, see Granger, Vandeventer, and Hamel (2001).

[9] The 'compare lists' facility in *WordSmith Tools* compares all the words in two word lists and reports all those which appear significantly more often in one than in the other. For this comparison, we used the FRIDA corpus and a corpus of essays written by French-speaking undergraduates in Romance languages at Louvain.

[10] Although L1-customization has not been implemented yet, it will be possible to do so in future because the FRIDA corpus is made up of three distinct subcorpora, covering three different categories of learners: native speakers of English, native speakers of Dutch, and learners from mixed mother tongue backgrounds.

REFERENCES

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-Aided Error Analysis. *System, 26* (2), 163-174.

Dulay, H., Burt, M., & Krashen, S. (1982). *Language Two.* New York: Oxford University Press.

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

Granger, S. (Ed.). (1998). *Learner English on computer*. London & New York: Addison Wesley Longman.

Granger, S., Meunier, F., & Watrin, P. (Forthcoming). Correction automatique du français langue étrangère : rêve ou réalité ? Analyse comparative de trois logiciels.

Granger, S., Vandeventer, A., & Hamel, M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO basé sur le TAL. *Traitement automatique des langues, 42* (2), 609-621.

James, C. (1998). *Errors in language learning and use. Exploring error analysis*. London & New York: Longman.

Milton, J., & Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. In L. Flowerdew & A. K. Tong (Eds.), *Entering text* (pp. 127-143). Hong Kong: The Hong Kong University of Science and Technology.

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Clevedon & Philadelphia: Multilingual Matters.

Scott, M. (1996). *WordSmith Tools.* Oxford: Oxford University Press.

Wehrli, E. (1997). *L'analyse syntaxique des langues naturelles: Problèmes et méthodes.* Paris: Masson.

APPENDIX A

Error Domains and Categories: Authentic Examples

<**F**>	
<**AGL**>	le **portebagages** (porte-bagages) est sur le toit
<**MAJ**>	Mme Thatcher incarnait la peur des **anglais** (Anglais)
<**DIA**>	Il existe une **ambiguité** (ambiguïté)
<**HOM**>	Ils **ce** (se) déshabillent
<**GRA**>	un **labirint** (labyrinthe)
<**M**>	
<**MDP**>	le sentiment de **malcontentement** (mécontentement)
<**MDS**>	… qui continue à **évolutionner** (évoluer)
<**MFL**>	Elle s'occupe des enfants et des **travails** (travaux)
<**MFC**>	Il n'a pas très bien **réussit** (réussi)
<**MCO**>	la participation de **celles-dernières** (celles-ci)
<**G**>	
<**CLA**>	L'unique chose **que** (qui) n'est pas bonne en Belgique...
<**AUX**>	Je m'**avais** (étais) très bien amusé
<**GEN**>	La protection sociale a été **amélioré** (améliorée)
<**MOD**>	Bien qu'ils **sont** (soient) pressés, …
<**NBR**>	Elle reprit ses **esprit** (esprits)
<**PER**>	J'espère **s'** (m') adapter rapidement
<**TPS**>	Un sondage qui **était** (a été) publié dans le Monde montre que …
<**VOI**>	Les éclipses **ont vues** (sont vues) comme des présages
<**EUF**>	En prenant **ce** (cet) aspect, ...

<**X**>

<**ORD**> Je peux **m’amuser bien** (bien m’amuser).

<**MAN**> Je crois **0** (qu’) ici il y a beaucoup plus de soirées

<**RED**> le domaine social **et** (,) économique et politique

<**COH**> A la métropole, il existe plus d’allocations et d’aide pour les chômeurs, les handicapés et les personnes âgées, **et** (tandis que) dans les îles, il n’y a pas beaucoup de soutien.

<**R**>

<**RLE**> **Quand même** (néanmoins), c’est une histoire lointaine

<**RSY**> Etant donné que **j’ai pas** (je n’ai pas) souvent présenté mes travaux …

<**Y**>

<**CLR**> Quand la famille d’un travailleur étranger vient le rejoindre, **on** (??) est obligé d’organiser toutes les formalités avant de quitter son pays.

<**LOU**> mais il y a des autres choses qui **m’ont donné une très grande surprise** (qui m’ont fort surpris)

<**Q**>

<**CON**> La langue devient plus française - (:) on l’appelle maintenant le créole francisé.

<**TRO**> Ce sont surtout les différences biologiques**,** (0) qui sont présentées dans la société avec la grande force de l’organisme humaine.

<**OUB**> Quand j’ai du temps libre **0** (,) je veux faire des choses reposantes …

<**Z**> **qunad** (quand); **ps** (pas)

APPENDIX B

Grammatical Categories

Grammatical Category		Tag
Adjective	Simple	ADJ
	Comparative	AJC
	Superlative	AJX
	Complex	AJL
Adverb	Simple	ADV
	Complex	AVL
Article	Definite	ADE
	Indefinite	AIN
	Partitive	APA
	Contracted	ACO
Conjunction	Coordinator	COC
	Simple subordinator	COS
	Complex subordinator	COL
Determiner	Demonstrative	DED
	Possessive	DEP
	Indefinite	DEI
	Exclamative-Interrogative	DEX
	Relative	DER
	Numeral	DEN
Noun	Common Simple	NOM
	Common Compound	NOC
	Common Complex	NOL
	Proper	NOP
Preposition	Simple	PES
	Complex	PEL

Grammatical Category		Tag
Pronoun	Demonstrative	POD
	Possessive	POP
	Personal	POO
	Indefinite	POI
	Exclamative-Interrogative	POX
	Numeral	PON
	Adverbial	POA
	Relative	POR
	Impersonal	POS
Verb	Finite simple	VSC
	Participle simple	VSP
	Gerund simple	VSG
	Infinitive simple	VSI
	Finite complex	VCC
	Participle complex	VCP
	Gerund complex	VCG
	Infinitive complex	VCI
Punctuation	Period	PUP
	Question mark	PUI
	Exclamation mark	PUE
	Comma	PUV
	Semi-colon	PUG
	Colon	PUD
	Suspension periods	PUS
	Parentheses	PUA
	Square brackets	PUC
	Quotation marks	PUL
	Dash	PUT
	Slash	PUO
Sequence		SEQ

AUTHOR'S BIODATA

Sylviane Granger is Professor of English Language and Linguistics at the University of Louvain (Belgium). She is the director of the *Centre for English Corpus Linguistics* where research activity is focused on the compilation and exploitation of learner and bilingual corpora. Her publications include *Learner English on computer* (Ed.) (Longman, 1998) and *Computer learner corpora, second language acquisition and foreign language teaching* (Granger, Hung, & Petch-Tyson (Eds.). Benjamins, 2002).

AUTHOR'S ADDRESS

Professor Sylviane Granger
Centre for English Corpus Linguistics
University of Louvain
Place Blaise Pascal 1
B-1348 Louvain-la-Neuve
Belgium
Phone: +32 10 474947
Fax: +32 10 474942
Email: granger@lige.ucl.ac.be