Système d'annotation du corpus d'apprenants roumains de FLE*

Mariana-Diana Câșlaru**

Error Tagging System for Romanian Corpus of FFL Learners

Abstract:
The interlanguage – dynamic and systematic language at the same time – is still a challenge for the researchers. Nowadays, we try to build software for the computer based corpus analysis and, in this sense (i.e. Free Text project), tagging tools for the learners' errors are developed. This paper presents an original error tagging system built on a French corpus written by Romanian and highlights the benefits of such an annotation.
Keywords: tagging system, tag, interlanguage, error, conformity with the norm

« Pour le linguiste, les textes sont une suite d'événements singuliers,
idiosyncrasiques et désordonnés, à travers lesquels
il cherche un ordre sous-jacent ». (Kraif, 2011 : 68)

Les productions des apprenants de langue étrangère sont à la fois systématiques et dynamiques et sont représentatives des systèmes provisoires qu'un apprenant développe en étudiant une langue seconde. Ces étapes transitoires n'ont pas de norme propre, mais on les juge par rapport à la norme de la langue cible[1]. Toute structure qui n'est pas conforme à la

* Paper presented at the International Conference "Romanian as a Foreign Language", 4th edition, Iași, October 30–31, 2015, organized in the framework of the Days of "Alexandru Ioan Cuza" University of Iași. The topic of the conference: *Romanian Language and Identity in the Continuous Cultural Reconfiguration of Europe*.

** Assistant Lecturer PhD, "Alexandru Ioan Cuza" University of Iași, dianacaslaru@yahoo.fr

[1] Il ne faut pas ignorer le fait que l'interlangue peut être analysée aussi dans ses propres termes, ce que Nemser et Corder précisent lorsqu'ils définissent l'interlangue : « La parole de l'apprenant devrait être étudiée non seulement par référence à la langue source et à la langue cible mais aussi dans ses propres termes » (Nemser, 1971 : 116) et « Il nous faut tenter de décrire cette langue dans ses propres termes, du moins dans un premier temps, et non pas en termes d'une autre langue » (Corder, 1980c : 29). Voir aussi l'article de Mariana-Diana Câșlaru, « La complexité et la conformité de l'interlangue des apprenants roumains de FLE », *Annals of 1 Decembrie 1918 University of Alba Iulia – Philology*, 14/3/2013, p. 483–496.

norme représente une erreur qui n'est plus considérée signe de l'enseignement ratée, mais indice de l'activité d'esprit de l'apprenant, hypothèse que l'apprenant a formulée sur le système linguistique de la langue cible et qui n'a pas été validée.

Afin d'analyser la conformité d'un système transitoire/ provisoire/ avec la norme et, par conséquent, de calculer la distance entre celui-ci et la norme de la langue cible, de trouver l'ordre dans le chaos d'une interlangue, il faut établir un système d'annotation qui puisse rendre le corpus exploitable.

Nous présentons dans cet article un système d'annotation sur trois niveaux qui a été employé dans une étude de cas sur les productions écrites des apprenants roumains de FLE qui étudiaient aussi l'italien et l'espagnol.Cet étiquetage nous permet de passer par les trois étapes de l'analyse d'erreurs telle qu'elle a été structurée par Pit Corder (1980a, 1980b). Le premier niveau d'annotation, qui décide de quelle catégorie et de quel type d'erreur il s'agit, correspond à l'identification et à la description de la forme erronée, tandis que les deux derniers niveaux, qui rendent compte de la source de l'erreur, correspondent à l'étape de l'explication de l'erreur.

Le premier niveau d'annotation

Pour le premier niveau, nous proposons un étiquetage des erreurs selon la typologie traditionnelle : *lexique, morphologie et syntaxe*. Chacune de ces catégories comprend plusieurs types d'erreurs. Le tableau ci-dessous a été inspiré par plusieurs études (Granger 2003, L'Haire et Vandeventer 2003, L'Haire 2007, Thouesny 2011), mais l'organisation des catégories et des types nous appartient. Nous avons construit le tableau de manière graduelle, parallèlement avec le travail d'annotation, en fonction des situations rencontrées.

Les objectifs formulés à ce niveau sont d'identifier les erreurs et de déterminer la nature de celles-ci.

Catégorie d'erreur		Type d'erreur	
<L>	Lexique	DIA	Diacritiques
		GRA	Graphie, forme
		MNS	Motnon standard
		MOT	Mot inapproprié
<M>	Morphologie	AUX	Auxiliaire
		CLA	Classe
		CON	Conjugaison
		EUF	Euphonie
		FLE	Flexion/déclinaison du pronom
		GEN	Accord genre
		MOD	Mode inapproprié
		NBR	Accord nombre

		TEM	Temps inapproprié
		VOI	Voix
<S>	Syntaxe	CHO	Mauvais choix (préposition, conjonction)
		OMS	Omission
		ORD	Ordre des mots/ syntaxe incorrecte
		SUP	Superflu

Tableau 1 – Catégories d'erreurs

Dans ce qui suit, nous donnerons quelques exemples d'erreurs pour chacune de ces catégories. Mentionnons d'emblée que cette présentation ne comprendra que des explications adéquates à ce niveau d'annotation. La distinction intralinguale/interlinguale (<O>, <IC>, <IS>) à l'intérieur de chacune de ces catégories (MNS, MOT, VOI, SUP, etc.) sera traitée au moment opportun.

- L'étiquette DIA désigne les erreurs d'accent, à savoir les mots qui ne reçoivent pas les accents imposés par la norme, ou ceux qui reçoivent des accents mal positionnés.

<IC><L><DIA><NOM>*ocean/océan

<O><L><DIA><NOM>*fôret/forêt

- L'étiquette GRA désigne les erreurs de graphie (a et b) ou les erreurs due à l'homophonie (c). Cette catégorie comprend aussi les erreurs de forme (d) :

a) <O><L><GRA><NOM>*perssonages/personnages

b) <IC><L><GRA><NOM>*discution/discussion

c) <IC><L><GRA><VBF>*e/est

d)<O><L><GRA><DED>*cet/ce temps

- L'étiquette MNS désigne les mots qui n'existent pas en français standard. Il s'agit soit d'une substitution (a), soit d'un calque trop éloigné du terme français standard qu'il veut remplacer (b et c) :

a) <IS><L><MNS><NOM>*window/fenêtre (substitution, anglais)

b) <IS><L><MNS><NOM>*vecin/voisins (substitution, roumain)

c) <IC><L><MNS><NOM>*felinaire/lanterne (calque du roumain *felinar*)

- L'étiquette MOT désigne les mots qui existent en français mais qui, du point de vue sémantique, sont inappropriés. Cette erreur peut être produite parfois sous l'influence des autres langues connues par le sujet (exemples c et d) :

a) <O><L><MOT><NOM>*tables/tableau,

b) <O><L><MOT><NOM>*pêches/poissons,

c) <IC><L><MOT><NOM>*chemin/foyer (du ro. *cămin* = fr. *foyer*)

d) <IC><L><MOT><NOM>*repas/repos (du ro. *repaos* = fr. *repos*)

Cette étiquette désigne aussi les cas d'erreurs intralinguales, seulement lorsque le sujet a utilisé un mot qui fait partie de la même classe grammaticale du mot correct, comme dans les exemples ci-dessus. Par contre, si le mot proposé fait partie d'une autre classe grammaticale, alors il s'agit une erreur de morphologie.

<O><M><CLA><ADV>*bonne/bien

- L'étiquette AUX désigne la confusion des auxiliaires nécessaires à la formation des temps passés.

<IC><M><AUX><VBF>*a/est apparue

<O><M><AUX><VBF>*était/avait participé

- L'étiquette CLA désigne les erreurs de classe grammaticale :

<O><M><CLA><VBF>*promenade/promène

<O><M><CLA><SUB>*qui/que

<IC><M><CLA><ADV>*grave/gravement

- L'étiquette CON désigne les terminaisons inexistantes dans la conjugaison des verbes (a) ou la confusion des terminaisons appropriées pour chaque personne (b)

a) <O><M><CON><VBF>*finie/finit

b) <O><M><CON><VBF>*avais/avait

- L'étiquette EUF désigne l'absence de l'élision ou de la contraction :

<O><M><EUF><ADE>*de le/du

<O><M><EUF><POO>*je/j'entre

<O><M><EUF><DEP>*sa/son amie

- L'étiquette FLE désigne l'erreur de déclinaison en fonction du cas. <O><M><FLE><POO>*lui/le

<O><M><FLE><POR>*qui/que.

- L'étiquette GEN désigne la confusion entre les formes de masculin et féminin.

<IC><M><GEN><ADE>*le/la

<O><M><GEN><ADJ>*vert/verte

<O><M><GEN><VBP>*colorées/colorés

<IC><M><GEN><POR>*lequel/laquelle

• L'étiquette MOD désigne les modes inappropriés (a), et les verbes non conjugués (d) :

a) <O><M><MOD><VBF>*avaient/aient

b) <O><M><MOD><VBI>*viens/venir

c) <O><M><MOD><VBP>*manger/mangé

d) <O><M><CON><VBF>*vivre/vivent

• L'étiquette NBR désigne la confusion entre les formes de singulier et celles de pluriel (a et b). De même, cette catégorie comprend aussi les cas des déterminants possessifs lorsque les sujets ont confondu les unipossessifs avec les pluripossessifs (c et d) :

a) <O><M><NBR><ACO>*au/aux

b) <O><M><NBR><NOM>*branche/branches

c) <O><M><NBR><DEP>*leur/leurs

d) <O><M><NBR><DEP>*notre/nos

• L'étiquette TEM désigne les temps inappropriés :

<O><M><TEM><VBF>*assistent/assistaient

<O><M><TEM><VBF>*donnait/donne

• L'étiquette VOI désigne les cas où le sujet confond la voix active et la voix passive. Mentionnons que dans notre analyse nous tiendrons compte de l'indication de Grevisse (2011) selon laquelle la voix réflexive n'est qu'un cas particulier de la voix active. Par conséquent, l'erreur du type *il *se joue/ il joue* n'est pas une erreur de morphologie concernant la voix, mais une erreur de syntaxe concernant un élément superflu dans la phrase. Notre corpus ne contient pas d'erreurs de ce type.

• L'étiquette CHO désigne le mauvais choix des prépositions ou des conjonctions. Nous avons placé ces erreurs dans la catégorie *syntaxe*, car ces mots établissent des relations entre les unités d'une phrase/proposition et ne peuvent entrer dans la phrase que joints à d'autres mots (Benveniste, 1966 : 125).

Partir <O><S><CHO><SUB>*de/pour les rencontrer

Commence <O><S><CHO><PRE>*de/à danser

• L'étiquette OMS désigne l'omission de certaines parties de la phrase : <IC><S><OMS><ADE>*0/la (maison)

<O><S><OMS><POO>*0/me (sentir)

• L'étiquette ORD désigne l'ordre inapproprié des mots :

<IC><S><ORD><ADJ>*bleues fleurs/fleurs bleues

<IC><S><ORD><POO>*aussi elle/elle aussi

• L'étiquette SUP désigne les parties de la phrase qui sont superflues :

<O><S><SUP><PRE>*dans/0 (le matin)

IC><S><SUP><POO>*se/0 (jouer)

- Même si les étiquettes OMS et SUP désignent la redondance ou l'absence d'un morphème grammatical, nous les avons placées dans la catégorie Syntaxe, car elles représentent des « erreurs locales à effets secondaires » (Boissière *et alii,* 2007 : 3) en influençant le bon fonctionnement de l'énoncé.

Toujours à ce niveau d'annotation, nous précisons aussi la catégorie grammaticale de la forme erronée, selon le tableau suivant que nous avons construit d'après le modèle proposé par l'équipe du projet Free Text de l'Université de Louvain, en vue de l'annotation du corpus FRIDA, et décrit par Silviane Granger (2003 : 479).

Catégorie grammaticale		Étiquette
Adjectif		ADJ
Adverbe		ADV
Article	Défini	ADE
	Indéfini	AIN
	Partitif	APA
	Contracté	ACO
Conjonction	Coordination	COC
	Subordination	SUB
Déterminant	Démonstratif	DED
	Possessif	DEP
	Exclamatif/ interrogatif	DEX
	Relatif	DER
	Numéral	DEN
	Indéfini	DEI
Nom		NOM
Préposition		PRE
Pronom	Démonstratif	POD
	Possessif	POP
	Personnel	POO
	Indéfini	POI
	Exclamatif/ Interrogatif	POX
	Numéral	PON
	Adverbial	POA
	Relatif	POR
	Impersonnel	POS
Verbe	Prédicatif	VBF
	Participe	VBP
	Gérondif	VBG
	Infinitif	VBI
Séquence		SEQ

Tableau 2 Catégories grammaticales

Le second niveau d'annotation

Au second niveau de l'étiquetage, nous reprenons toutes les erreurs annotées avant et décidons de leur source. Cette fois-ci, l'objectif est d'expliquer l'erreur et d'inférer les éléments qui l'ont déterminée. Cette fois-ci, nous utilisons la typologie de C. Richards (1970) qui fait la distinction entre l'erreur *interlinguale*, d'une part, et l'erreur *intralinguale* et *développementale*, d'autre part. L'erreur interlinguale est causée par l'interférence de la langue maternelle dans la langue cible : l'apprenant utilise une règle de la langue source lorsqu'il produit des énoncés en langue cible.

Par contre, l'erreur intralinguale et l'erreur développementale ne reflètent pas l'incapacité de l'apprenant à séparer deux langues, mais montrent la compétence de l'apprenant à un certain moment de l'acquisition. L'origine de ces erreurs se trouve dans la structure de la langue cible même et est liée aux stratégies que l'apprenant emploie afin de l'acquérir. Les erreurs intralinguales se produisent à cause de la généralisation erronée, de l'application incomplète des règles et de la méconnaissance des conditions qui régissent l'application d'une certaine règle (Richards, 1970 : 6). Les erreurs développementales montrent que l'apprenant essaie de construire des hypothèses sur la langue cible selon son expérience limitée. Ces erreurs sont causées par la compréhension erronée des distinctions à l'intérieur de la langue cible.

Comme la langue cible est à la fois source des erreurs intralinguales et des erreurs développementales, lors du second niveau de l'étiquetage des erreurs, nous garderons la dénomination *erreur intralinguale* pour les deux.

La pratique nous a mené à détailler les types d'erreurs à ce niveau. Certaines erreurs sont vraiment difficiles à insérer dans une rubrique précise. Dans ces cas, nous avons ajouté le type « *erreurs ambiguës* » proposé par Dulay et Burt (1974).

Au troisième niveau, l'étiquetage s'opère seulement sur les erreurs interlinguales, en les classifiant en deux catégories : *substitutions* et *calques*, selon la typologie de Trence Odlin (1989). Les *substitutions* impliquent l'usage des formes de la langue maternelle de l'apprenant dans les énoncés produits en langue étrangère. L'apprenant remplace les structures de la langue cible qui lui sont encore inconnues par des structures de la langue maternelle. T. Odlin (1989 : 37) propose l'exemple du mot suédois *bort* qui signifie *loin* (en français) et *away* (en anglais). Un natif suédois apprenant d'anglais a écrit : « Now I live home with my parents. But some-times I must go bort ».

Les *calques* sont des erreurs qui reflètent une structure de la langue maternelle. T. Odlin (1989 : 37) propose l'exemple d'un enfant bilingue

(anglais – espagnol) qui a dit : « Let's quikly put the fire out / Vamos rapido a poner el fuego afuera. » Il a fait la traduction littérale de l'anglais « put the fire out » qui normalement se traduit « extinguir el fuego ». L'ordre erroné des mots peut être aussi le résultat d'un calque selon Dong Juan et Han Ge-ling (2009 : 12). Ils donnent l'exemple d'un apprenant chinois qui a écrit en anglais : « The taxi driver was the only responsible person for the accident ». Le sujet transfère la structure chinoise *attribut + nom* en anglais où l'attribut doit suivre le nom, dans ce cas.

Description de l'étiquette

En tenant compte de ce système d'annotation, nous avons fait l'étiquetage manuellement avec le logiciel Microsoft Word. Au premier niveau d'annotation, l'étiquette comporte trois acronymes porteurs d'information concernant le type d'erreur : le premier fait référence à la catégorie de l'erreur (lexique, morphologie, syntaxe), le second montre le type de l'erreur (mode inapproprié, genre, graphie, ordre des mots, etc.) et le troisième précise la catégorie grammaticale de la forme erronée (voir tableau 2). Chaque acronyme est encadré par des flèches pour des raisons d'efficience au moment du traitement automatique du corpus, les logiciels étant, de cette manière, capables de reconnaitre mieux des éléments demandés.L'étiquette précède l'erreur qui est encadrée par une étoile (*) et une barre oblique (/). Voici un exemple :

Le <L><GRA><NOM>*perssonage/personnage <S><CHO><PRE>*dans/de l'image est la princesse qui est très belle et elle aime beaucoup <L><GRA><VBI>*joyer/jouer avec les <L><GRA><NOM>*extratcrestres/extraterrestres, parce qu'elle croit qu'ils <M><CON><VBF>*exist/existent et dans <L><GRA><DED>*cet/ce moment elle

Figure 1 – Exemple d'étiquetage : niveau 1

Ensuite, au second niveau de l'annotation, nous ajoutons à l'étiquette une lettre supplémentaire qui fait référence à la source de l'erreur en cause, à savoir I = erreur interlinguale, O = erreur intralinguale et B = erreur ambiguë.

Elle arrive au <I><L><MNS><NOM>*camin/foyer, entre dans l'<I><L><GRA><NOM>*apartament/appartement et parle au <I><L><DIA><NOM>*telephone/téléphone <I><L><MNS><ADV>*again/de nouveau. Elle part et arrive <O><M><EUF><ACO>*au/à l'<O><L><DIA><NOM>*hopital/hôpital

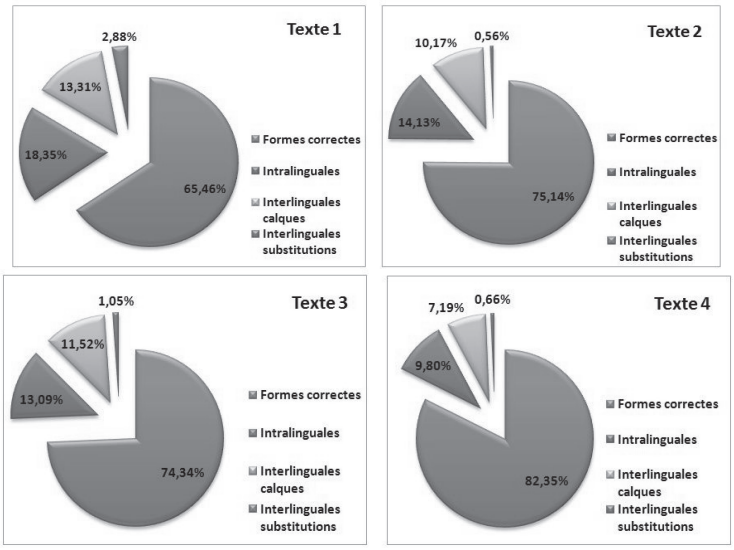
Figure 2 – Exemple d'étiquetage : niveau 2

Enfin, le troisième niveau d'étiquetage, qui s'opère seulement sur les erreurs interlinguales, dispose à son tour des lettres symboles pour désigner le calque (C) ou la substitution (S) :

Elle arrive au <IS><L><MNS><NOM>*camin/foyer, entre dans
l'<IC><L><GRA><NOM>*apartament/appartement et parle au
<IC><L><DIA><NOM>*telephone/téléphone <IS><L><MNS><ADV>*again/de nouveau.
Elle part et arrive <O><M><EUF><ACO>*au/à l'<O><L><DIA><NOM>*hopital/hôpital po

Figure 3 – Exemple d'étiquetage : niveau 3

Une fois le corpus annoté, on peut en tirer profit et calculer la conformité de l'interlangue avec la norme[2] ou quantifier l'influence que telle ou telle type d'erreur a sur l'interlangue d'un certain sujet à un moment donné, comme dans l'exemple suivant :



[2] Pour voir les formules, consulter la thèse de Mariana-Diana Câşlaru, *L'interlangue des apprenants roumains de FLE au carrefour des langues romanes (études de cas sur des apprenants roumains étudiant aussi l'italien et l'espagnol),* http://www.theses.fr/2013AVIG1128.

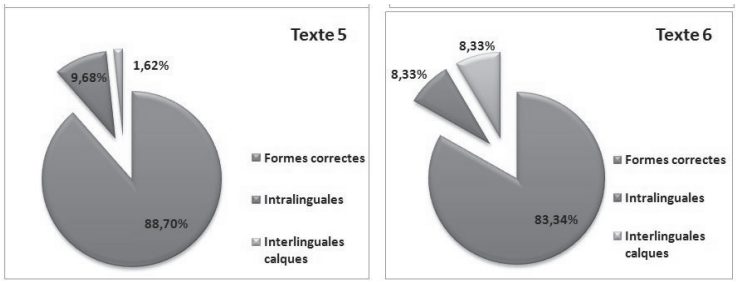


Figure 4 - Pourcentage des erreurs selon leur source

Il s'agit d'un sujet qui a produit six textes écrites à des moments différents, au cours d'une année. Selon les diagrammes ci-dessus, l'interlangue du sujet en cause présente des erreurs interlinguales qui affectent moins sa conformité avec la norme que les erreurs intralinguales. Le premier enregistre le nombre le plus grand de substitutions. Dans les quatre premiers textes, le nombre des substitutions continue à varier. Par contre, elles sont absentes dans les textes 5 et 6, ce qui traduit le fait que le sujet ne se contente plus de combler les vides de ses connaissances linguistiques en introduisant des mots étrangers tels quels, mais qu'il commence à vérifier des hypothèses ; il en résulte les calques.

Conclusion

Sans doute, comme l'on a déjà dit maintes fois[3], l'annotation du corpus suivie par l'analyse des erreurs et de leur influence sur la conformité de l'interlangue avec la norme, ne représente pas une recherche exhaustive, elle n'est que la première étape dans l'analyse de la langue de l'apprenant.Le système transitoire est beaucoup plus que sa conformité avec la norme d'un autre système linguistique (celui de la langue cible) ; elle a aussi une autre dimension que nous appelons complexité et dont l'analyse est beaucoup plus minutieuse à faire, mais qui enrichit et raffine beaucoup l'étude d'une interlangue.

Cependant, l'annotation d'un corpus d'apprenant est indispensable à l'analyse ultérieure de celui-ci. Dans un état initial, d'habitude, l'interlangue de l'apprenant comprend des éléments de la langue cible mais aussi des éléments de la langue maternelle ou des autres langues étrangères connues par le sujet, ou même des éléments d'origine ambigüe. Etablir un système d'annotation et étiqueter les éléments

[3] Voir Câşlaru Mariana-Diana, *La complexité et la conformité de l'interlangue des apprenants roumains de FLE* , « Annals of 1 Decembrie 1918 University of Alba Iulia – Philology », 14/3, 2013, p. 483–496.

d'interlangue selon ce système nous permet d'extraire l'ordre du désordre et d'analyser de manière quantitative, mais aussi qualitative, les aspects difficiles dans l'apprentissage d'un certain sujet.

RÉFÉRENCES :

Boissière, Ph., Bouraoui, J.-L., Vella, F., Lagarrigue, A., Mojahid, M., Vogouroux, N., Nespoulous, J.-L, *Méthodologie d'annotation des erreurs en production écrite. Principes et résultats préliminaires,* TALN, Toulouse, 12–15 juin, 2007.

Câşlaru Mariana-Diana, *L'interlangue des apprenants roumains de FLE au carrefour des langues romanes (études de cas sur des apprenants roumains étudiant aussi l'italien et l'espagnol)*, http://www.theses.fr/2013AVIG1128, 2013.

Câşlaru Mariana-Diana, *La complexité et la conformité de l'interlangue des apprenants roumains de FLE*, in « Annals of 1 Decembrie 1918 University of Alba Iulia – Philology », 14/3, 2013, p. 483–496.

Corder, Pit, *Que signifient les erreurs des apprenants ?*, « Langages » 57, Apprentissage et connaissance d'une langue étrangère, 1980a, p. 9–15.

Corder, Pit, *Dialecte idiosyncrasique et l'analyse d'erreurs,* « Langages » 57, Apprentissage et connaissance d'une langue étrangère, 1980b, p. 17–28.

Corder, Pit, *La sollicitation de données d'interlangue*, « Langages » 57, Apprentissage et connaissance d'une langue étrangère, 1980c, p. 29–38.

Dong, Juang et Han, Ge-ling, *Negative Transfer in Chinese College Students' English Writing*, « Sino-US English Teaching », 6/8, 2009, p. 8–25.

Dulay, Heidi et Burt, Marina, *You Can't Learn Without Goofing. An Analysis of Children's Second Language* ***Errors***, dans J. Richards (ed.), *Error Analysis: Perspectives on Second Language Acquisition*, London, Longman, 1974, p. 95–123.

Granger, Sylviane, *Error-tagged learner corpora and CALL: a promising synergy*, « CALICO », 20/3, 2003, p. 465–480.

Kraif, Olivier, *Les concordances pour l'observation des corpus : utilité, outillage, utilisabilité,* dans J. Chuquet (dir.), *Le langage et ses niveaux d'analyse- cognition, production des formes, production du sens*, Presses Universitaires de Rennes, 2011, p. 67–79.

L'Haire, Sébastien, *FipsOrtho : A spell checker for learners of French,* « ReCALL », 19/2, 2007, p. 137–161.

L'Haire, Sébastien, Vandeventer, Faltin, Anne, *Error Diagnosis in the FreeText Project*, « Calico Journal », 20/3, 2003, , p. 481–495.

Nemser, William, *Approximative systems of foreign language learners*, « International review of applied linguistics in language teaching », IX, 1971, p. 115–124.

Odlin, Terence, *Language transfer: cross-linguistic influence in language learning*, Cambridge University Press, 1989.

Richards, Jack C., *A Non-Contrastive Approach to Error Analysis*, Document présenté à la Convention TESOL, San-Francisco, 1970.