Recueil, traitement et analyse des corpus d'apprenants : Défis méthodologiques et enjeux

Gaëtanelle Gilquin,
Université catholique de Louvain

JECA, 18 mars 2022

1

Quelle est la première question à se poser avant de commencer à constituer un corpus (d'apprenants) ?

"Existe-il un corpus qui répond à mes besoins ?"

2

Une multitude de corpus d'apprenants

- Les corpus d'apprenants ont commencé à apparaître au début des années 1990 (*International Corpus of Learner English*, *Longman Learners' Corpus*)
- Depuis lors, de nombreux corpus d'apprenants ont été constitués, dont certains ont été rendus accessibles à la communauté
- Majorité de corpus écrits et de corpus d'anglais (L2), mais d'autres types de corpus disponibles

3

Learner corpora https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Use the query box below to search for specific keywords (e.g. languages, task type, medium).

Show All entries Search: French written

Corpus	Target language	First language	Medium	Text type / task type	Proficiency level	Size in words	Project director	Availability
The "Dire Autrement" corpus	French (Second Language)	Mainly L1 speakers of English	Written	Narrative, injunctive, persuasive and informative texts		c. 50,000	Marie-Josée Hamel, Jasmina Milicevic (Dalhousie University, Canada)	Available after registration
French Interlanguage Database (FRIDA)	French	Various	Written	Free compositions: descriptive, argumentative and narrative texts, news & mail	Intermediate		Sylviane Granger (Centre for English Corpus Linguistics, Université catholique de Louvain, Belgium)	
The Learner Corpus French (LCF)	French	Dutch	Written	Argumentative essays, informative texts, journalistic texts, formal letters, summaries, written compositions by Flemish students of French	Intermediate to advanced	c. 500,000	Hans Paulussen (K.U.Leuven/Ugent /Lessius, Belgium)	Under development
The Lund CEFLE Corpus (Corpus Écrit de Français Langue Étrangère)	French	Swedish	Written	Descriptive and narrative essays, picture-based stories	Various	c. 100,000	Malin Ägren (Lund University, Sweden)	A sub-part of the corpus is available online.

4

La constitution de corpus d'apprenants

- Nécessaire si le corpus souhaité n'existe pas encore, par exemple :
  - langue (L1/L2) pas/peu représentée
  - variété très spécifique (ex. production écrite d'apprenants de l'anglais en pharmacologie)
  - variété locale (ex. travaux de ses propres étudiants de Master)
- Importance de la constitution des corpus : 'Garbage in, Garbage out' (cf. Sinclair 1991, p.9)

5

Les étapes

1. Recueil
2. Traitement
3. Analyse

- Chacune de ces étapes peut représenter des défis méthologiques majeurs (accent sur ce qui fait la spécifité des corpus d'apprenants)
- Illustration avec certains des projets de corpus d'apprenants initiés au CECL : ICLE, ICLE+30, LONGDALE, PROCEED, LINDSEI
- Surtout l'anglais, mais applicable à d'autres langues

6

1 RECUEIL

7

L'objectif

- Constituer un corpus dans un but bien précis – et uniquement dans ce but (ex. corpus pour étudier l'impact d'une pratique pédagogique)
- Créer une ressource polyvalente qui servira à plusieurs chercheurs, projets, etc. et donc qui devra pouvoir répondre à une grande variété de questions de recherche (ex. PROCEED)

8

8

Le contexte

- A la différence d'autres corpus écrits comme les corpus de presse, les corpus d'apprenants sont presque toujours recueillis en simultané
- Les corpus d'apprenants sont souvent recueillis par les enseignants-chercheurs dans leur propre environnement (contexte universitaire)
- Des projets collaboratifs permettent de multiplier les contextes et donc les profils d'apprenants (cf. ICLE : 30 équipes internationales)

9

9

La représentativité

- Presque toujours nécessaire d'échantillonner la population d'apprenants
- Souvent échantillonnage de commodité, pas nécessairement représentatif de la population
  - Étudiants volontaires
  - Étudiants d'un certain cours/enseignant
- Certains corpus visent un échantillonnage plus large, par ex. grâce à la collecte de données en ligne (cf. CEDEL2)

10

10

Le recrutement

- Difficulté de recruter des apprenants volontaires en-dehors du cadre des cours
- Incitants : « As incentive for participation, we're going to hold a lottery. The prizes are 2 gift certificates from the local bookshop, worth 500 kr each »
- Fiabilité des volontaires : « I have to track down a student who didn't show up yesterday »
- Recueillir la permission des apprenants (RGPD)

11

11

L'authenticité et les genres de l'écrit académique

- Un corpus d'apprenants n'est presque jamais entièrement authentique (dans le sens de "people going about their normal business", Sinclair 1996)
- Seuls certains genres académiques peuvent être produits par des apprenants d'un profil défini (en particulier, selon leur niveau de langue)
- Éviter de collecter un genre de texte parmi des apprenants qui n'ont pas l'habitude de produire ce genre de texte

12

12

Le cas des corpus longitudinaux

- Projets de longue durée (financement, personnel)
- Perte de participants (attrition)

Trajectory	Number of students
Y1	86
Y1–Y2	17
Y1–Y2–Y3	66
Y1–Y3	15
Y2	10
Y2–Y3	16
Y3	27
Total	237

LONGDALE (Paquot et al. 2021, p.129)

13

Les métadonnées

- Indispensable de collecter des métadonnées riches et détaillées, étant donné l'hétérogénéité des profils d'apprenants et l'influence de nombreuses variables sur la langue d'apprenants (Ellis 1994, Granger 1998)
- Sélection de variables potentiellement pertinentes, cf. Lozano (2021) (ex. interviewer dans LINDSEI)
- Niveau de compétence : si possible inclure le résultat d'un test standardisé, ex. LexTALE (limite des indicateurs de substitution, cf. Myles 2015, p.316)
- Impossible d'inclure toutes les variables pouvant influencer la langue d'apprenants (Granger 2004, p.126)

14

> "Corpus-building is of necessity a marriage of perfection and pragmatism" (McEnery et al. 2006, p.73)

15

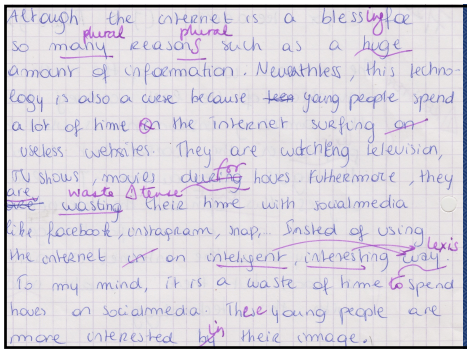
2 TRAITEMENT

16

La numérisation

- Aujourd'hui, de nombreux corpus d'apprenants rassemblent des textes produits directement sous format électronique (cf. ICLE+30)
- Certains contextes où l'écriture à la main est encore de mise (ex. apprenants débutants, examens) => travail manuel de transposition
- Crucial de reproduire le texte exactement à l'identique – y compris les formes non-standard

17


Although the internet is a blessing for so many reasons such as a huge amount of information. Nevertheless, this technology is also a curse because young people spend a lot of time on the internet surfing on useless websites. They are watching television, TV shows, movies during hours. Furthermore, they are wasting their time with socialmedia like facebook, instagram, snap. Instead of using the internet in an intelligent, interesting way. To my mind, it is a waste of time to spend hours on socialmedia. These young people are more interested by their image.

18

Although the internet is a bless for so many reason such as a huge amount of information . Nevertheless , this technology (techno-logy?) is also a curse because young people spend a lot of time in the internet surfing on useless websites. They are watching television, TV shows , movies during hours . Furthermore , they are wasting their time with socialmedia (?) like facebook, instagram, snap,... Instead of using the internet in an (on?) intelligent , interesting way. To my mind, it is a waste of time spend hours on socialmedia (?). This young people are more interested by their image.

19

19

L'anonymisation : textes

- So we got aquatinted a little bit . " Hi , my name is <name> and I went to high school in <place> . " (ICLE-BG)
- The people who only saw my name often thought that I am a boy . (ICLE-KR)
- The next day there I was sitting in class , I believe it was called Integrated Skills Course or some similar quasi intellectual and pretentious name . A very charismatic assistant professor was sharing her excitement , she was so glad to have us all there and boy did she sound British ! (ICLE-BG)

20

20

L'anonymisation : métadonnées

- Supprimer le nom n'est pas toujours suffisant, cf. Rocher et al. (2019) : identification d'un individu sur base de 16 attributs (dont la nationalité)
- Données potentiellement sensibles, ex.
  - Langue maternelle : « if someone answers Kurdish or Sami, then that will tell us their ethnicity »
  - Niveau de compétence : « It could be easily turned into a ranking list of the different subcorpora and even the learners from different countries »

21

21

Le balisage (1): les citations

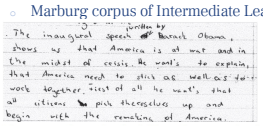
- Importance de baliser tout texte écrit par autrui (⇔ corpus d'écrits académiques d'experts)
- Exemple:
  - After about two hundred years an English writer George Orwell extended the statement in his novel Animal Farm : <quote> " All men are created equal but some are more equal than others . " </quote> (ICLE-CZ)

22

22

Le balisage (2): les révisions

- Marburg corpus of Intermediate Learner English (Kreyer 2015)



The inaugural speech <d>of</d> <ad>, written by</ad> Barack Obama, <lb></lb> shows us that America is at war and in <lb></lb> the midst of crisis. He want's to explain, <lb></lb> that America need to stick as well as to <lb></lb> work together. First of all he want's that <lb></lb> all citizens <d><ur>2 letters</ur></d> pick theresel ves up and <lb></lb>begin with the remaking of America.

Key: **d**: deleted text; **ad**: added text; **lb**: line break; **ur**: unreadable text.

23

23

L'annotation (1)

- La plupart des outils d'annotation actuels ont été créés pour annoter les variétés écrites natives / standard des langues (Gilquin & De Cock 2011)
- Exemple : étiquetage morpho-syntaxique d'ICLE avec CLAWS
  - I could realized<VVD> this fact (ICLE-GE)
  - Nowaday<NN1>, many educational programmes … (ICLE-CN)
  - … dreadful punishment againts<VVZ> Joan (ICLE-SP)

24

24

L'annotation (2)

- Résultats de l'étiquetage morpho-syntaxique et de l'analyse syntaxique de corpus d'apprenants généralement assez satisfaisants
  - Structure relativement simple des phrases (Meunier 1998, Huang et al. 2018)
  - Meilleurs résultats pour des niveaux de compétence avancés (Geertzen et al. 2014)
  - Peut varier selon les textes : 95%-99.1% de précision avec CLAWS sur ICLE (Granger et al. 2009, p.16)
  - Peut varier selon les catégories, ex. 100% pour JJR vs 33% pour VB0 avec CLAWS sur LINDSEI

25

25

L'annotation (3): les erreurs

- L'annotation automatique d'erreurs est pour l'instant limitée à des phénomènes spécifiques, ex. prépositions (De Felice & Pulman 2009), articles (Rozovskaya & Roth 2010), orthographe (Rayson & Baron 2011)
- Comment identifier les erreurs ?
  - Études pilotes autour de l'identification d'erreurs par des locuteurs natifs : parfois plus du double d'erreurs identifiées par texte (Dagneaux et al. 2008, p.7)

26

26

L'annotation (4): les erreurs

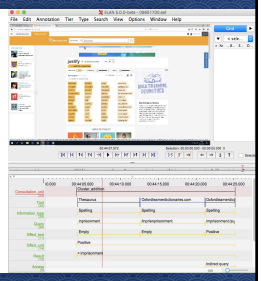
- Comment distinguer les erreurs de "maladresses?"
  - Louvain Error Tagging Manual (Dagneaux et al. 2008) : 'Errors' vs 'Infelicities' (ex. couldn't $could not$; him him/her) – sauf connecteurs et ponctuation ; quid des collocations (ex. high $heavy$ responsibilities) ?
- Comment déterminer la meilleure correction ?
  - Sometimes we cannot blame the masses for being (LS) narrow $small-minded?$...
  - If we were to assume that technological progress and human imagination are two irreconcilable things, then we would (LS) imminently $?$ be taking on a Romantic or even Luddite attitude.

27

27

L'annotation (5): autres

- Toute couche d'annotation augmente la valeur du corpus et permet de répondre à davantage de questions
- Impossible d'inclure tous les types d'annotation potentiellement utiles => à ajouter selon les besoins
- Process Corpus of English in Education (PROCEED ; Gilquin 2022a) : annotation de l'utilisation d'outils d'aide à la rédaction (Gilquin & Laporte 2021)



28

28

"In practice, most annotation schemes … aim at the maximum potential utility, tempered by the practicalities of annotating the text" (McEnery & Wilson 2001, p.34)

29

29

3 ANALYSE

30

30

L'objet de recherche (1): *corpus-driven*

- Privilégier une approche centrée sur le corpus / la langue d'apprenants, ex.
  - Recherche du mot *library* : 40 orthographes différentes dans un corpus d'anglais d'apprenants japonais (Mitton & Okada 2007) : *libelary, liberary, liburally, liburary, liveraly, liverary, liverely*, etc. (=>VariAnt, Anthony 2017)
  - Recherche de connecteurs sur base d'une liste pré-établie (Granger & Tyson 1996) => quid des connecteurs mal orthographiés (ex. *altough, inspite, therefor*) ou non-standard (ex. *on the other side, according to me*) ?
- "Recall problem" / "You-Don't-Know-What-You're-Missing problem" (Ball 1994, p.295)

31

31

L'objet de recherche (2): annotation

- Commencer par une étude pilote pour évaluer la précision de la recherche automatique (Granger 1997)
  - Recherche du passif (Granger 1997), ex.
    - *The reader is lead<noun> to believe that…*
    - *The laws aren't uphold<Vinf> that efficiently*
  - Recherche des constructions causatives avec *cause, get, have* et *make* (Gilquin 2016)
    - Inclusion de toutes les étiquettes correspondant à une forme verbale non-personnelle (infinitif, participe passé et présent), ex. *cause to decrease the crime rate* ; *makes people trying to find one*

32

32

Les facteurs d'influence (1)

- Les nombreux éléments du profil socio-linguistique de l'apprenant (et leur interaction !)
  - Souvent impossible d'effectuer un contrôle strict dans la sélection des données à cause de la taille du corpus (Callies 2015, p.52)
  - Interactions parfois difficiles à interpréter, cf. Deshors (2021) à propos des verbes à particule en anglais dans VOICE : type d'objet direct X rôle de l'orateur

33

33

Les facteurs d'influence (2)

- La tâche, y compris l'intitulé / le sujet

Paquot (2013, p.399) : 14 des 34 blocs lexicaux distinctifs de ICLE-FR (parmi ICLE) sont liés au sujet de la création de l'Europe, ex.

- *Europe will be united against USA and Japan.*
- *Each country will keep its own identity, currency, institutions and constitution.*

Bell et al. (2021, p.223-4) : Image de deux policiers, d'un enfant et de sa mère

- Moins d'opportunités d'utiliser des mots féminins que masculins
- Incertitude quant au caractère correct ou incorrect des pronoms décrivant l'enfant

34

34

Les facteurs d'influence (3)

- La longueur des textes
  - La longueur des phrases tend à augmenter proportionnellement à la longueur des textes (Buttery & Caines 2012)
  - Les adverbes ont proportionnellement plus de chances d'être utilisés dans des textes plus longs (ibid.)
  - La nature et la fréquence des blocs lexicaux varie selon la longueur des textes (et le nombre de textes par corpus) (Pan et al. 2020)

35

35

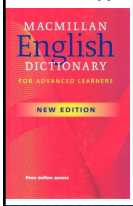
Le choix de la norme (1) (Gilquin 2022b)

- Étudiants natifs ?
  - "Native-speaking students do not necessarily provide models that everyone would want to imitate" (Leech 1998, p.xix)
- Experts (natifs ou non-natifs) ?
  - "both unfair and descriptively inadequate" (Lorenz 1999, p.14)
  - "unrealistic standard of 'expert writer' models" (Hyland & Milton 1997, p.184)
- Étudiants natifs si évaluation ; experts si but formatif (Ädel 2006, p.206-207)

36

36

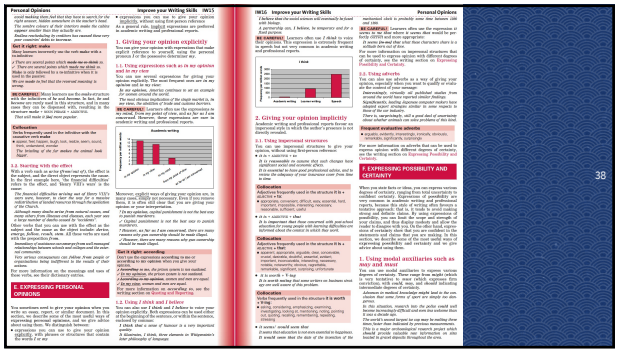
Le choix de la norme (2)



- Ressource pédagogique pour aider les apprenants à améliorer leurs compétences rédactionnelles en anglais académique
  - 100 notes 'Get it right' sous les entrées du dictionnaire correspondantes
  - *Improve your writing skills*: 12 sections axées sur des fonctions rhétoriques importantes en anglais académique (Gilquin et al. 2007a, 2007b)
- Basée sur l'analyse d'ICLE et sa comparaison avec LOCNESS, mais surtout avec le BNC

37

37



38

38

Le circuit long

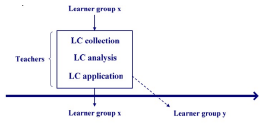
- Corpus pour utilisation pédagogique différée (Granger 2004)
- Risque que le circuit soit interrompu car acteurs différents

39

39

Le circuit court

- Corpus pour utilisation pédagogique immédiate (Granger 2004)
- 'Corpus d'apprenants local' (Seidlhofer 2002)
- Ampleur limitée, mais sur mesure



40

40

L'évolution diachronique

- Date de péremption des corpus d'apprenants ?
  - Contextes d'acquisition
  - Mode d'écriture
  - Format des corpus
- ICLE+30 (cf. Gilquin 2021, à paraître)
  - Recueil de données similaires à ICLE 30 ans plus tard
  - Données plus actuelles
  - Analyse diachronique de la langue d'apprenants en comparant ICLE et ICLE+30

41

41

> "Every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way" (Fillmore 1992, p.35)

42

42

4 CONCLUSION

43

Recueillir et traiter son corpus pour mieux l'analyser

- Les défis et les difficultés liés au recueil, au traitement et à l'analyse des corpus d'apprenants sont nombreux, mais…
- La constitution de corpus d'apprenants est une expérience enrichissante
- Constituer et traiter un corpus permet de mieux connaître ses données, ce qui peut faciliter leur exploitation
- Service à la communauté, mutualisation des ressources

44

44

Quelques lectures utiles

- Bell, Philippa & Caroline Payant. 2021. Designing learner corpora: Collection, transcription, and annotation. In Nicole Tracy-Ventura & Magali Paquot (eds) *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 53–67). Abingdon: Routledge.
- Gilquin, Gaëtanelle. 2015. From design to collection of learner corpora. In Sylviane Granger, Gaëtanelle Gilquin & Fanny Meunier (eds) *The Cambridge Handbook of Learner Corpus Research* (pp. 9–34). Cambridge: Cambridge University Press.
- Tono, Yukio. 2016. What is missing in learner corpus design? In Margarita Alonso-Ramos (ed.) *Spanish Learner Corpus Research: Current Trends and Future Perspectives* (pp. 33–52). Amsterdam: John Benjamins.

45

45

Références (1)

•Ahel, A. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.
•Anthony, L. 2017. VariAnt (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software
•Ball, C. N. 1994. Automated text analysis: Cautionary tales. *Literary and Linguistic Computing* 9(4): 295-302.
•Bell, P., L. Collins & E. Marsden. 2021. Building an oral and written learner corpus of a school programme. Methodological issues. In B. Le Bruyn & M. Paquot (eds) *Learner Corpus Research Meets Second Language Acquisition* (pp. 214-242). Cambridge: Cambridge University Press.
•Buttery, P. & A. Caines. 2012. Normalising frequency counts to account for 'opportunity of use' in learner corpora. In Y. Tono, Y. Kawaguchi & M. Minegishi (eds) *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* (pp. 187-204). Amsterdam: John Benjamins.
•Callies, M. 2015. Learner corpus methodology. In S. Granger, G. Gilquin & F. Meunier (eds) The Cambridge Handbook of Learner Corpus Research (pp. 35-56). Cambridge: Cambridge University Press.
•Dagneaux, E., S. Denness, S. Granger, F. Meunier, J. Neff & J. Thewissen. 2008. *The Louvain error tagging manual. Version 1.3*. Centre for English Corpus Linguistics, Université catholique de Louvain.
•De Felice, R. & S. Pulman. 2009. Automatic detection of preposition errors in learner writing. *CALICO Journal* 26(3): 512-528.
•Deshors, S. 2021. Phrasal verbs in English as a Lingua Franca: a case for corpus-based multifactorial analysis. Presentation at *Workshop on Multiword Units in Multilingual Speakers*, Eberhard Karls Universität Tübingen, Germany, 7 June 2021.
•Ellis, R. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
•Geertzen, J., T. Alexopoulou & A. Korhonen. 2014. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. I. Miller, K. Martin, C. M. Eddington, A. Henery, N. Marcos Miguel, A. M. Tseng, A. Tuninetti, & D. Walter (eds) *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines* (pp. 240-254). Somerville: Cascadilla Proceedings Project.
•Fillmore, C. J. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics". In J. Svartvik (ed.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Stockholm, 4-8 August 1991* (pp. 35-60). Berlin: Mouton de Gruyter.
•Gilquin, G. 2016. Input-dependent L2 acquisition: Causative constructions in English as a foreign and second language. In S. De Knop & G. Gilquin (eds) *Applied Construction Grammar* (pp. 113-148). Berlin: de Gruyter.
•Gilquin, G. 2021. Hic sunt dracones: Exploring some terra incognita in learner corpus research. In A. Čermáková & M. Malá (eds) *Variation in Time and Space: Observing the World through Corpora* (pp. 65-86). Berlin: De Gruyter.

46

46

Références (2)

•Gilquin, G. 2022a. The Process Corpus of English in Education: Going beyond the written text. *Research in Corpus Linguistics* 10(1): 31-44. Available at http://rcl.ucll.nco.es/first-view/174-Article%20Text-1066-1-10-20210407.pdf.
•Gilquin, G. 2022b. One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching* 55(1): 87-99.
•Gilquin, G. A paraître. Diachronic learner corpus research: Examining learner language through the lens of time. In S. Flach & M. Hilpert (eds) *Broadening the Spectrum of Corpus Linguistics: New Approaches to Variability and Change*. Amsterdam: John Benjamins.
•Gilquin, G. & S. De Cock. 2011. Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics* 16(2): 141-172.
•Gilquin, G., S. Granger & M. Paquot. 2007a. Improve your writing skills (Writing sections). In M. Rundell (Editor in chief) *Macmillan English Dictionary for Advanced Learners (Second Edition)* (pp. IW1-IW29). Oxford: Macmillan Education.
•Gilquin, G., S. Granger & M. Paquot. 2007b. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6(4): 319-335.
•Gilquin, G. & S. Laporte. 2021. The use of online writing tools by learners of English: Evidence from a process corpus. *International Journal of Lexicography* 34(4): 472-492.
•Granger S. 1997. Automated retrieval of passives from native and learner corpora: precision and recall. *Journal of English Linguistics* 25(4): 365-374.
•Granger, S. 1998. The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (ed.) *Learner English on Computer* (pp. 3-18). London: Longman.
•Granger, S. 2004. Computer learner corpus research: Current status and future prospects. In U. Connor & Th. A. Upton (eds) *Applied Corpus Linguistics: A Multidimensional Perspective* (pp. 123-145). Amsterdam: Rodopi.
•Granger, S., E. Dagneaux, F. Meunier & M. Paquot. 2009. *International Corpus of Learner English, Version 2: CD-ROM and Handbook*. Louvain-la-Neuve: Presses universitaires de Louvain.
•Granger, S. & S. Tyson. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15(1): 17-27.
•Huang, Y., A. Murakami, T. Alexopoulou & A. Korhonen. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics* 23(1): 28-54.
•Hyland, K. & J. Milton. 1997. Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6(2): 183-205.
•Kreyer, R. 2015. *The Marburg Corpus of Intermediate Learner English (MILE)*. In M. Callies & S. Götz (eds) *Learner Corpora in Language Testing and Assessment* (pp. 13-34). Amsterdam: John Benjamins.
•Leech, G. 1998. Preface. In S. Granger (ed.) *Learner English on Computer* (pp. xiv-xx). London: Longman.
•Lorenz, G. 1999. *Adjective Intensification - Learners versus Native Speakers. A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.

47

47

Références (3)

•Lozano, C. 2021. CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*. OnlineFirst. https://doi.org/10.1177/0267658321105025
•McEnery, T. & A. Wilson. 2001. *Corpus Linguistics: An Introduction. Second Edition*. Edinburgh: Edinburgh University Press.
•McEnery, T., R. Xiao & Y. Tono. 2006. *Corpus-Based Language Studies: An Advanced Resources Book*. London: Routledge.
•Meunier, F. 1998. Computer tools for interlanguage analysis: A critical approach. In S. Granger (ed.) *Learner English on Computer* (pp. 19-37). London: Longman.
•Mitton, R. & T. Okada. 2007. The adaptation of an English spellchecker for Japanese writers. Paper presented at the *Symposium on Second Language Writing, 15-17 September 2007*, Nagoya, Japan. Available at http://eprints.bbk.ac.uk/592-5/592.pdf.
•Myles, F. 2015. Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin & F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research* (pp. 309-332). Cambridge: Cambridge University Press.
•Pan, F. R. Reppen & D. Biber. 2020. Methodological issues in contrastive lexical bundle research: The influence of corpus design on bundle identification. *International Journal of Corpus Linguistics* 25(2): 215-229.
•Paquot, M. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics* 18(3): 391-417.
•Paquot, M., H. Naets & S. Gries. 2021. Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb object structures in LONGDALE. In B. Le Bruyn & M. Paquot (eds) *Learner Corpus Research Meets Second Language Acquisition* (pp. 122-147). Cambridge: Cambridge University Press.
•Rayson, P. & A. Baron. 2011. Automatic error tagging of spelling mistakes in learner corpora. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (eds) *A Taste for Corpora. In Honour of Sylviane Granger* (pp. 109-126). Amsterdam: John Benjamins.
•Rocher, L. , J. M. Hendrickx & Y.-A. de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10: 3069.
•Rozovskaya, A. & Roth, D. 2010. Training paradigms for correcting errors in grammar and usage. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 154-162). Los Angeles: Association for Computational Linguistics.
•Seidlhofer, B. 2002. Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (eds) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 213-234). Amsterdam: John Benjamins.
•Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
•Sinclair, J. 1996. Preliminary recommendations on corpus typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards). Available at www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

48

48