# L'utilisation de corpus pédagogiques pour l'enseignement et la recherche: la question de l'acquisition lexicale

Heather E. Hilton, Ronald Peereman, Michael Gauthier

10èmes Journées Internationales de la Linguistique de corpus
26-28 novembre 2019, Grenoble, France

# L'utilisation de corpus pédagogiques pour l'enseignement et la recherche : la question de l'acquisition lexicale

Heather Hilton [1], Ronald Peereman [2] et Michael Gauthier [1]
CRTT, Université Lyon 2
LPNC, Université Grenoble Alpes
heather.hilton@univ-lyon2.fr, ronald.peereman@univ-grenoble-alpes.fr, michael.gauthier.uni@gmail.com

## 1 Introduction

Dans les premières décennies du XX$^{e}$ siècle, le psychologue américain Edward Thorndike a rassemblé le premier grand corpus constitué pour répondre à des besoins éducatifs : des textes de toutes sortes (littéraires, commerciaux, journalistiques, scolaires), contenant 4,5 millions de mots (Thorndike 1921). Il est difficile pour nous, en 2019, d'imaginer la minutie nécessaire pour élaborer manuellement à partir de ce grand corpus les premières listes de fréquence en anglais, rassemblées dans la célèbre série des *Teacher's Word Books* (Thorndike 1921 ; Thorndike 1931 ; Thorndike & Lorge 1944). Ces listes ont permis l'élaboration d'un syllabus lexical, pour l'apprentissage structuré et progressif de la lecture dans les écoles élémentaires américaines. Considérées comme « l'une des ressources scientifiques les plus utiles jamais développée » (Goodenough 1950 : 296), les listes de Thorndike ont également servi au développement d'outils psychométriques pour mesurer les connaissances lexicales et des niveaux de compétence en lecture.

En collaboration avec Thorndike (Fawcett, Palmer, Thorndike & West 1936), Michael West a élargi et ajusté le corpus de base (le portant à 5 millions de mots), pour identifier les mots les plus utiles (« *of greatest general service* ») pour l'apprentissage de l'anglais langue étrangère ou seconde par des apprenants adultes (West 1936 ; West 1953). A la suite de ce travail didactique, des initiatives parallèles ont été entreprises dans d'autres pays européens après la Seconde Guerre Mondiale, dans un contexte de mobilité internationale et de promotion soutenue de l'enseignement des langues (par exemple, Gougenheim et al. 1956, pour le français langue étrangère). Ces travaux lexicologiques ont fourni aux auteurs des manuels de langues un syllabus lexical, permettant l'introduction progressive des mots selon la fréquence de leur utilisation, et donc une graduation des supports (textes et enregistrements), pour un apprentissage structuré et donc optimisé.

Lors de la « révolution » communicative des années 1980 (en Europe et aux Etats-Unis), la notion d'un programme lexical structuré fut abandonnée en didactique des langues (voir, par exemple, Auroux 1985, pour une vision résolument anti-lexicale de la compétence communicative). L'attention méthodologique depuis cette époque est focalisée sur les « activités langagières » en classe de langue (compréhension, expression, interaction), et le programme est basé sur des « actes de parole » ou différentes fonctions interactionnelles du langage. Cette centration didactique sur l'utilisation du langage a eu comme effet de déstructurer la programmation des éléments formels de la langue à apprendre (vocabulaire, morphologie, syntaxe, prononciation) :

manuels et enseignants abordent les formes en fonction des besoins communicatifs de telle situation interactionnelle, et non plus selon des critères linguistiques (comme la fréquence d'un mot ou d'une forme grammaticale : Meara 1980 ; Swan 1985 ; Arnaud et al. 1985 ; Nordlund 2016 : 48-50). La dernière décennie en France a vu le retour des préoccupations lexicales dans les classes de français langue maternelle (et notamment l'importance du vocabulaire dans l'apprentissage de la lecture, Dehaene 2011), mais ce renouveau d'intérêt pour le vocabulaire n'est pas encore reflété dans les textes qui régissent l'enseignement des langues vivantes en France[1].

## 2 Corpus et méthodologie

### 2.1 Corpus

Dans ce contexte, un groupe de chercheurs dans quatre universités françaises a élaboré un projet visant à analyser de plus près l'acquisition lexicale dans les classes de langue vivante en France, à l'école élémentaire et au collège. La première tâche de ce projet porte sur l'analyse d'un grand corpus de manuels utilisés pour enseigner l'anglais dans les quatre années du collège (60 manuels, 15 pour chaque année de collège), à l'image de la base lexicale *Manulex*, compilée sur 54 manuels scolaires utilisés dans l'enseignement élémentaire en France (Lété *et al.* 2004). Trente-deux manuels ont été numérisés jusqu'ici, générant un corpus de 725 720 mots, avec en moyenne environ 181 000 mots par année de collège. Un deuxième volet de cette tâche consiste à filmer quatorze leçons d'anglais en collège (trois ou quatre leçons par année de collège), générant un petit corpus oral parallèle, qui complète le corpus écrit et permet quelques comparaisons lexicologiques entre manuels et leçons.

### 2.2 Méthodologie

Dans cette communication, nous présenterons la méthodologie utilisée pour numériser les manuels et établir des listes de mots selon leur fréquence et leur fonction grammaticale : les listes sont annotées et lemmatisées selon CLAWS (Garside & Smith 1997) et *Stanford NLP* (Toutanova *et al.* 2003). Nous résumerons également les méthodes et outils utilisés pour transcrire le corpus des leçons avec le logiciel EXMARaLDA (Schmidt *et al.* 2014) et l'étiqueter avec *WMatrix* (Rayson 2008).

## 3 Résultats

Nos premières analyses du corpus écrit révèlent une grande et surprenante disparité lexicale entre les manuels. Alors que 20 000 types sont dénombrés (incluant 20% de sigles et noms propres –un lexique très étendu, pour les niveaux européens A2 et B1 visés en collège), seulement 3500 types (17,5%) sont partagés entre les quatre niveaux scolaires. Ce ratio n'est que légèrement plus élevé (24,27%) lorsque l'analyse porte sur les lemmes, sans prise en compte des noms propres et sigles. De façon inattendue, les manuels d'une même année (les huit manuels

1. Les programmes du Socle commun insistent sur l'importance du vocabulaire 80 fois dans les sections consacrées à l'enseignement du français, de l'histoire, des arts plastiques... mais une seule fois dans les sections dédiées à l'enseignement des langues vivantes (Ministère de l'éducation nationale 2015).

de 5$^{e}$ analysés jusqu'ici, par exemple) ne partagent que 5,5% de leurs types ; dans un contexte pédagogique structuré, on s'attendrait à un recouvrement nettement plus élevé (Nordlund 2016). Plus surprenant encore, 186 mots (7%) du *New General Service* List (Browne *et al.* 2013) sont absents de ces manuels de collège (des mots comme *ally*, *overall*, *sufficient*, *income*, *strengthen*, *tire*...) ; pourtant, la totalité des 2801 mots de cette liste de base (les mots incontournables pour une utilisation réceptive ou productive de l'anglais) devrait logiquement figurer –avec une fréquence élevée –dans des supports à ces niveaux élémentaires (Nordlund 2016, pour des résultats semblables dans des manuels d'anglais utilisés en Suède). De façon plus rassurante, la grande majorité (85 %) des mots entendus en classe se retrouve dans les manuels du même niveau, mais nous trouvons aussi (à l'image du corpus écrit) que ce taux de recouvrement lexical est quasi identique, quel que soit le niveau des manuels que l'on compare aux leçons.

Les premières analyses de ce corpus de manuels d'anglais donnent donc des résultats inattendus, qui soulèvent des questions didactiques de fond. En conclusion, nous évoquerons quelques retombées d'un programme lexical diffus et sans doute trop étendu : le faible niveau des élèves français en anglais langue étrangère (compréhension de l'oral et de l'écrit, expression écrite), selon l'étude européenne *Surveylang* (European Commission 2012 ; Beadle & Scott 2014) ; leurs connaissances lexicales limitées en anglais L2 à l'arrivée dans l'enseignement supérieur (Hilton 2019 : 25). Nous mentionnerons également les démarches expérimentales basées sur les listes du corpus, qui auront comme objectif de mesurer l'émergence des connaissances lexicales en anglais L2 chez des collégiens français (tâche de décision lexicale). Notre communication illustrera donc deux utilisations possibles des corpus : didactique (la programmation des contenus lexicaux en langue étrangère) et expérimentale (la conception d'outils pouvant mesurer les acquis lexicaux).

## Références bibliographiques


Arnaud, P. J. L., Béjoint, H. & Thoiron, P. (1985). A quoi sert le programme lexical ? *Les Langues Modernes 3/4*, 72-85.

Auroux, S. (1985). Le droit à l'oubli : réponse à Arnaud, Béjoint et Thoiron. *Les Langues Modernes 3/4*, 86-91.

Beadle S, and Scott D (2014). *Languages in Education and Training : Final Country Comparative Analysis*, Report n° J9241. European Commission.

Browne, C., Culligan, B. & Phillips, J. (2013). *New General Service List*. Tokyo.

Dehaene, S. (dir ; 2011). *Apprendre à lire : Des sciences cognitives à la salle de classe*. Paris : Odile Jacob.

European Commission (2012) *First European Survey on Language Competences : Final Report.* Brussels : European Commission Education and Training Division.

Faucett, L. W., Palmer, Thorndike, E. L., & West, M. P. (1936). *Interim Report on Vocabulary Selection for the Teaching of English as a Foreign Language*. London.

Garside, R. & Smith, N. (1997). A hybrid grammatical tagger : CLAWS4. Dans R. Garside, G. Leech & A. McEnery (dir), *Corpus annotation : Linguistic information from computer text corpora.* London : Longman, 102-121.

Goodenough, F. L. (1950). Edward Lee Thorndike : 1874-1949. *The American Journal of Psychology* 63(2), 291-301.

Gougenheim, G., Michea, R., Rivenc, P. & Sauvageot, A. (1956). *L'élaboration du français élémentaire : Étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Paris : Didier.

Hilton, H. E. (2019). *Sciences cognitives et didactique des langues, Rapport d'expertise.* Paris : CNESCO.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX : A grade-level lexical database from French elementary-school readers . *Behavior Research Methods, Instruments, & Computers, 36*, 156-166.

Meara, P. M. (1980). Vocabulary acquisition : A neglected aspect of language learning. *Language Teaching and Linguistics Abstracts, 13*, 221-46.

Ministère de l'éducation nationale (2015). *Programmes pour le cycle 2, 3, 4.* Paris : MEN.

Nordlund, M. (2016). EFL textbooks for young learners : a comparative analysis of vocabulary. Education Inquiry•, 7(1), 47-68.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics.* 13(4), 519-549.

Schmidt, T., & Wörner, K. (2014). EXMARaLDA. In U. Gut, J. Durand & G. Kristofferse (dir), *Handbook on Corpus Phonology.* Oxford : Oxford University Press, 402-419.

Swan, M. (1985). A critical look at the Communicative Approach. *ELT Journal, 39*(1-2), 2-12 ; 76-87.

Thorndike, E. L. (1921, 1931). *A Teacher's Word Book : the Twenty Thousand Words Found Most Frequently in General Reading for Children and Young People.* New York : Teachers College.

Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York : Teachers College.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, 252-259.

West, M. P. (1936, 1953). *A General Service List of English Words.* London : Longman.