Les corpus et les TIC comme aide à la découverte des langues.

Alex Boulton

Les corpus et les TIC comme aide à la découverte des langues

PAR **ALEX BOULTON,** ATILF - CNRS & UNIVERSITÉ DE LORRAINE

Introduction

Traditionnellement, l'activité d'enseignement consiste à transmettre des savoirs en s'appuyant sur des exercices (plus ou moins contrôlés) et des activités (plus ou moins ouvertes) afin de tester l'acquisition de ces savoirs et ensuite de développer leur mise en pratique. Cette caricature volontairement réductrice représente un aspect important de nos activités dans la plupart des méthodes et approches explorées à ce jour. Toutefois, on peut retourner la situation en demandant aux apprenants de créer leurs propres savoirs. Ceci est à la base de toute approche s'appuyant sur le constructivisme, la découverte, l'apprentissage par la tâche.

Une approche déductive traditionnelle – centrée sur l'enseignant, passant des généralités (« règles ») à leur application spécifique – contraste avec une approche inductive où l'on commence par des occurrences spécifiques pour arriver à des constats généralisables. Cette approche inductive nécessite en première instance une exposition régulière et massive à la langue, l'un des avantages notables d'une acquisition informelle avec les pratiques courantes d'activités langagières en ligne (ex. Kusyk & Sockett, 2012). L'inconvénient est que le processus peut être long, et ne se prête pas facilement au repérage (« *noticing* ») des éléments de langue en tant que tels puisque l'accent est mis en priorité sur le contenu et la communication. L'élément manquant serait un moyen d'aider l'apprenant à mieux en profiter en attirant son attention sur des points langagiers utiles et récurrents en contexte. Ainsi, le recours aux outils et techniques de la linguistique de corpus peut s'avérer bénéfique dans une approche que nous appellerons « apprentissage sur corpus » (ASC ; Boulton & Tyne, 2014 ; cf. le « *data-driven learning* » de Johns, 1990).

Le concept central semble alléchant : l'apprenant adopte le rôle d'un « chercheur » qui teste et affine ses hypothèses en explorant la langue pour devenir un véritable « Sherlock Holmes » (Johns, 1997, p. 101). L'enseignant devient guide ou directeur de recherches, et la langue regagne une place centrale. Boulton et Cobb (2017) déclinent les arguments en cinq catégories qui reflètent les théories linguistiques et psycholinguistiques, de l'apprentissage et de l'acquisition, ainsi que l'évolution des TIC et l'émergence des natifs du numérique. Les quelques lignes qui suivent esquissent les principaux socles théoriques qui étayent l'ASC.

Les « règles » traditionnelles représentant un concept abstrait, intellectuellement difficile à assimiler et appliquer ; l'ASC s'aligne sur le constructivisme et fait appel à nos capacités naturelles à détecter les régularités dans des données complexes – l'objet langue qui est dynamique, probabiliste et interactif. Ainsi, l'approche réduit la charge cognitive, tout en faisant appel au traitement cognitif profond essentiel à un apprentissage évolutif. L'ASC facilite le développement des compétences de repérage, en privilégiant l'organisation en « *chunks* » (séquences de mots) qui réunissent lexique et grammaire. Le travail rééquilibre des approches « *top-down* » (sens) et « *bottom-up* » (forme), en redonnant de l'importance à la langue elle-même, les formes et sens en contexte. Enfin, l'ASC s'appuie sur des pratiques courantes chez les apprenants qui se servent déjà des TIC (technologies de l'information et de la communication) pour chercher des réponses à leurs questions réelles, notamment grâce à Google (ou un autre

moteur de recherche) comme « concordancier » pour interroger le web comme « corpus » d'usages attestés. L'enseignant peut aider l'apprenant à mieux l'utiliser ; ce travail peut servir de tremplin vers une approche sur corpus plus approfondie.

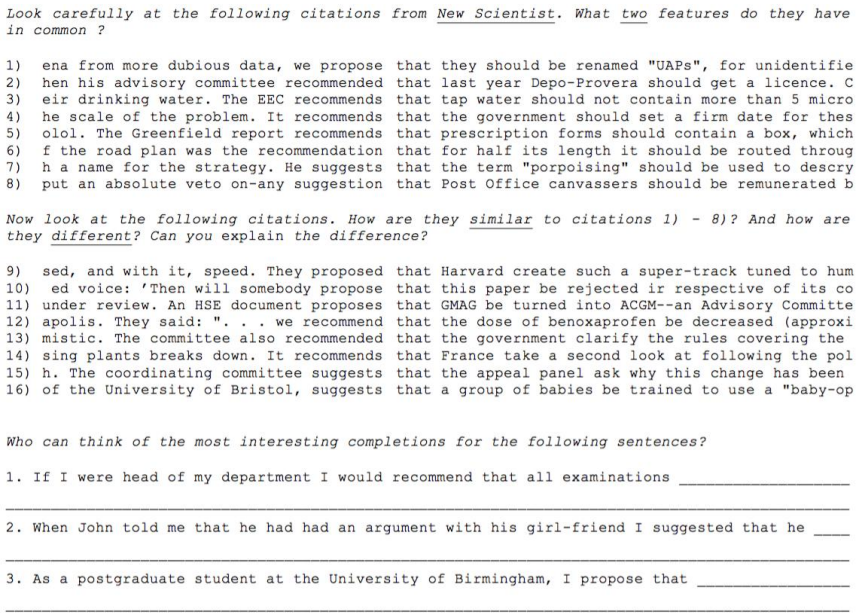
Nonobstant les avantages rapidement esquissés ci-dessus, il reste des contre-arguments de poids. On peut citer le fait que certains apprenants et enseignants se sentent mal à l'aise avec les TIC en cours, même s'ils se servent quotidiennement d'applications complexes. Le langage présent dans un corpus, le plus souvent authentique et écologique, peut se révéler difficile pour de nombreux apprenants à des niveaux peu élevés. La présentation des données sous forme de lignes tronquées (ou KWIC – *key words in context* ; voir Figure 1) est utile pour le repérage des « *patterns* » (usages récurrents) mais ne se prête pas à une lecture linéaire, et va à l'encontre d'un travail sur le sens à des fins communicatives (surtout pour l'oral). Ainsi, l'ASC nécessiterait un entraînement long et laborieux, et les processus seraient chronophages par rapport à une approche déductive où les réponses sont fournies directement.

Pour ces raisons, certains collègues qui pratiquent l'ASC préfèrent définir des limites : uniquement auprès d'apprenants de niveaux avancés avec des besoins précis et spécifiques ; avec des supports imprimés ou des corpus de textes choisis pour leur pertinence et à un niveau de langue approprié ; avec des présentations autres que les KWIC ; à des fins de référence plutôt que pour un apprentissage proprement dit. Quant à la question du temps, il s'agit surtout d'un investissement pour devenir « meilleur apprenant » – plus sensible au fonctionnement de la langue, capable de repérer des éléments récurrents et de se poser des questions utiles. Ces objections méritent d'être prises au sérieux dans la proposition d'activités et dans les recherches effectuées.

L'ASC classique

À quoi ressemble donc l'ASC ? Sous sa forme classique, on recherche un mot ou groupe de mots (ou plus précisément une chaîne de caractères) qui apparaît au milieu de la présentation avec un contexte minimal à gauche et à droite pour réunir lexique et grammaire. L'exemple en Figure 1 est tiré de Johns (1990 : 29) suite à une recherche pour *that* précédé par un verbe dans un corpus d'anglais scientifique. La requête est formulée par l'enseignant qui choisit et organise les lignes pertinentes et en prépare un document de travail incorporant des consignes afin de focaliser l'attention sur un phénomène donné. Dans cet exemple, la tâche est de trouver les points communs entre les deux séries de lignes, de remarquer les *patterns* qui les distinguent – l'utilisation de *should* dans le premier (*should be renamed, should get a licence*…) par opposition à la forme de base du verbe dans le deuxième, quel que soit le sujet ou le temps (*be rejected, take*…).

Figure 1. Exemple traditionnel de l'ASC



Look carefully at the following citations from New Scientist. What two features do they have in common ?

1) ena from more dubious data, we propose that they should be renamed "UAPs", for unidentifie
2) hen his advisory committee recommended that last year Depo-Provera should get a licence. C
3) eir drinking water. The EEC recommends that tap water should not contain more than 5 micro
4) he scale of the problem. It recommends that the government should set a firm date for thes
5) olol. The Greenfield report recommends that prescription forms should contain a box, which
6) f the road plan was the recommendation that for half its length it should be routed throug
7) h a name for the strategy. He suggests that the term "porpoising" should be used to descry
8) put an absolute veto on-any suggestion that Post Office canvassers should be remunerated b

Now look at the following citations. How are they similar to citations 1) - 8)? And how are they different? Can you explain the difference?

9) sed, and with it, speed. They proposed that Harvard create such a super-track tuned to hum
10) ed voice: 'Then will somebody propose that this paper be rejected ir respective of its co
11) under review. An HSE document proposes that GMAG be turned into ACGM--an Advisory Committe
12) apolis. They said: ". . . we recommend that the dose of benoxaprofen be decreased (approxi
13) mistic. The committee also recommended that the government clarify the rules covering the
14) sing plants breaks down. It recommends that France take a second look at following the pol
15) h. The coordinating committee suggests that the appeal panel ask why this change has been
16) of the University of Bristol, suggests that a group of babies be trained to use a "baby-op

Who can think of the most interesting completions for the following sentences?

1. If I were head of my department I would recommend that all examinations _____
_____
2. When John told me that he had had an argument with his girl-friend I suggested that he ____
_____
3. As a postgraduate student at the University of Birmingham, I propose that _____
_____

Cette approche se distingue d'une démarche déductive où l'enseignant présente une règle : l'entrée dans un manuel d'usage très répandu (Swan, 2016 : 232-234) compte une cinquantaine de lignes – utile pour certains, mais abstraite et potentiellement difficile à assimiler. Avec l'activité proposée par Johns, c'est l'apprenant qui crée des règles qui ont un sens pour lui : le type de verbe impliqué, les constructions, le sens, le registre, etc. Ces éléments essentiels de l'ASC ont peu changé en 30 ans (voir Boulton, 2017, pour un recensement historique). Les différences se trouvent principalement dans l'esthétique et l'ergonomie mais surtout dans la rapidité des réponses : Aston (1996 : 179) pouvait lancer une requête dans les 100 millions de mots du British National Corpus et avoir le temps de déjeuner avant de récupérer le résultat. Aujourd'hui, le même corpus peut être consulté en une fraction de seconde. Autre différence majeure, l'existence aujourd'hui de nombreux corpus de toutes sortes, ainsi que la possibilité de créer facilement son propre corpus grâce aux immenses quantités de texte sur le web.

Différents gratuiciels facilitent la navigation au sein d'un corpus, parfois consultables en ligne où l'on dépose son corpus à chaque fois (ex. le Compleat Lexical Tutor de Tom Cobb[1]), parfois téléchargeables pour un travail hors ligne (ex. AntConc de Laurence Anthony[2]) ; d'autres corpus en ligne sont interrogeables grâce à une interface intégrée (ex. celle de Mark Davies à Brigham Young University[3]). Les exemples les plus connus sont accompagnés par des tutoriels vidéo, officiels ou pas, qui couvrent les fonctionnalités de base.

Si le format KWIC représente certainement l'image de marque de la linguistique de corpus, les logiciels permettent bien d'autres fonctionnalités. Ainsi, avec différents outils on peut aussi visualiser des phrases complètes, des extraits plus longs voire la position du mot recherché dans un texte complet. On peut également voir sa répartition au sein d'un corpus, ou sa distribution dans différents sous-corpus. On peut créer des listes de mots par fréquence, ou encore des listes de « *chunks* ». Ainsi, on voit que les mots les

1 https://www.lextutor.ca/conc
2 http://www.laurenceanthony.net/software/antconc
3 https://corpus.byu.edu

plus fréquents sont extrêmement fréquents et méritent une attention particulière, tandis que d'autres, selon la loi de Zipf, sont rencontrés bien moins souvent et se prêtent à des stratégies indirectes comme l'inférence. D'autres mots encore, même rares dans la langue en général, peuvent se révéler importants pour des besoins précis (pour un travail disciplinaire, par exemple), ou bien dans un seul texte. On peut chercher les occurrences d'un mot donné dans des séquences figées, ou encore ses collocations plus souples. Il est aussi possible de comparer deux corpus pour identifier leurs spécificités. Les « *taggers* » et « *lemmatisers* » permettent respectivement d'étiqueter son corpus selon la partie du discours (ex. *walk* comme nom ou comme verbe) et de regrouper les mots d'une même racine (ex. *walk, walks, walking, walked*).

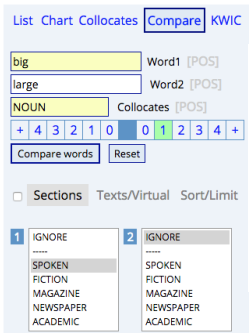
La lexico-grammaire est certainement le domaine le plus souvent abordé dans l'ASC alors qu'elle avait longtemps été mise de côté pour un enseignement plus centré sur la communication. L'ASC la fait revivre précisément car il la replace en contexte de communication. Par ailleurs, l'ASC est de plus en plus souvent entrepris pour un travail sur le discours et les fonctions pragmatiques. Mais puisque les requêtes se font nécessairement sur du texte écrit (y compris des transcriptions), l'ASC est moins souvent utilisé pour l'oral malgré quelques tentatives dans ce sens.

Les outils ASC

Il est impossible de montrer l'ensemble des possibilités de requêtes avec un corpus, mais quelques exemples donneront une idée du potentiel de l'approche. Ce qui suit porte exclusivement sur l'anglais, la langue enseignée par le présent auteur, et pour laquelle il existe beaucoup d'outils, d'études et d'aides (Bennett, 2010 ; Reppen, 2010). Quelques exemples pour le français, l'allemand et l'espagnol, sont fournis dans Boulton et Tyne (2014).

Une question fréquente concerne la distinction entre synonymes proches. Le *Oxford advanced learner's dictionary*,[4] par exemple, définit *big* en anglais comme « *large in size, degree, amount, etc.* », et *large* comme « *big in size or quantity* ». Il est clair que les deux sont très proches, mais les entrées ne répondent pas à la question des différences. L'interface au Corpus of Contemporary American English (COCA)[5] permet de comparer les noms qui suivent directement ces deux adjectifs pour repérer des différences. Ainsi, dans la Figure 2, on voit clairement que dans le langage parlé, *large* est souvent suivi de noms représentant une quantité mesurable, tandis que *big* signale une valeur subjective.

Figure 2. Comparaison de synonymes proches dans le COCA



WORD 1 (W1): BIG (5.11)

	WORD	W1	W2	W1/W2	SCORE
1	DEAL	2605	2	1,302.5	255.1
2	FAN	531	0	1,062.0	208.0
3	DAY	465	0	930.0	182.1
4	BROTHER	408	0	816.0	159.8
5	TROUBLE	374	0	748.0	146.5
6	SURPRISE	340	0	680.0	133.2
7	NIGHT	255	0	510.0	99.9
8	BUCKS	204	0	408.0	79.9
9	CHALLENGE	200	0	400.0	78.3
10	WIN	192	0	384.0	75.2

WORD 2 (W2): LARGE (0.20)

	WORD	W2	W1	W2/W1	SCORE
1	MEASURE	135	2	67.5	344.7
2	EXTENT	175	3	58.3	297.9
3	QUANTITIES	56	1	56.0	285.9
4	DEGREE	99	2	49.5	252.8
5	QUANTITY	24	0	48.0	245.1
6	AMOUNTS	201	6	33.5	171.1
7	SUM	26	1	26.0	132.8
8	PROPORTION	25	1	25.0	127.7
9	AMOUNT	194	8	24.3	123.8
10	INTESTINE	12	0	24.0	122.5

4 https://www.oxfordlearnersdictionaries.com
5 https://corpus.byu.edu/coca ; créé par Mark Davies à Brigham Young University (BYU).

Cette fonctionnalité ne présente que les différences statistiques, et très souvent les mots sont effectivement interchangeables avec peu de distinctions au niveau du sens. Toutefois, une recherche par registre (Figure 3) met en évidence que *big* est cinq fois plus fréquent que *large* dans le sous-corpus de l'anglais parlé (644 vs 126 occurrences par million de mots), alors que *large* est quatre fois plus fréquent que *big* dans la partie académique du COCA (368 vs 88). (Les barres sont de longueur relative et non absolue.)

Figure 3. Distribution dans les sous-registres de COCA

« big »

SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
SPOKEN	75,266	116.7	644.68	
FICTION	55,924	111.8	500.01	
MAGAZINE	57,496	117.4	489.94	
NEWSPAPER	60,383	113.0	534.38	
ACADEMIC	9,842	111.4	88.34	

« large »

SPOKEN	14,740	116.7	126.25	
FICTION	24,615	111.8	220.08	
MAGAZINE	46,740	117.4	398.28	
NEWSPAPER	29,505	113.0	261.12	
ACADEMIC	41,012	111.4	368.12	

Comme nous l'avons dit, un atout des corpus est d'observer les mots en contexte. Par exemple, quels adverbes en anglais servent ordinairement à intensifier le sens du verbe *advise* (conseiller) ? Une simple requête ADVERBE+*advise* dans le COCA montre que la seule option fréquente est *strongly advise*. En cliquant dessus, on trouve les occurrences en contexte, ce qui permet de vérifier le sens et d'analyser l'usage. Dans l'échantillon en Figure 4, on remarque la structure *advise someone to do something* dans la moitié des lignes, le plus souvent avec *not*. Ainsi, consulter un corpus permet de constater non pas ce qui est possible mais ce qui est fréquent et usuel, contribuant au développement d'une production plus naturelle. Ce type d'activité n'est pas sans rappeler l'exemple de Johns en Figure 1 ; l'enseignant peut toujours créer un exercice sur papier en sélectionnant les lignes, ou bien transmettre le lien de la requête[6] aux apprenants pour qu'ils voient la requête et les résultats. Étant donné la disponibilité des corpus BYU et la rapidité et facilité de l'interface, les apprenants peuvent y accéder directement pour formuler leurs propres requêtes et explorer leurs propres questions.

Figure 4. Échantillon de concordance dans le COCA

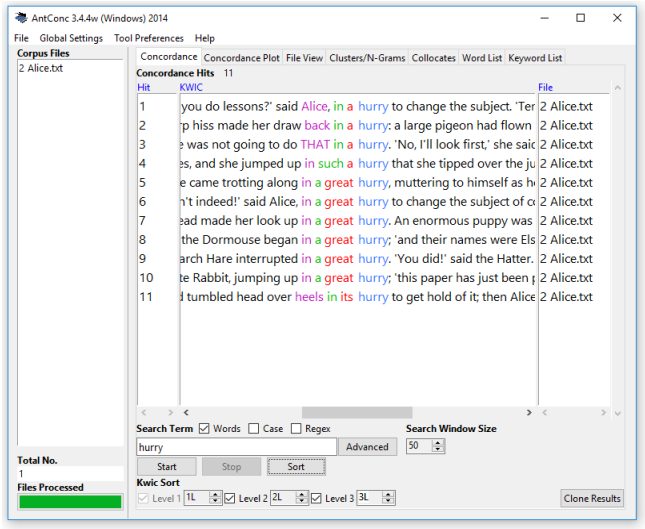
then looked at Mattie, her expression hardening. " I strongly advise you not to refuse medical care for your child. " # * * *
the bar isn't a safe place. The fliers " strongly advise " people, especially women, to avoid the bar. # Kennedy declined to
must make an exception. " # " I would strongly advise against special exceptions, " Dr. Plecker said. # " Good God, man
ment, Mars One says on its website that it will " strongly advise the settlement habitants not to attempt to have children, " given the ur
ent with your thinking it creates anxiety, stress and fear. We strongly advise people who are feeling anxious about the future that the b
, I'd like to meet my granddad? MELISSA-MOORE-1KE# I would strongly advise that she doesn't. JUJU-CHANG-1-ABC-# (Off-camera) Why
you can hear me, be advised that we strongly advise you not to attempt a rescue on your own. However, in case you
down, " Singh repeated. " Sir, I would strongly advise you not to question the courage of my soldiers. We left most of our
children) forbade us to go. " I strongly, strongly advise against any type of travel, Ella, " said Dr. Tomaszewski, one of
rial. Almost any structure can serve satisfactorily for housing poultry. I strongly advise leaving an earth floor in the coop and covering it

Travailler avec un grand corpus générique comme le COCA peut ne pas convenir à tous, mais les outils et l'approche peuvent s'appliquer à des collections de textes simplifiés,[7] à des corpus plus modestes créés spécifiquement pour un public cible – par exemple, l'ensemble des textes dans un manuel scolaire, voire un seul texte à étudier. L'exemple

6 Pour cette requête : https://corpus.byu.edu/coca/?c=coca&q=63282623.
7 Le Compleat Lexical Tutor offre 1,25 million de mots de « *graded readers* ».

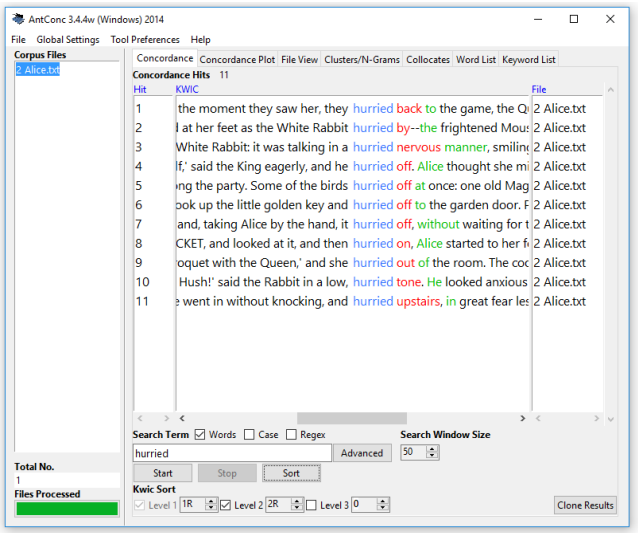
suivant porte sur le livre des *Aventures d'Alice au pays des merveilles* en anglais, disponible gratuitement sur le site du Project Gutenberg,[8] et le logiciel AntConc à télécharger. Pour commencer, une simple liste de fréquence permet de cibler les mots qui seront rencontrés souvent dans le texte ; ceux-ci donnent une idée du contenu et méritent ainsi une attention particulière avant, pendant ou après la lecture. Un exemple : le mot *hurry* (hâte) qui se présente dès la première page n'est pas forcément connu de tous. Plutôt que de recourir à un dictionnaire pour une définition générale, l'apprenant peut chercher chaque instance dans le texte pour voir les sens et usages à travers les multiples exemples et, de cette façon, ne pas s'éloigner du livre. Parmi les 25 occurrences de hurr* (terme à chercher) se trouvent 11 de *hurry*, à chaque fois un nom – et à chaque fois dans la structure *in a hurry*, souvent complétée par un autre mot dont six fois *in a great hurry* (Figure 5). Quant à la forme *hurried* (Figure 6), elle est toujours suivie d'une particule dans ce livre (*back, by, off, on, out, upstairs*), sauf quand elle a une fonction d'adjectif (lignes 3 et 10). Ces constats ne reflètent que le seul livre à l'étude, un avantage non négligeable pour le travail immédiat ; pour généraliser, un travail complémentaire peut s'avérer souhaitable.

Figure 5. Concordance de « hurry » dans Alice avec AntConc



[8] https://www.gutenberg.org. Écrit par Lewis Carroll en 1865 et rendu populaire à travers plusieurs films.

Figure 6. Concordance de « hurried » dans Alice avec AntConc



L'ASC accessible

L'une des critiques souvent formulées envers l'ASC est que ce sont des chercheurs qui sont à l'origine des ressources et des techniques proposées ; ainsi, les apprenants doivent maîtriser les outils et interfaces conçus par et pour des experts. La question se pose alors : est-il possible de rapprocher l'ASC des apprenants plutôt que d'exiger qu'ils s'approprient tous les éléments de la linguistique de corpus ? (Voir les exemples dans Boulton, 2015.)

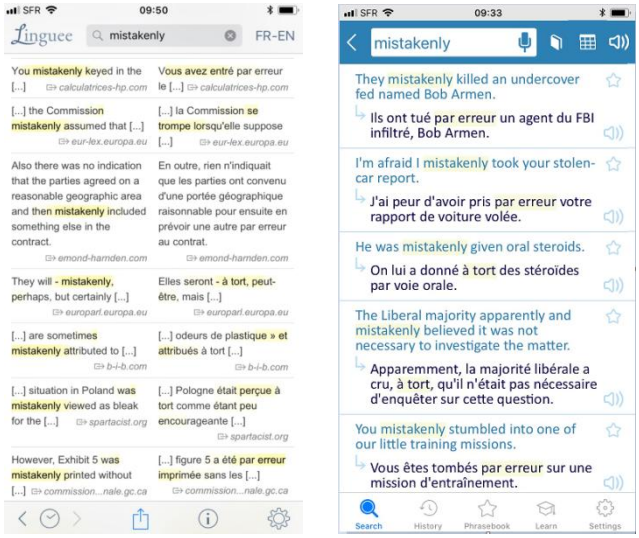
La simple fonction « rechercher » (CTRL+F) dans un seul texte électronique permet de retrouver les occurrences répétées, par exemple de *hurry / hurried* dans *Alice au pays des merveilles*. L'apport du concordancier permet des fonctionnalités supplémentaires mais ne change pas le principe fondamental.

De même, les moteurs de recherche comme Google sont bien connus de tous. Toutefois, la plupart des utilisateurs ignorent que ces outils permettent, entre autres, de filtrer les requêtes de différentes manières. Sensibiliser les étudiants à un meilleur usage de Google peut leur servir tout au long de la vie et bien au-delà de l'apprentissage des langues. Geiller (2014) décrit en détails son usage avec un public en classe prépa. L'une des meilleures astuces est l'utilisation des guillemets pour forcer Google à rechercher exactement la chaîne de mots demandée. Les guillemets peuvent être complétés par l'usage d'un astérisque comme joker pour représenter n'importe quel mot dans la chaîne. Une requête "*play a * role in*", inspiré par le cas d'une étudiante qui sur-utilisait l'adjectif *important*, fournit de nombreuses occurrences d'autres possibilités en contexte (ex. *key*, *critical*). Comme avec tout travail sur corpus, l'outil ne propose pas « la » bonne réponse mais des exemples d'usages attestés. La tâche de l'apprenant est de trier et d'explorer les informations fournies afin de choisir en toute connaissance de cause une réponse à sa question. Les résultats (« *snippets* ») de Google ne sont pas sans rappeler la présentation KWIC, tout comme les moteurs de recherche de nombreux sites web.

Certains éléments linguistiques sont difficiles à appréhender en raison des différences dans la couverture lexicale ou les normes grammaticales L1-L2. L'adverbe anglais *mistakenly*, par exemple, se traduit dans un dictionnaire en ligne principalement par des locu-

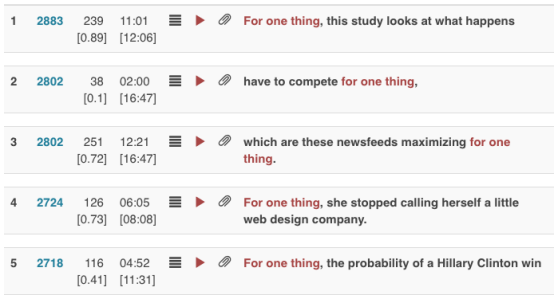
tions adverbiales, à savoir « à tort » ou « par erreur ».[9], Or, de nombreux exemples peuvent s'avérer plus utiles qu'une traduction. La Figure 7 présente les premiers résultats de deux applications, Linguee[10] et Reverso Context,[11] ici sur smartphone. Ces deux outils offrent un accès à des textes dans de nombreuses langues à côté de traductions réelles, par opposition à des traducteurs automatiques comme ceux de Google ou DeepL. Encore une fois, il n'y a pas de garantie quant à la qualité des traductions individuelles, mais celles-ci donnent une indication des traductions fréquentes et peuvent être source d'inspiration, à explorer en cas de doute.

Figure 7. Contextes parallèles dans Linguee et Reverso



Le dernier exemple sert à souligner un travail à l'oral, tiré des conférences TED[12] dont les transcriptions ont été intégrées à un concordancier.[13] Ce site est surtout précieux pour l'alignement texte et vidéo. Par exemple, on peut rechercher une chaîne de mots comme *for one thing* qui dépasse le purement lexical et voir une concordance, ou bien l'écouter en visionnant la locutrice à l'endroit précis où elle s'en sert, appuyée par la transcription interactive (Figure 8).

Figure 8. Concordance et transcription interactive dans TED



9 https://www.wordreference.com est l'un des dictionnaires préférés des étudiants dans notre université.
10 https://www.linguee.fr
11 http://context.reverso.net/traduction
12 https://www.ted.com
13 https://yohasebe.com/tcse, par Yoichiro Hasebe.

237	10:57	Now, as you can imagine,
238	10:58	the answers aren't quite as simple as that.
239	11:01	For one thing, this study looks at what happens
240	11:04	to thousands and thousands of children on average,
241	11:06	but that doesn't necessarily say what will help my child or your child
242	11:10	or any individual child.

Les recherches en ASC

Pour tout innovation pédagogique, il ne suffit pas de se contenter d'un potentiel théorique ou d'une expérience personnelle ; il convient aussi d'évaluer l'apport réel dans des études de recherche. Malgré les plaintes récurrentes concernant le manque de recherches dans le domaine, une collection personnelle compte actuellement plus de 300 études qui visent à évaluer différents aspects de l'ASC sur près de 30 ans. Boulton et Cobb (2017) ont tenté une méta-analyse rigoureuse sur plus de 3000 participants dans 88 échantillons uniques relevés dans 64 études différentes qui correspondaient aux critères retenus, avec des données quantitatives permettant de chiffrer l'apport d'un travail sur corpus comme aide à l'apprentissage ou comme outil de référence auprès d'apprenants L2. Il faut souligner que cette méta-analyse, de par sa nature, ne traite que les résultats quantifiables ; d'autres études s'intéressent aux comportements ou aux représentations des apprenants et nécessitent une analyse plus qualitative.

La méta-analyse se base sur le *d* de Cohen pour l'ensemble des études, une statistique qui mérite une brève discussion pour bien la comprendre. Il s'agit d'une valeur calculée selon une formule simple et transparente : elle compare deux moyennes en tenant compte de la variation dans chacune. Ainsi, une valeur de 1,00 signifie que la moyenne au post-test se situe à un écart-type de celle du pré-test (intra-groupe), ou encore que le groupe expérimental dépasse d'un écart type le groupe témoin (inter-groupe). Il est assez évident qu'une comparaison intra-groupe donne un résultat plus important car on teste un fait de langue avant et après une intervention – il serait difficile d'imaginer que la moyenne n'augmente pas ou de peu dans une étude de ce type. Par contre, avec une comparaison entre deux groupes, le groupe témoin n'est pas généralement un véritable groupe de contrôle car il reçoit un enseignement et non un traitement-zéro. Par conséquent, il est tout à fait possible qu'il progresse grâce à l'enseignement offert, le plus souvent classique, sur la même variable pour rapprocher voire dépasser le groupe expérimental. Plonsky et Oswald (2014) ont rassemblé les résultats de 91 méta-analyses dans le domaine de l'acquisition L2 et trouvent des moyennes de 1,0 et 0,6 pour les deux protocoles respectifs ; leur suggestion est de proposer le quartile supérieur (c'est-à-dire, qui représente les 25 % des résultats les plus élevés) comme seuil pour parler d'un « grand » effet, à savoir 1,4 et 0,9.

Les résultats de la méta-analyse de Boulton et Cobb montrent un grand effet dans les deux catégories d'études : d = 1,50 entre les pré- et post-tests ; d = 0,95 entre les

groupes témoin et expérimental. Ainsi, de toutes les méta-analyses recensées par Plonsky et Oswald en acquisition L2, plus des trois quarts obtiennent des résultats inférieurs à l'ASC. Si ce constat global est plus qu'encourageant pour notre domaine, il serait extrêmement superficiel de vouloir réduire toutes les études à un seul chiffre qui ne saurait rendre compte de la variation inévitable entre elles. Une deuxième étape consiste alors à calculer la taille de l'effet pour différentes sous-catégories représentées. La conclusion est tout aussi frappante : « L'ASC fonctionne plutôt bien dans presque tous les contextes où il a fait l'objet de nombreuses évaluations » (p. 386) avec un effet de taille moyenne ou grande. Si les études obtiennent des résultats variables, pour l'instant il n'a pas été possible d'identifier les raisons qui l'expliquent. Tout enseignant souhaitant tenter l'expérience peut se sentir encouragé à le faire, tout en restant sensible au contexte local, aux participants et à leurs besoins, pour adopter et adapter l'ASC selon les besoins.

Il est à noter que la quasi-totalité des études dans cette méta-analyse porte sur l'anglais comme langue cible et se déroule dans un milieu universitaire. Pour les apprenants jeunes en contexte scolaire, il n'y a que 4 études intra-groupe et 6 inter-groupe dont 4 à Taiwan et d'autres en Allemagne, en Grèce ou au Portugal. Si celles-ci témoignent encore d'un grand effet (1,56 et 1,41 respectivement), il est clair qu'il nous faut plus de recherches auprès de ces jeunes publics. Il serait tentant de penser que le niveau de compétence en langue pourrait être proportionnel à l'âge des participants : pour les 8 et 7 études auprès d'apprenants de niveaux « faible » et pré-intermédiaire, on retrouve encore un effet de grande taille (1,10 et 1,72 respectivement). Toutefois, il est notoirement difficile de se fier aux niveaux affichés dans les publications en didactique. Burston et Arispe (2016), dans leur analyse des recherches en apprentissage des langues assisté par ordinateur, découvrent que dans 50 % des cas, les publics de niveau « avancé » correspondent au plus à un niveau B1 du CECR.

Conclusions

L'approche ASC représente un potentiel important pour l'apprentissage des langues étrangères et s'appuie sur des concepts d'actualité, dont l'apprendre à apprendre et la transférabilité des savoirs et savoir-faire, l'autonomie et l'apprentissage tout au long de la vie, l'individualisation et le constructivisme, l'authenticité et la contextualisation des données langagières, et bien sûr les TIC. Sous sa forme prototypique, l'ASC permet un travail pointu pour les utilisateurs qui ont des besoins précis et pérennes, qui peuvent investir un peu de temps pour s'approprier la technologie (les concordanciers ou autres logiciels) et qui ont un niveau en langue et une maturité nécessaires pour maîtriser les techniques de formulation des requêtes et d'interprétation des données récupérées. En même temps, on trouve des parallèles dans bon nombre d'outils simples déjà connus des apprenants. Ceux-ci peuvent servir de tremplin vers l'ASC classique pour certains profils d'apprenants (voire d'enseignants), mais le travail apporte des bénéfices immédiats et ces outils peuvent suffire à d'autres publics.

Avec l'ASC, la langue – et surtout la lexico-grammaire – retrouve sa place au centre de la scène. Qui plus est, l'approche permet de dépasser une vision normative de la langue pour dévoiler sa vraie nature floue et probabiliste. Tout aussi central est l'apprenant qui joue un rôle plus actif, qui travaille avec une langue authentique, forme et teste des hypothèses pour repérer les tendances pour devenir plus sensible et « meilleur apprenant ». Tout ceci n'est pas sans conséquences pour l'enseignant qui abandonne son rôle

traditionnel tout-puissant, mais peut retrouver un certain réconfort dans la reconnaissance qu'il n'est pas possible de tout savoir – ni sur la langue, ni sur les processus d'apprentissage. De cette façon, l'ASC est bien plus qu'un outil et relève presque d'une philosophie qui a toute sa place dans la formation initiale de l'enseignant qui, tout comme Johns (1990, p. 19), peut enfin admettre à ses étudiants : je ne suis pas sûr, allons voir ensemble.

Références bibliographiques

ASTON, Guy. The British National Corpus as a language learner resource. In BOTLEY, Simon, Julia GLASS, Anthony McENERY et Andrew WILSON. *Proceedings of TaLC 1996. UCREL Technical Papers*. 1996, vol. 9, p. 178-191.

BENNETT, Gena R. *Using corpora in the language learning classroom*. Ann Arbor: University of Michigan Press, 2010. 134 p.

BOULTON, Alex. Applying data-driven learning to the web. In LEŃKO-SZYMAŃSKA, Agnieszka et Alex BOULTON. *Multiple affordances of language corpora for data-driven learning*. Amsterdam : John Benjamins, 2015, p. 267-295.

BOULTON, Alex. Research timeline : corpora in language teaching and learning. *Language Teaching*. 2017, vol. 50, p. 483-506.

BOULTON, Alex et Tom COBB. Corpus use in language learning : a meta-analysis. *Language Learning*. 2017, vol. 67, p. 348-393.

BOULTON, Alex et Henry TYNE. *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues*. Paris : Didier, 2014. 309 p.

BURSTON, Jack et Kelly ARISPE. The contribution of CALL to advanced-level foreign/second language instruction. In PAPADIMA-SOPHOCLEOUS, Salomi, Linda BRADLEY et Sylvie THOUËSNY. *CALL communities and culture*. Dublin : Research-publishing.net, 2016, p. 61-68.

GEILLER, Luc. How EFL students can use Google to correct their "untreatable" written errors. EUROCALL Review. 2014, vol. 22, p. 26-45.

JOHNS, Tim. From printout to handout : grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*. 1990, vol. 10, p. 14-34.

JOHNS, Tim. Contexts: the background, development and trialling of a concordance-based CALL program. In WICHMANN, Anne, Steven FLIGELSTONE, Tony McENERY et Gerry KNOWLES. *Teaching and language corpora*. Harlow : Addison Wesley Longman, 1997, p. 100-115.

KUSYK, Meryl et Geoffrey SOCKETT. From informal resource usage to incidental language acquisition : language uptake from online television viewing in English. *ASp*. 2012, vol. 62, p. 45-65.

PLONSKY, Luke et Frederick OSWALD. How big is 'big'? Interpreting effect sizes in L2 research. *Language learning*, 2014, vol. 64, p. 878-912.

REPPEN, Randi. *Using corpora in the classroom*. Cambridge : Cambridge University Press, 2010. 104 p.

SWAN, Michael. *Practical English usage* (4e édition). Oxford : Oxford University Press, 2016. 768 p.