

**Quels corpus pour quels objectifs d'apprentissage ?
Étude de l'utilisation comparée de corpus, par des
apprenants LANSAD pour la rédaction de courriels
professionnels.**

Corinne Landure

Université de Lorraine - IUT d'Epinal, Laboratoire ATILF-CNRS

Mots-clés

Apprentissage sur corpus - corpus – générique – spécifique - langue de spécialité

Keywords

Data-driven learning – corpus - generic - specific - language for specific purposes

Résumé

Cette étude comparative rend compte de l'évolution des consultations provenant respectivement d'un corpus générique, d'un corpus spécifique et d'un dictionnaire en ligne, et permet ainsi d'en déterminer l'impact sur les productions écrites d'apprenants LANSAD, notamment pour la production de courriels professionnels en anglais. Notre choix portant sur la méthodologie d'analyses mixtes permet de croiser les résultats quantitatifs, provenant des analyses longitudinales et séquentielles des consultations et des productions, avec les résultats qualitatifs, provenant des comportements et commentaires des étudiants. Cette étude nous informe de l'efficacité des consultations effectuées sur le corpus spécifique qui ont été plus ciblées et raisonnées que celles accomplies sur le corpus générique. Elle souligne également l'implication régulière des apprenants exploitant les corpus et leur motivation constante tout au long de cette expérimentation.

Abstract

This comparative study reports on the development of learners' consultation practices using corpora or traditional resources, and provides insights into the impact of the approach on their writing, more specifically for producing emails in English with a professional objective. The mixed methods approach triangulates the quantitative results from longitudinal and sequential analyses of the consultations and the productions with the qualitative results of learners' behavior and comments. This study highlights the relative efficiency of the specific corpus in terms of consultations, thanks to more targeted and reasoned ones. It also shows that corpus use led to greater investment and involvement of the learners, as well as increased levels of motivation throughout the experiment.

Introduction

Le *data-driven learning*, ainsi baptisé par Johns en 1991 et nouvellement francisé en « apprentissage sur corpus » (ASC par Boulton & Tyne (2014)), est une approche centrée sur l'apprenant qui devient lui-même acteur et responsable de son apprentissage dans la mesure où il exploite directement les données des corpus, cible ses propres recherches en fonction de ses besoins langagiers, linguistiques et de ses objectifs, et interprète lui-même les résultats de ses recherches.

Cette approche se caractérise notamment par sa souplesse d'utilisation qui se reflète dans ses concepts, ses aspects, ses usages. L'apprentissage sur corpus se décline également en fonction de différentes variables : publics, besoins, objectifs, spécialités, niveaux, etc. L'exposition à de nombreuses données authentiques, leur exploration à l'aide d'un concordancier, l'individualisation des consultations, l'apprentissage par la découverte, l'observation de la langue en cotexte (ou contexte minimal, c'est-à-dire l'environnement immédiat d'un item) et en contexte (environnement plus large), le repérage des régularités, la prise de conscience linguistique et l'induction concourent à l'efficacité de l'apprentissage mais également au développement de l'autonomie de l'apprenant en favorisant la prise en charge par l'apprenant de son apprentissage et en augmentant sa motivation (Duda & Tyne, 2010). L'ASC n'est pourtant pas la « révolution » annoncée (Johns, 1990 : 27), pas plus qu'une rébellion contre les méthodes dites traditionnelles, mais il s'apparente davantage à une évolution dans la manière de considérer l'apprentissage et l'apprenant ainsi que l'enseignement et l'enseignant. Il n'a donc pas vocation à se substituer aux approches et méthodes traditionnelles. En revanche, il peut les compléter, apporter une complémentarité afin de faciliter l'apprentissage d'une L2. L'objectif d'une telle approche est d'éliminer les intermédiaires, « *attempt to cut out the middleman as far as possible* » (Johns, 1990 : 30), pour ainsi donner à l'apprenant les moyens de conduire son propre apprentissage par une participation active passant par la découverte, l'observation et l'induction linguistique.

L'objet de cette étude vise à évaluer de quelles façons des apprenants LANSAD (LANGues pour Spécialistes d'Autres Disciplines) utilisent directement les corpus afin d'améliorer leurs productions écrites à visée professionnelle. Nous émettons deux hypothèses : l'utilisation de corpus comme ressource de référence et d'aide à la production écrite serait plus efficace que l'utilisation d'une ressource plus conventionnelle, mais aussi l'exploitation d'un corpus spécifique au domaine et au genre correspondant aux besoins de l'apprenant LANSAD pourrait s'avérer plus efficace que la consultation d'un corpus générique. Nous supposons qu'un corpus qui combine à la fois la langue de spécialité et le genre est un facteur de motivation parce qu'il répond aux intérêts des apprenants et à leurs besoins spécifiques.

Nous nous attacherons à vérifier ces hypothèses en répondant aux questions suivantes :

Comment évoluent au fil des séances les consultations provenant respectivement de l'utilisation d'un dictionnaire bilingue, d'un corpus générique et d'un corpus spécifique et quel en est l'impact sur les productions écrites des apprenants ?

Quelles tendances se profilent au regard des comportements et commentaires individuels et collectifs tout au long de cette expérience ?

1. Méthode

1.1. Participants

La partie de l'expérience que nous présentons dans cet article s'échelonne sur quatre séances de travaux pratiques (TP) de deux heures et a été conduite dans un IUT, composante de l'Université de Lorraine. Les 97 étudiants âgés de 18 à 20 ans participant à cette étude sont en première année de Techniques de Commercialisation (TC) et ont à leur actif entre cinq et huit années d'anglais. La spécificité de cette population réside autant dans l'hétérogénéité de leur parcours scolaire que de leur niveau d'anglais. En effet, ces étudiants aux parcours d'études très diversifiés sont titulaires d'un baccalauréat général, technique ou professionnel. D'autre part, les différences de niveau en anglais sont assez conséquentes au vu de l'amplitude des résultats obtenus à un test informel du TOEIC administré au début de l'année universitaire (ceux-ci s'échelonnant de 130 à 645 points sur 990). Le niveau moyen en anglais est modeste et s'élève à 287 sur 990 points, ce qui correspond approximativement au niveau A2 du CECR. Sur les 97 étudiants répartis en 10 groupes de TP, nous n'avons pris en compte que les résultats des étudiants présents aux quatre séances dédiées à l'expérimentation, soit 67 participants. Nous avons alloué à chaque groupe de façon aléatoire une ressource à consulter : 4 groupes (29 étudiants) ont exploité le corpus générique (COCA), 3 groupes (20 étudiants) ont utilisé le corpus spécifique (SPE) et 3 groupes (18 étudiants) le dictionnaire en ligne *WordReference* (WR).

1.2. Instruments

Trois instruments de recherche ont été utilisés pour la partie de l'étude présentée ici. Il s'agit en premier lieu de résultats linguistiques comparant, sur trois séances effectuées à quelques jours d'intervalle, l'impact de la consultation directe d'un dictionnaire bilingue, d'un corpus générique et d'un corpus spécifique pour la rédaction de courriels professionnels. L'analyse qualitative, quant à elle, s'appuie sur des réponses à un questionnaire complété par les étudiants évaluant l'utilité et la facilité d'utilisation de la ressource allouée, le niveau de collaboration au sein du groupe et de sollicitations d'aide auprès de l'enseignante, et enfin le ressenti des étudiants en termes d'apprentissage. L'analyse se fonde également sur les comportements et commentaires des apprenants relevés et enregistrés tout au long de cette expérimentation.

1.3. Déroulement

La partie de l'expérimentation présentée plus principalement dans cet article s'est déroulée sur 10 semaines environ, l'intervalle entre chaque séance (S1, S2, S3, S4) étant de 4 à 15 jours selon les groupes. La première séance (S1) se concentre sur l'initiation pratique à la ressource assignée à chaque groupe. Après une brève présentation théorique de celle-ci, les étudiants la découvrent à travers une formation pratique et directe, supervisée et guidée par l'enseignante. Ils peuvent ainsi se

familiariser avec l'interface et les différentes fonctions de la ressource. Pour ce faire, chaque participant a un ordinateur et un accès à Internet. L'enjeu ne se situe pas au niveau des compétences technologiques, les étudiants ne sont pas « technophobes » et comprennent assez vite les différentes fonctionnalités. En revanche, la difficulté réside dans la manière de se servir de la ressource comme aide efficace à la rédaction ; en d'autres termes, en quoi les différentes fonctions vont-elles pouvoir concrètement les aider dans l'achèvement de leur tâche ? Le véritable défi pour l'enseignante, lors de cette séance de formation, est de les faire réfléchir sur la langue, son utilisation et son usage, ensuite de les amener à se poser les « bonnes » questions afin d'utiliser la ressource efficacement et ainsi de pouvoir mieux appréhender et interpréter les résultats. Lors de cette séance d'initiation, l'enseignante les encourage, aide, guide, conseille et leur propose notamment de suivre une feuille de route, un peu à la manière des différentes phases proposées par Holec (1990) pour promouvoir l'autonomie de l'apprenant, intégrant ici les étapes nécessaires pour une consultation de corpus fructueuse :

1. Cibler ses consultations: savoir ce que l'on cherche
2. Sélectionner la fonction adéquate pour une consultation pertinente
3. Observer le cotexte (environnement immédiat) et le contexte (environnement plus large)
4. Lister ses observations pour en dégager les régularités et les spécificités
5. Induire : en tirer des conclusions
6. Vérifier pour valider
7. Généraliser quand cela est possible
8. Adopter en sachant pourquoi

Pour les séances suivantes (S2, S3 et S4), chaque groupe se voit attribuer des tâches identiques. Ces tâches consistent en la rédaction de courriels professionnels liés à leur domaine de spécialité en suivant les scénarios préalablement établis par l'enseignante (annexe 1). Il s'agit d'en suivre les grandes lignes, ce qui laisse une certaine latitude à l'étudiant quant aux choix des formulations, du registre et de la forme à adopter. L'un des objectifs est de savoir s'ils vont attacher autant d'importance au contenu du message qu'à sa dimension sociale et culturelle (usage, formulations, etc.). Nous avons exclu toute intervention ou participation de notre part pouvant influencer les résultats de l'expérimentation, tant au niveau des consultations que des résultats des productions.

La deuxième séance est dédiée à la rédaction du premier courriel (C1). Les étudiants ont pour tâche de rédiger le courriel sur Word en s'aidant uniquement de la ressource qui leur a été attribuée. Ils ont pour consigne ensuite de l'envoyer à l'enseignante avec l'historique des consultations (H1) qu'ils ont préalablement copié-collé.

La troisième séance concerne à nouveau la rédaction du premier courriel (C1), les productions du C1 sont indépendantes les unes des autres afin de pouvoir évaluer les modifications linguistiques apportées par l'étudiant d'une séance à l'autre, et concerne également la rédaction d'un deuxième courriel (C2). Ils sont invités à la fin de

la séance à faire parvenir à l'enseignante les courriels C1 et C2 ainsi que les historiques H1 et H2.

La quatrième séance consiste en la production d'un troisième courriel (C3) et porte également sur la rédaction du premier et deuxième courriel (C1 et C2). Ils font également parvenir à l'enseignante les trois courriels (C1, C2, C3) accompagnés des trois historiques correspondants (H1, H2, H3).

1.4. Ressources

1.4.1. Dictionnaire bilingue en ligne : *WordReference*

Le dictionnaire bilingue *WordReference* (WR) donne la signification du mot recherché dans différents contextes et procure des exemples avec leur traduction. WR permet aussi l'accès à l'historique des mots recherchés par l'utilisateur. Cette option offre aux participants la possibilité d'enregistrer l'historique des items consultés après chaque séance et de l'envoyer par courriel à l'enseignante en même temps que leurs productions. De cette façon, il est possible de retracer le processus de consultation de chaque utilisateur lors des différentes séances (séances 2, 3 et 4) et ainsi d'évaluer les besoins des apprenants via les consultations. Nous envisageons d'examiner ce que celles-ci apportent au niveau des productions écrites successives. Trois groupes utilisent *WordReference* et nous servent ainsi de groupe de contrôle pour l'analyse comparative des résultats provenant des corpus générique et spécifique.

1.4.2. Corpus générique : COCA

COCA est attribué à 4 groupes. Ce corpus, gratuit et accessible en ligne, ne nécessite qu'une simple inscription pour un accès permanent au site. COCA est composé d'extraits de textes provenant de transcriptions orales, livres de fiction, magazines, journaux, écrits universitaires. On peut y trouver des fonctions de base comme *List* qui donne la fréquence, *Chart* qui indique la fréquence dans différentes sections et sous-sections, *Kwic* qui renseigne sur la classe grammaticale, *Compare* qui permet d'effectuer diverses comparaisons, et encore *Collocates* qui fournit les collocations potentielles du mot recherché. Ce site propose également des fonctions plus sophistiquées et de nouvelles applications sont ajoutées régulièrement. Dans tous les cas, l'apprenant peut choisir de n'utiliser que celles avec lesquelles il se sent à l'aise et qui répondent à ses besoins. COCA offre aussi la possibilité de recenser toutes les consultations dans un historique.

1.4.3. Corpus spécifique : *SPE* (corpus de courriels professionnels)

Notre choix d'employer le terme « corpus spécifique » tient au fait que nous souhaitons éviter toute ambiguïté avec l'appellation « corpus spécialisé » qui désigne les « corpus de discours académique » (Chambers, 2010 : 11). Les « corpus spécifiques » quant à eux représentent le genre discursif se rapportant à toute activité linguistique à visée professionnelle, excluant toutes activités linguistiques universitaires (dissertations, articles, etc.) et la langue générale.

Pour rappel, l'hypothèse émise préalablement sous-tend que l'utilisation d'un corpus spécifique au domaine et aux besoins de l'apprenant LANSAD pourrait s'avérer plus probant que la consultation d'un corpus générique. Afin de pouvoir vérifier cette hypothèse, il convient tout d'abord de créer un corpus de courriels qui sera associé à un concordancier. Aucune entreprise contactée, pour des raisons de confidentialité, n'est disposée à nous faire part de ses courriels. La seule solution pour mener à bien notre projet est alors de trouver suffisamment de courriels libres de droit, consultables via Internet. Notre recherche s'est révélée doublement fructueuse puisque nous avons pu obtenir un corpus constitué non seulement de courriels en langue anglaise mais aussi de courriels professionnels. Ce corpus de courriels professionnels a été constitué en 2004 et 2005 par un chercheur américain Y. Wu impliqué dans des travaux ayant pour objectif l'évaluation des systèmes d'extraction d'informations. Nous avons contacté ce chercheur pour l'informer de notre projet et lui demander l'autorisation d'utiliser dans le cadre de nos recherches les 12 656 courriels compilés par ses soins.

Pour exploiter ce corpus de courriels qui consiste en 2 933 544 tokens dont 34 920 types, nous utilisons le logiciel *AntConc*, créé par L. Anthony³⁴. Ce choix est motivé par deux raisons essentiellement : sa gratuité et sa relative facilité d'utilisation pour des utilisateurs non aguerris à l'utilisation de concordanciers. De plus, il propose toute une gamme de fonctionnalités : concordances (recherche par mot), *concordance plot* (emplacement du mot dans les fichiers), collocations (associations plus ou moins directes), *clusters* (localisation des mots récurrents), *word list* (liste de mots et information sur leur fréquence), etc. Les principes de consultation de base sont identiques aux corpus BYU dont COCA. À la différence de WR et COCA, il n'y a pas de fonction générant automatiquement l'historique des consultations effectuées pendant la séance. Il est donc demandé aux apprenants de recopier au fur et à mesure leurs recherches sur un document Word.

L'un des objectifs est d'estimer l'effet de chacune de ces ressources en matière de consultation et d'aide à la rédaction en procédant à l'étude comparative de leur utilisation. Il va de soi que dans un contexte normal, ces mêmes étudiants auraient sans doute combiné plusieurs ressources afin de réaliser leur tâche. Mais notre étude a pour objectif principal d'évaluer ce qu'apporte effectivement la consultation de chacune d'entre elles à la production écrite.

2. Résultats

2.1. Consultations

Pour l'analyse des données des consultations, nous avons utilisé le site *Compleat Lexical Tutor*³⁵ créé par Tom Cobb. Ce site gratuit et accessible en ligne est destiné à tout public : apprenants, enseignants et chercheurs. Il propose plusieurs logiciels à des fins pédagogiques ou de recherches. Nous avons utilisé certaines fonctions, notamment *Text-Based Concordances* et *Text-lex Compare*, que nous avons adaptées pour l'analyse des données des corpus de courriels et d'historiques.

³⁴AntConc. Concordancier : <http://www.antlab.sci.waseda.ac.jp/software.html>

³⁵LexTutor. T. Cobb. <http://www.lextutor.ca/vp/eng>

2.1.1. Analyse comparative du nombre de consultations en fonction des ressources

Nous comptabilisons le nombre de consultations et obtenons pour COCA un total de 4 051 consultations pour 29 étudiants, soit une moyenne de 140 consultations. Pour le corpus SPE, nous relevons un total de 860 consultations pour 20 participants, soit une moyenne de 43 consultations. Pour le dictionnaire WR, nous en dénombrons 947 pour 18 étudiants, soit une moyenne de 53 consultations. Nous remarquons qu'il y a environ trois fois plus de consultations sur COCA que sur le corpus SPE et que le nombre de consultations du groupe de contrôle se rapproche davantage de celui du corpus SPE que de COCA. Comment expliquer un tel écart du nombre de consultations entre les corpus générique et spécifique ?

2.1.2. Analyse longitudinale des consultations

Afin d'aller plus loin dans l'analyse, nous comptabilisons les consultations des six productions des trois ressources. On constate que les moyennes des recherches sur COCA sont plus importantes en nombre pour chacune des six productions (fig. 1). On note que la moyenne des consultations portant sur le premier jet du premier courriel atteint un pic de 54 items consultés, soit quatre fois plus que pour le corpus SPE (13) et environ trois fois plus que pour le dictionnaire (17). Le deuxième pic concerne la première production du deuxième courriel, avec 24 requêtes pour COCA et 11 chacune pour le corpus SPE et WR. Le troisième pic porte sur la dernière séance comprenant l'unique rédaction du troisième courriel. On peut en déduire que chaque nouveau courriel a nécessité des consultations plus nombreuses, et ce pour les trois ressources.

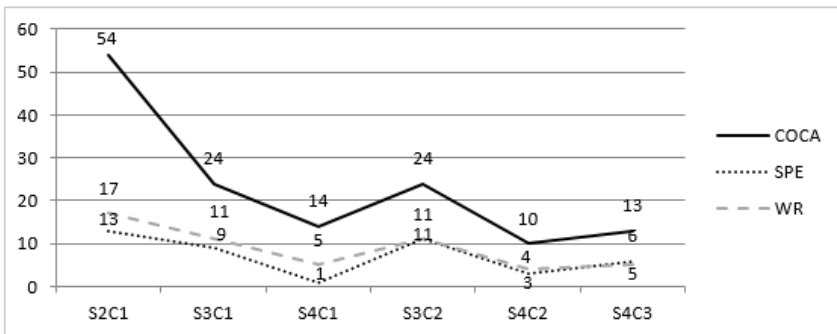


Figure 1 : Analyse longitudinale du nombre moyen des consultations

On note également que la dégressivité des consultations se fait en fonction de la répétition de la même tâche dans le temps. Répéter la même tâche aurait un effet sur la baisse du nombre de consultations, ce qui laisse entendre que les recherches effectuées précédemment ont laissé une trace dans la mémoire des apprenants. En revanche, même s'il y a eu moins de recherches, les apprenants ont tout de même

effectué des consultations, peut-être par nécessité de conforter leurs connaissances, de consulter un point particulier, etc.

2.1.3. Analyse séquentielle des consultations

Nous choisissons de prendre un autre angle d'observation des données de consultations en les analysant de manière séquentielle, en fonction des séances. Bien entendu, nous ne pouvons pas comparer le nombre de consultations du C1 avec celles du C2 ou C3, puisque le texte est différent et plus long dans le C1 que dans les deux autres. En revanche, nous pouvons comparer les consultations des différents courriels d'une même séance en fonction de la ressource utilisée. La moyenne des consultations des séances 2 (C1), 3 (C1+C2) et 4 (C1+C2+C3) est plus importante pour COCA, alors que les deux autres ressources obtiennent des résultats assez similaires (Fig. 2).

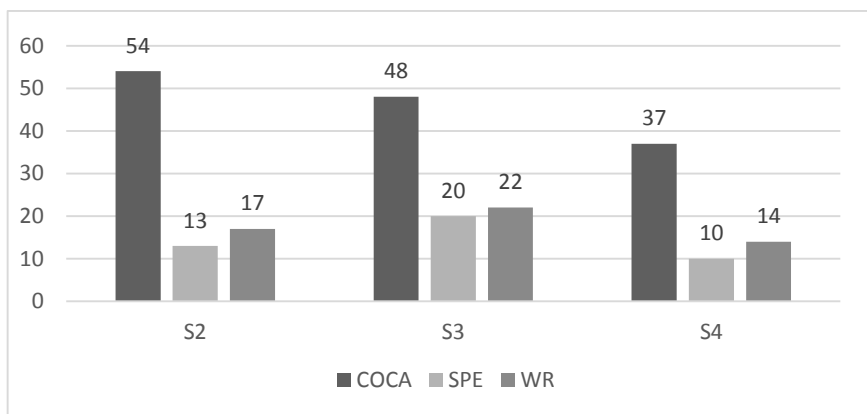


Figure 2 : Nombre moyen de consultations par séquence

Alors qu'on observe une dégressivité du nombre moyen de consultations d'une séance à l'autre pour COCA, on note davantage de consultations au S3 par rapport au S2 et S4 pour le corpus SPE et WR. Cela peut sembler normal étant donné qu'il y a deux courriels à rédiger au S3 au lieu d'un seul au S2. Mais alors comment expliquer qu'au S4 le nombre de consultations chute, tant pour le corpus SPE que pour WR, alors qu'il y a trois courriels à produire ? Nous pouvons émettre au moins deux hypothèses pour expliquer ce phénomène. La première se rapporterait au niveau de difficulté du courriel 2, ce qui aurait pu nécessiter davantage de consultations. La deuxième concernerait le premier courriel qui aurait requis à nouveau l'attention des apprenants avec pour conséquence de nouvelles recherches. En revanche, nous ne parvenons toujours pas à expliquer l'écart entre le nombre de consultations effectuées sur COCA et celles effectuées sur le corpus SPE et WR.

2.2. Production

Afin de pouvoir étudier les productions des apprenants, nous concevons un corpus de courriels d'apprenants reprenant toutes leurs productions. Ce corpus comprend environ 26 000 mots au total, soit 7 125 mots pour WR, 11 379 mots pour COCA et 7 494 mots pour le corpus SPE. Nous exploitons ce corpus avec l'outil *Lex Tutor* en téléchargeant les documents dans la fonction *Text-Based Concordances*³⁶. Cette fonction permet de déterminer le nombre total de mots. Elle propose également un index des concordances pour chaque mot avec une classification des mots-types par ordre alphabétique en précisant également leur fréquence dans le document. Il suffit de cliquer sur un mot dans cette liste pour qu'il apparaisse dans les lignes de concordance. Nous répertorions ensuite pour chaque fichier le nombre de tokens ainsi que le nombre de types. On constate que le nombre de mots du premier courriel (S2C1, S3C1, S4C1), toutes ressources confondues, est plus important que pour les autres courriels, ce qui est logique puisque le scénario du C1 est plus long. On remarque cependant une dégressivité du nombre de mots des productions C1 au fil des séances pour toutes les ressources. Cela voudrait dire que la rédaction devient plus concise, soit en adoptant une formulation plus précise ou fluide ou au contraire en omettant certains éléments du message.

2.2.1. Nombre moyen de mots par production et par ressource

Afin de pouvoir comparer le nombre de mots par production et par ressource, nous calculons la moyenne des mots produits par le nombre d'étudiants. La Figure 3 montre une production en termes de mots assez cohérente quelle que soit la ressource ; nous notons cependant une production supérieure pour les C1 avec le dictionnaire et une production supérieure pour les C2 et C3 avec COCA.

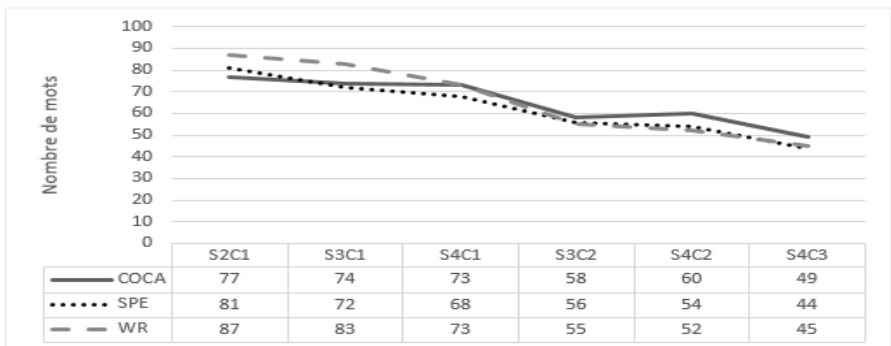


Figure 3 : Moyenne du nombre de mots / production

³⁶ http://www.lextutor.ca/concordancers/text_concord/

Si nous prenons en compte la moyenne des mots de toutes les productions, nous obtenons respectivement 392 mots pour COCA, 375 mots pour SPE et 396 mots pour WR, soit une différence maximum de 5%. Nous pouvons en conclure qu'il y a une certaine cohérence dans le nombre de mots des productions quelle que soit la ressource utilisée.

2.2.2. Rapport entre le nombre moyen de consultations et le nombre moyen de mots des productions

Nous remarquons (Fig. 4) une disproportion du rapport moyen du nombre de consultations sur COCA au regard du nombre de mots des productions à l'aide de COCA par rapport aux autres ressources. Le rapport moyen est de 35% pour COCA (139 consultations pour 392 mots produits), 11.5% pour le corpus SPE (43 consultations pour 375 mots produits) et 13 % pour le dictionnaire WR (53 consultations pour 396 mots produits).

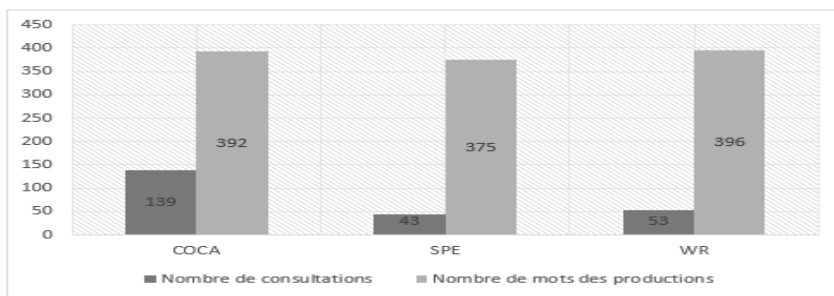


Figure 4 : Moyenne du nombre des consultations / la moyenne du nombre des mots des productions

On peut avancer que seule la spécificité de COCA pourrait expliquer la différence conséquente du nombre de consultations alors que le nombre de mots des productions est similaire aux autres ressources. Nous tenterons de déterminer, à l'aide des études qualitatives, quelles caractéristiques propres à cette ressource peuvent expliquer une telle disparité et dans quelle mesure la qualité de la production pourrait s'en trouver améliorée puisque, comme nous l'avons vu, il n'y pas d'incidence sur la quantité, c'est-à-dire le nombre de mots des productions. On pourrait donc s'attendre à ce que cette différence profite à la qualité des productions. On pourrait également se demander si COCA a été sur-consulté. Quelles en seraient les raisons et les conséquences sur la qualité des productions et l'efficacité de l'apprentissage ? On peut tout aussi bien se poser la question inverse et se demander pourquoi le corpus SPE et WR n'ont pas été davantage consultés, pour quelles raisons et avec quelles conséquences sur les productions et l'apprentissage ?

2.2.3. Rapport type/token

En comparant les mots-types (*types*) par rapport aux mots des productions (*tokens*), nous pouvons évaluer la diversité du lexique. Nous procédons au calcul de ce ratio pour toutes les productions grâce à la fonction de *Text-Based Concordances de Lextutor* qui nous fournit le nombre de mots (*tokens*) et les mots-types (*types*). COCA obtient un ratio *type/token* de 16%. On constate que le corpus SPE atteint 20% et le corpus dictionnaire 24%. Pour autant, que le rapport *type/token* soit plus important en utilisant WR (24%) n'est pas très étonnant en soi puisqu'il s'agit d'un dictionnaire bilingue et que les mots recherchés le sont principalement dans la langue maternelle de l'apprenant. Il est donc difficile dans ce cas de comparer l'efficacité du dictionnaire avec celle des corpus, puisque les recherches sur corpus sont faites dans la langue cible.

En revanche, nous pouvons comparer les deux corpus : COCA (corpus générique) et SPE (corpus spécifique). On s'aperçoit que le ratio *type/token* est plus important pour le corpus SPE que pour COCA et ce pour toutes les productions (Fig. 5).

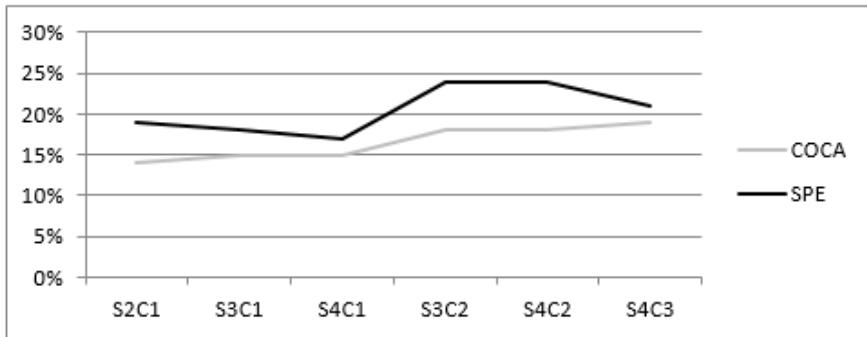


Figure 5 : Rapport *type/token*

On constate également une légère baisse lors des deuxième et troisième rédactions du premier courriel pour le corpus SPE, avec un pic pour le deuxième courriel et à nouveau une légère baisse lors de la rédaction du troisième courriel. Pour COCA, la progression est constante sur toutes les séances.

2.2.4. Étude comparative des consultations et des productions de COCA et du corpus SPE

Afin d'affiner nos résultats, nous comparons les items des consultations effectuées sur COCA et sur le corpus SPE avec ceux des productions des différentes séances. Nous ne pouvons donc pas prendre en compte, pour cette partie de l'étude, le dictionnaire *WordReference*, étant donné que les consultations effectuées sur cette ressource l'ont été majoritairement en français. En fait, cette étude comparative se fonde sur le rapport des consultations et des productions effectuées dans la langue cible. Pour ce faire, nous reprenons le sous-corpus de production de courriels et nous le rapprochons du sous-corpus des historiques de consultation. Ces derniers reprennent toutes les consultations des étudiants en fonction du courriel, de la séance et de la

ressource. Pour réaliser ces sous-corpus, nous procédons à un copier-coller des consultations répertoriées dans l'historique correspondant à chaque production et envoyé par les étudiants à l'enseignante après chaque séance. Nous attirons l'attention sur le fait que le nombre de consultations n'est pas équivalent au nombre de mots car une même consultation peut comporter une chaîne de mots, comme : *is possible, come with, at the same hour...*

2.2.5. Outils d'analyse quantitative des productions

Pour rapprocher un corpus d'historiques de consultations d'un corpus de productions de courriels, nous utilisons la fonction *Tex Lex Compare* de *Lextutor*. Elle nous permet d'obtenir des données précises et des indices fiables notamment au niveau des historiques de consultations et des productions de courriels. Nous pouvons ainsi déterminer les items consultés que l'on retrouve dans les productions (c/p), les items consultés qui y sont absents, c'est-à-dire non produits (c/np) ainsi que les items non consultés mais présents dans les productions (nc/p). Cette fonction donne également des indications sur le comportement des apprenants lors de la phase de consultation et permet ainsi de mieux appréhender leur manière de procéder, autant lors de cette phase que lors de la phase de production. En principe, s'il y a consultation, c'est parce que l'étudiant veut conforter une intuition, vérifier une connaissance ou trouver une réponse à sa question. Si la recherche y répond de façon satisfaisante, l'item sera recyclé dans la production et apparaîtra dans la colonne *Shared* de la fonction *Tex Lex Compare*, si au contraire la recherche n'aboutit pas ou que l'apprenant ne trouve pas de réponse appropriée, le mot consulté sera abandonné, donc absent de la production, et nous le trouverons dans *Unique to first*. Enfin, nous pouvons également prendre connaissance de la liste des mots présents dans les productions mais qui n'ont pas été consultés dans *Unique to second*. Il peut s'agir d'acquis antérieurs à l'étude pour les premiers jets, c'est-à-dire principalement pour les productions S2C1, S3C2 et S4C3, dont les items seraient connus ou prétendument connus des apprenants avant l'expérimentation ou encore d'acquis provenant de précédentes consultations, notamment pour les rédactions du S3C1, S4C1 et S4C2.

2.2.6. Étude comparative de l'impact des consultations sur les productions

Pour estimer l'impact des consultations sur les productions, nous procédons à l'étude comparative des différentes catégories (c/np, c/p, nc/p) par ressource (Fig. 6) et commençons avec la moyenne des items consultés non produits (c/np).

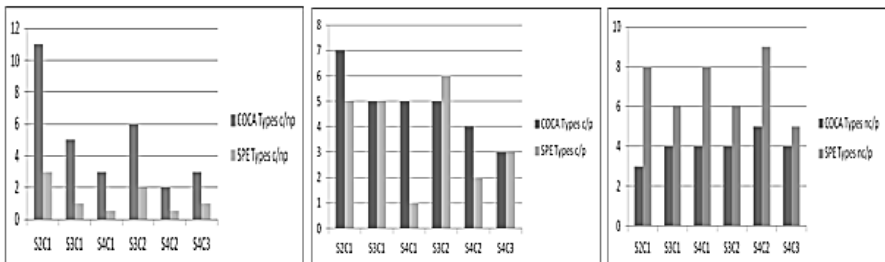


Figure 6 : Moyenne des mots-types c/np, c/p et nc/p / production

On constate d'emblée un plus grand nombre de types *c/np* pour COCA que pour le corpus SPE et ce pour toutes les séries de production. Nous pouvons avancer plusieurs explications à ce phénomène, la première étant que si les consultations ont été plus nombreuses, elles ont peut-être été aussi plus rapides et moins ciblées. Cela signifierait alors que les étudiants ont enchaîné les consultations sans prendre le temps de réfléchir à leurs objectifs de recherche, ni prendre le temps d'interpréter les données. Inversement, on peut imaginer que si les consultations ont été nombreuses, c'est notamment parce que les étudiants sont performants et s'impliquent complètement dans cette tâche. On peut aussi attribuer cette prolifération de consultations aux qualités de la ressource, à ses fonctions simples et multiples qui facilitent la recherche et l'interprétation des données, et qui fournissent des résultats pertinents et rapides. Ou, au contraire, on peut mettre l'accent sur les limites de la ressource, par exemple, la complexité de ses fonctions qui ne permet pas toujours une compréhension aisée des données, d'où des consultations répétitives, signes d'incompréhension des apprenants pour ce type d'approche. À ce stade de l'analyse, il est difficile de se prononcer pour l'une ou l'autre de ces hypothèses. On remarque cependant que les items consultés et non produits sont plus nombreux lors du premier jet. Cela veut bien dire que les items qui ont été consultés une première fois et non repris dans les productions ne font pas l'objet, pour 50% d'entre eux, d'une nouvelle consultation lors de la séance suivante pour un même courriel (Fig. 6). Donc, il y a bien une certaine cohérence dans les recherches effectuées par les apprenants, ce ne sont pas seulement des requêtes effectuées au hasard. Si les consultations sont plus ou moins bien ciblées, elles sont effectuées avec sérieux ; on ne peut donc mettre en cause un manque d'implication des étudiants ou un comportement négatif, sinon nous aurions relevé des incohérences à ce niveau.

En revanche, que penser du nombre peu élevé des types *c/np* du corpus SPE par rapport à COCA ? On pourrait émettre plusieurs hypothèses également, comme mettre en avant un comportement plus raisonné des apprenants qui se traduit par des recherches plus ciblées et pertinentes, ou au contraire, citer un manque d'investissement des apprenants pour l'approche sur corpus. On pourrait tout aussi bien mettre en cause la faible performance de la ressource, ses fonctions trop complexes, ses possibilités de consultation limitées, ou inversement louer ses atouts permettant une recherche efficace. Tout comme pour COCA, on remarque que les consultations non reprises dans les productions diminuent d'une séance à l'autre pour la rédaction d'un même courriel, ce qui traduit également une certaine cohérence dans les consultations effectuées par les apprenants.

En ce qui concerne les mots-types consultés et produits (*c/p*), on constate que l'écart entre COCA et le corpus SPE est peu important sauf pour la séance 4 correspondant au troisième jet du C1 (S4C1) et au deuxième jet du C2 (S4C2). Cela signifierait peut-être que les apprenants utilisant COCA auraient moins retenu des consultations précédentes ou auraient procédé à de nouvelles consultations.

Quant aux mots-types non consultés mais produits (*nc/p*), on s'aperçoit que pour toutes les productions la moyenne des *nc/p* est plus importante pour le corpus SPE que pour COCA. Cela signifie que pour les premiers jets (S2C1, S3C2 et S4C3), les

apprenants utilisant le corpus SPE se sont davantage appuyés sur leurs acquis antérieurs à l'étude que les apprenants utilisant COCA, ce qui peut en partie expliquer l'écart des consultations entre COCA et le corpus SPE que nous avons relevé précédemment. En revanche, pour les rédactions intermédiaires, c'est-à-dire les deuxièmes jets du S3C1 et S4C2 et le troisième jet du S4C1, il pourrait s'agir d'acquis provenant des séances précédentes, ce qui peut signifier que les apprenants utilisant le corpus SPE auraient mieux retenu, voire davantage profité en termes d'acquisition des consultations précédentes, peut-être aussi parce que celles-ci étaient moins nombreuses et donc plus ciblées.

2.2.7. Répartition lexicale des productions dans les listes de fréquence

Nous nous intéressons dans cette partie à évaluer le lexique composant les productions des étudiants grâce à la fonction *VocabProfile* de *LexTutor* qui nous fournit les listes de fréquence à partir d'un corpus. Dans notre cas, il s'agit des 20 listes représentant les 20 000 mots les plus fréquents du *British National Corpus* – BNC 20. Cette fonction permet de cibler plus facilement les mots peu courants d'un texte ou encore les mots appartenant à un domaine spécialisé et/ou un genre spécifique.

La Figure 7 récapitule le pourcentage moyen de mots-types répartis dans les listes de fréquence pour COCA et le corpus SPE. Quelle que soit la ressource, les taux de répartition des mots-types dans les listes de fréquence sont à peu près équivalents. On s'aperçoit que 73% des mots pour COCA et 74% des mots pour le corpus SPE se trouvent dans la liste des 1 000 mots les plus fréquents (K-1 Words) et que 11% (COCA) et 10% (corpus SPE) font partie de la liste des 2 000 (K-2 Words). Le restant se répartit essentiellement dans les cinq listes suivantes à hauteur de 1 à 3%. On remarque que la *Off-List* contient entre 10 et 11% des mots-types des productions, il s'agit pour la plupart de noms propres, de mots mal orthographiés, de fautes de frappe, de mots étrangers ou de mots inventés par les apprenants.

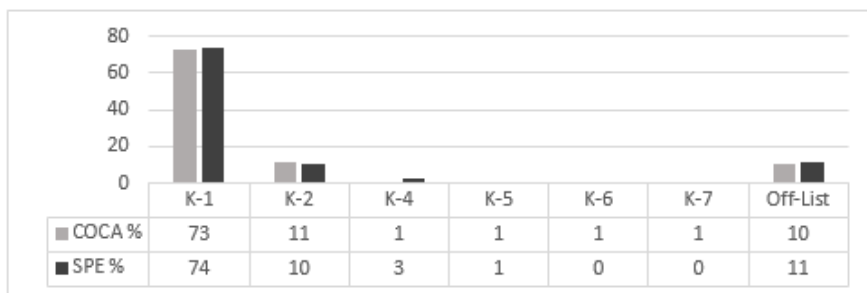


Figure 7 : Taux de répartition des mots-types dans les listes de fréquence

Pour conclure, à la lumière des données collectées (Fig. 8), on peut dire que le nombre de tokens et types consultés, de tokens et types produits et de consultations est

plus important pour COCA que pour le corpus SPE. On remarque pour autant que l'écart de mots-types produits provenant des deux ressources n'est pas très important au regard de la différence du nombre de types consultés. On peut s'étonner que le nombre de tokens soit plus important que le nombre de consultations, mais comme nous l'avons expliqué précédemment, une même consultation peut porter sur une chaîne de mots (par exemple, pour une consultation portant sur « *at the same hour* », nous avons quatre mots).

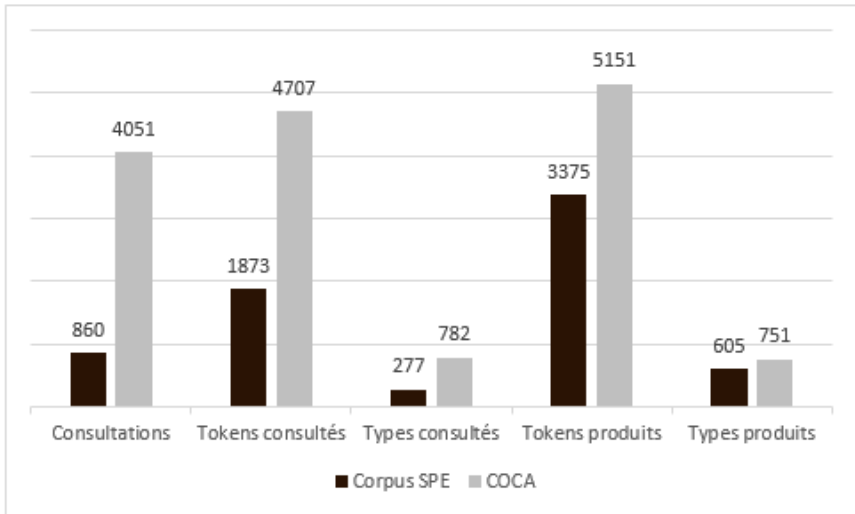


Figure 8 : Consultations et productions / ressource

Ces différences conséquentes du nombre de consultations, de tokens et de types consultés n'ont pas eu d'impact sur la richesse et la diversité lexicale de la production avec COCA puisque le ratio type/token est de 16% pour ce dernier, et de 20% pour le corpus SPE.

3. Réponses au questionnaire et commentaires des apprenants

À la fin de cette expérimentation, les étudiants ont complété un questionnaire à échelle de Likert paire évaluant la ressource utilisée (Fig. 9). En réponse à la question portant sur l'utilité de la ressource allouée pendant l'expérience, les étudiants ayant utilisé l'une ou l'autre des ressources l'ont trouvée unanimement très utile. Les corpus, qui sont pourtant nouveaux pour nos apprenants, ont ainsi obtenu un score identique au dictionnaire bilingue qui est une ressource plus traditionnelle et communément utilisée par les étudiants. Ceux-ci se sont ensuite prononcés sur le caractère accessible, aisé de la ressource. 92% des étudiants ayant utilisé le dictionnaire le trouvent facile d'usage, 80% des apprenants exploitant le corpus SPE l'estiment accessible, quant à COCA, 62%

le considèrent comme facile à utiliser. À la question portant sur la collaboration au sein du groupe, 63% des étudiants consultant le dictionnaire ont déclaré avoir collaboré avec les autres participants pendant les séances, le taux s'élève à 87% pour les apprenants utilisant le corpus et à 88% pour ceux travaillant sur COCA.

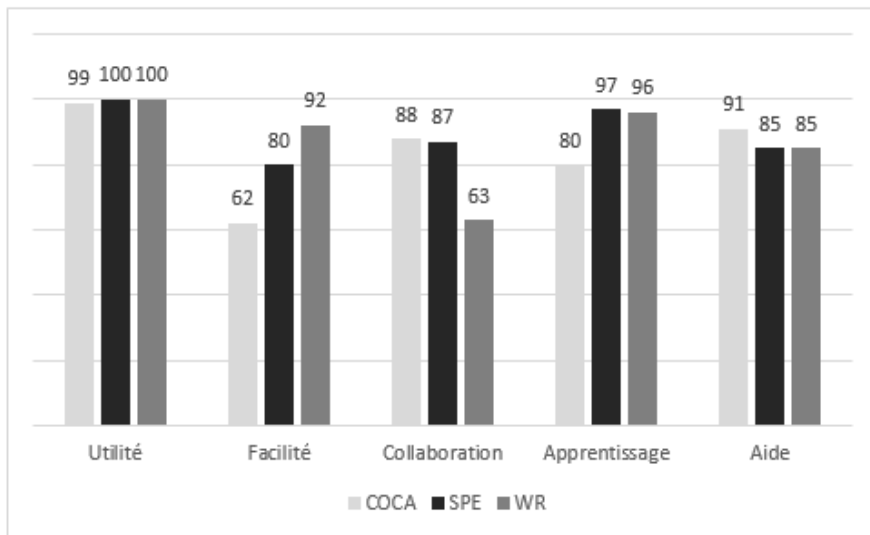


Figure 9 : Évaluation de la ressource utilisée

Quant à la question visant à évaluer le ressenti des étudiants en termes d'apprentissage, 96% des étudiants ayant utilisé *WordReference* estiment avoir appris lors de ces séances. Le résultat est pratiquement identique pour les étudiants ayant exploité le corpus SPE (97%) ; quant à COCA, 80% déclarent avoir le sentiment d'avoir appris. La question suivante concerne l'aide sollicitée auprès de l'enseignante lors des séances dédiées à la consultation. À la question s'ils avaient demandé de l'aide à l'enseignante, la réponse est positive pour 85% des étudiants utilisant le dictionnaire, elle est du même ordre pour les étudiants consultant le corpus SPE et de 91% pour COCA. Pour autant, l'aide qu'ils ont reçue n'a pas concerné les recherches ou les résultats de celles-ci. Elle s'est concentrée exclusivement sur les points techniques liés au matériel et sur le plan organisationnel lié à l'expérimentation, c'est-à-dire l'envoi des courriels avec les productions et les historiques principalement. Notre sentiment est que les sollicitations relevaient plus d'une attente de confirmation, d'un souhait de partage que d'un besoin réel d'assistance ou d'une réponse « d'expert ».

Les problèmes rencontrés par les étudiants exploitant les corpus ont concerné principalement la traduction. Comment trouver un mot en anglais lorsqu'on ne le connaît pas ou plus et qu'on ne peut pas le chercher dans un dictionnaire bilingue pour les

besoins de l'expérience ? Première stratégie constatée lors de cette expérience, l'étudiant demande à son voisin et, si ce dernier ne connaît pas la réponse, il demande au groupe. Deuxième stratégie, l'apprenant lance la requête avec le mot en français en espérant ainsi trouver l'existence de son équivalent transparent anglais. Troisième stratégie, il va chercher à l'exprimer autrement, trouver un synonyme, un mot dont le sens s'en rapproche. Quatrième stratégie, il essaie de trouver la réponse en mettant un mot qui lui est souvent associé. Cinquième stratégie utilisée par les étudiants exploitant le corpus SPE, ils lisent quelques courriels dans leur intégralité pour éventuellement y trouver des phrases ou formulations qui pourraient leur convenir et ainsi s'en inspirer pour leur propre production.

Lors de l'entretien collectif, la plupart des étudiants utilisant le corpus SPE ont approuvé l'étudiant 42 disant que « c'est une expérience très intéressante et à reproduire, je trouve ce cours très intéressant car il nous amène à réfléchir sur comment trouver quelque chose, c'est une bonne aide pour l'utilisation des mots ». Ils ont également souscrit à la déclaration de l'étudiant 35 : « au début, on a l'impression que le logiciel est vraiment compliqué mais au fur et à mesure que l'on pratique on s'y habitue. On apprend sur les méthodes de recherches et sur des tournures de phrases spécifiques en anglais ». Ils ont tous reconnu, comme l'a exprimé l'étudiante 8, que « ce logiciel est très enrichissant. Il m'a aidée dans l'apprentissage des tournures de phrases pour rédiger des courriels professionnels ».

Quant aux utilisateurs de COCA, ils trouvent la ressource utile, mais certains remarquent, comme l'étudiant 14, que « cette ressource est très intéressante mais difficile à utiliser quand on a peu de vocabulaire ». Un autre participant ajoute que « l'utilisation du corpus est très utile et nous aide beaucoup mais parfois lorsqu'on a vraiment besoin d'un mot et que nous ne le trouvons pas, cela peut être très gênant pour le travail ». L'étudiant 74, quant à lui, déclare que la ressource est « utile et agréable, elle reste cependant difficile d'accès. Je ne sais pas si j'utiliserai COCA pendant ma vie professionnelle, bien qu'il puisse servir de coup de pouce lors de la rédaction d'une lettre en anglais ». L'étudiant 79 pense que « la méthode est assez utile lorsqu'on doit rédiger, elle nous permet d'être autonome et de retenir ce que l'on ne savait pas ou de choisir une expression quand on hésite ».

4. Discussion

L'étude des consultations nous a renseignée sur les pratiques de recherche des apprenants utilisant les corpus. Si nous avons noté quelques différences de pratique au niveau des fonctions proposées selon le logiciel ou concordancier utilisé, la démarche est restée pratiquement identique. En règle générale, elle a débuté par la sélection de l'item à rechercher. Ensuite est intervenue la recherche elle-même, impliquant le choix de la fonction par rapport à l'objectif fixé. Après que les résultats se sont affichés, les apprenants ont observé l'item dans son co-texte et contexte, ont repéré les similitudes, les régularités et les différences. Suite à cette étape, ils ont émis des hypothèses, et ont pu éventuellement procéder à une vérification de celles-ci avant de prendre une décision et de valider ou d'invalider le résultat.

Quel que soit le corpus utilisé et malgré la séance de formation préalable, les fonctions du concordancier et du logiciel ont été sous-exploitées. Plusieurs facteurs peuvent expliquer ce sous-emploi. Premièrement, il se peut que la formation n'ait pas été suffisante et qu'elle n'ait pas permis aux apprenants de se sentir à l'aise avec les diverses possibilités que leur offraient les outils. Deuxièmement, il se peut que les tutoriels écrits (COCA) ou oraux (AntConc) n'aient pas été suffisamment utilisés parce qu'en anglais, et donc difficiles à comprendre pour nos apprenants d'un niveau modeste et non spécialistes. Troisièmement, le caractère « artificiel » des cours a certainement contribué à ce sous-emploi. Afin de ne pas fausser les résultats de l'expérimentation, l'enseignante a volontairement adopté une position non participative. Elle a privilégié l'observation des apprenants dans leurs pratiques sur corpus et n'est pas intervenue pour les guider dans leurs recherches.

Les consultations sur le corpus générique ont été plus nombreuses que sur le corpus spécifique alors que la tâche à effectuer était la même pour tous. Pour autant, cet écart important du nombre de consultations sur COCA n'a pas profité à la production, ni en quantité ni en qualité, puisque la production totale en nombre de mots est sensiblement identique à celle du corpus spécifique et que la qualité de la production en termes de richesse lexicale est inférieure à celle du corpus SPE.

Nous avons constaté, en observant les historiques de consultations du corpus SPE, que les étudiants les avaient annotés. Les commentaires sont rédigés à côté de l'item consulté, il y est indiqué dans la plupart des cas le motif de la consultation et la procédure pour parvenir au résultat en utilisant le logiciel *AntConc* (cf. annexe 2). Pourquoi les étudiants exploitant le corpus spécifique ont-ils noté leurs démarches de recherche et pas les autres ? Nous pensons que cela peut provenir de la manière dont est généré l'historique. S'il est automatique comme pour le corpus générique, l'apprenant ne peut rien ajouter ; s'il est manuel, comme c'était le cas pour le corpus spécifique, l'apprenant peut rédiger des annotations à chaque nouvelle consultation. Nous avons également observé que les apprenants ayant utilisé le corpus spécifique ont pris le temps de lire quelques textes en entier et ont étudié plus longuement le sens du message, ce qui pourrait expliquer l'écart du nombre de consultations entre les deux ressources. En conclusion, ce constat nous conduit à réfléchir sur les textes constituant les corpus et l'importance des données des corpus.

Effectivement, bien que les pratiques sur corpus comme ressource de référence ou d'apprentissage soient assez proches quel que soit le type de corpus utilisé, il n'en demeure pas moins que la généralité ou la spécificité des données composant les corpus a un impact sur la consultation, qui elle-même peut avoir un impact sur l'apprentissage. Pour rappel, notre hypothèse est que plus les données des corpus se concentrent sur un genre, plus les apprenants sont intéressés, motivés et apprennent. Cette recherche a permis de vérifier nos suppositions et a fait émerger des différences dans la manière de faire des étudiants lors des consultations, ce qui a eu des conséquences directes sur les diverses productions. Le corpus spécifique a permis également à l'apprenant de prendre connaissance de courriels authentiques. L'intérêt et la curiosité les ont poussés à les lire, la lecture leur a fait découvrir des mots, des formulations, des phrases qui revenaient de manière récurrente dans les courriels. Cette

prise de conscience les a incités à s'en inspirer, à imiter ou à copier des morceaux choisis par eux et pour leur production. Pour Maingueneau (1992 : 117) cela s'apparente à des « modes de gestion de l'intertexte » propre à une communauté ou à un genre. De plus, ces formulations ont l'avantage d'être facilement reconnaissables par des apprenants L2, de correspondre de façon appropriée à un usage spécifique et d'être prêtes à l'emploi.

Conclusion

L'apprentissage sur corpus est une approche centrée sur l'apprenant qui peut conduire à l'autonomisation de l'apprenant et à une collaboration active des étudiants favorisant ainsi le développement et l'échange de stratégies communes. Il se distingue par son adaptabilité à différents contextes, environnements, utilisations, pratiques d'enseignement, contraintes institutionnelles et organisationnelles. Dans le cadre de cette étude, nous avons conçu un dispositif, combinant un corpus de courriels professionnels et un concordancier téléchargeable, que nous avons utilisé en classe. Pour autant, notre intention est d'encourager l'apprentissage sur corpus non seulement en classe mais également à l'extérieur des salles de cours, notamment dans les centres de langues de l'enseignement supérieur, dans le cadre d'un apprentissage en auto-direction, et aussi dans les entreprises pour la formation en langues sur objectifs spécifiques. Pour ce faire, nous pensons qu'une adaptation didactique du dispositif à d'autres contextes et environnements est nécessaire, notamment par la création ou l'aménagement d'un concordancier conçu pour des apprenants et par la conception d'une gamme de corpus spécifiques à différents domaines et genres discursifs à visée professionnelle. Nous sommes consciente de l'enjeu d'une telle démarche « innovante », tant au niveau du concept que de l'approche. Nous sommes convaincue que ce projet, qui vise à apporter aux enseignants de langues de spécialité, tout comme à leurs apprenants, de nouvelles manières d'approcher la langue, en complément d'autres ressources, peut offrir de nouvelles perspectives pour l'enseignement et l'apprentissage d'une L2.

Annexe 1. Scénarios

Séance 2.

C1.Vous avez rendez-vous avec une future cliente Madame Jane Brown, directrice des achats de la société SMB située à Londres, le jeudi 14 février 2013 à 16 heures. Pour des raisons personnelles, vous êtes obligé(e) d'annuler ce rendez-vous et vous lui proposez de le reporter au mercredi 20 février même heure. Vous lui demandez également si votre collaborateur, monsieur Pierre Gris, peut vous accompagner à ce rendez-vous.

Séance 3.

Rédiger à nouveau le courriel C1.

C2. Puis, vous rédigez un courriel à Mme Brown disant :

Que vous joignez un fichier avec les nouveaux prix.

Qu'une remise de 6,5% sera appliquée sur toutes commandes dépassant 10 000 livres sterling et passées avant le 30 avril.

Séance 4.

Rédiger à nouveau les courriels C1 et C2.

C3. Puis, vous rédigez un courriel à Mme Brown disant :

Que vous accusez réception de sa commande du 1^{er} avril 2013 s'élevant à 15 000 euros TTC.

Que la livraison aura lieu dans dix jours et que la facture se trouve en pièce jointe.

Annexe 2. Exemples d'historique de consultations du corpus SPE

Étudiant 26

command : ok but useless in the context

enclosed : ok

more information : ok

facture : no answer

Étudiant 30

apologies pour l'orthographe

report* : pour le mot après

too (traduction de « aussi », si c'était employé)

faithfully (pour voir si employé lorsqu'on connaît ou non le nom de la personne)

appointment* (orthographe + ce qui suit pour voir s'il y a une date)

prices (pour voir si éventuellement un courriel mentionnerait « pièce jointe »)

price (trouver peut-être comment on dit réduction)

Étudiant 42

file : pour trouver pièce jointe

euro : savoir si existe en anglais

pounds : pour trouver TTC

all tax : sans réponse

discount : trouver l'ordre des mots

Étudiant 54

chercher fichier, écrit to join, trouvé « file »

écrit file, trouvé « file is included »

chercher appliqué, écrit appli, trouvé « applications » en déduit « applicated »

chercher livres sterling, écrit sterling, trouvé « pounds sterling »

vérifier « you will receive » trouvé OK

Étudiant 77

would like pour savoir si derrière il faut mettre quelque chose : trouvé TO

4pm : at devant

personals : pas de s

chercher all the products : trouvé phrase complète « the listing of names and price of all the products »

BIBLIOGRAPHIE

- Boulton, A., & Tyne, H. (2014). *Des documents authentiques aux corpus : Démarches pour l'apprentissage des langues*. Paris : Didier.
- Chambers, A. (2010). L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé. *Revue Française de Linguistique Appliquée (Apprendre à écrire : l'apport des nouvelles technologies)*, 15 : 9-20.
- Conseil de l'Europe. (2000). *Un cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer*. Strasbourg : Conseil de l'Europe. Récupéré de http://www.coe.int/t/dg4/linguistic/Source/Framework_FR.pdf
- Duda, R., & Tyne, H. (2010). Authenticity and autonomy in language learning. *Bulletin suisse de linguistique appliquée*, 92 : 86-106.
- Holec, H. (1990). Des documents authentiques, pour quoi faire ? *Mélanges Pédagogiques*, 65-74.
- Johns, T. (1990). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, 10: 14-34. Version révisée Dans T. Johns & P. King (Dirs.), 1991. *Classroom Concordancing, English Language Research Journal*, 4: 27-45.
- Johns, T. (1991). Should you be persuade: two examples of data-driven learning. Dans T. Johns & P. King (Dirs.), *Classroom Concordancing. English Language Research Journal*, 4: 1-16.
- Mainueneau, D. (1992). Le 'tour' ethnolinguistique et l'analyse du discours. *Langages*, 105 : 114-125.

Liste de corpus, outils et ressources en ligne

AntConc. Concordancier : <http://www.antlab.sci.waseda.ac.jp/software.html>

Compleat Lexical Tutor (LexTutor): <http://www.lextutor.ca>.

Corpus of Contemporary American English (CoCA). Corpus d'anglais américain contemporain : <http://corpus.byu.edu/coca/>

Autres corpus BYU : <http://corpus.byu.edu/>

LexTutor. Site de Tom Cob : <http://www.lextutor.ca/vp/eng>

WordReference. Online Language Dictionaries: <http://www.wordreference.com>

W3C Corpus: <http://research.microsoft.com/en-us/um/people/nickcr/w3c-summary.html>
et http://universal.elra.info/product_info.php?cPath=42_43&products_id=2153