Studii de lingvistică 7, 2017, 13 - 26

À la croisée des corpus et de la phraséologie : une proposition d'outil informatique

At the crossroads of corpora and phraseology:
a proposal for a computer tool

Jean-Pierre Colson[1]

Abstract: Corpora are currently enjoying ever-increasing success, and are no longer solely the domain of corpus linguists. They provide us with a picture of language as described by John Sinclair: on average about half of any text consists of phraseological units in the broadest sense. It would therefore be unrealistic to teach foreign languages without taking this aspect into account.
However, the exploitation of corpora for the didactics of phraseology is fraught with obstacles: the frequency of phraseological units is low, and huge corpora are necessary in order to guarantee sufficient numbers of relevant examples; simple concordances of phraseological units are sometimes contradictory or hard to interpret; finally, the available tools for manipulating such vast collections of texts are not always user-friendly, and require that learners should have a flair for computing and/or statistics.
In this paper we propose new techniques to build on a notion introduced by Michael Lewis in 1993: phraseological awareness raising, based on recurrent comparison with corpus examples. We also present an experimental tool, the *IdiomSearch* application (J. P. Colson 2016), that provides learners with immediate feedback on the presence of phraseological units in an input text, as well as a ranking on the basis of frequency and fixedness.

Key words: phraseology, foreign language teaching, corpora, statistical semantics, collocations.

1. Introduction

Notre propos n'est pas de retracer ici l'histoire de la linguistique de corpus ni d'analyser en détails les différents courants qui la composent. Rappelons simplement que les corpus linguistiques ont été critiqués par d'éminents linguistes, notamment par Chomsky

[1] Université catholique de Louvain, Louvain School of Translation and Interpreting (LSTI) ; jean-pierre.colson@uclouvain.be.

(1957) et Abercrombie (1963). Leur définition même reste quelque peu controversée, car même les spécialistes de la linguistique de corpus (McEnery *et al.* 2006) nuancent leur propos, et reprennent des critères assez diversifiés :

> There are many ways to define a corpus […], but there is an increasing consensus that a corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety. (McEnery *et al.* 2006: 5)
> 'Il y a diverses manières de définir un corpus […], mais un consensus de plus en plus large se dégage pour le définir comme une collection de textes (1) *lisibles par la machine* et (2) *authentiques* (en incluant les transcriptions de données orales), qui est (3) *échantillonnée* pour être (4) *représentative* d'une langue ou d'une variété linguistique particulières.' (notre traduction)

Il est frappant de constater que les corpus linguistiques, souvent au sens le plus général de vastes collections de textes, ne sont plus l'apanage des linguistes de corpus : on les voit apparaître dans nombre de disciplines linguistiques qui se démarquaient autrefois nettement de la linguistique de corpus. Peut-être faut-il y voir l'influence plus générale des multimédias et des ressources informatiques sur les sciences humaines.

La pragmatique, par exemple, longtemps adversaire des corpus, s'est engagée dans la double voie des études basées sur les corpus ("corpus-based pragmatics") et dérivées des corpus "corpus-driven pragmatics" (Huang 2012 : 18). Même la linguistique cognitive, dont les liens historiques avec la grammaire générative sont bien établis, s'est récemment tournée vers les corpus, comme en témoignent par exemple les travaux de Geeraerts (2002), Storms *et al.* (2015), Peirsman *et al.* (2015).

Il suffit d'ailleurs d'utiliser un moteur de recherche sur la Toile pour constater la fréquence élevée du terme *corpus* adossé à bien des domaines de la linguistique[2].

La phraséologie, l'étude des expressions figées au sens large, depuis les simples routines et formules communicatives jusqu'aux expressions idiomatiques et aux proverbes (Burger 1998), constitue également aujourd'hui un thème récurrent dans les recherches linguistiques, et les points de contact avec les corpus sont nombreux. Un des pères fondateurs de la linguistique de corpus britannique, John Sinclair, a souligné l'importance du *principe idiomatique* (« idiom principle », Sinclair 1991), en vertu duquel environ la moitié de tous les textes est composée de structures idiomatiques au sens large. Il

[2] Citons par exemple, en anglais : *corpus-based semantics, corpus-based translation studies, corpus-based lexicology, corpus-based language teaching.*

s'ensuit que les texts sont largement *délexicalisés*[3] :

> Most normal text is made up of the occurrence of frequent words, and the frequent senses of less frequent words. Hence, normal text is largely delexicalized, and appears to be found by exercise of the idiom principle, with occasional switching to the open-choice principle. (Sinclair 1991: 113)
> 'L'essentiel d'un texte normal est constitué des occurrences de mots fréquents, et des sens courants des mots moins fréquents. En conséquence, un texte normal est largement délexicalisé, et se révèle issu du principe idiomatique, avec des recours occasionnels au principe du libre choix.' (notre traduction)

Notons que, dès les années 1990, Sinclair insiste sur l'importance des structures délexicalisées dans les textes, ce qui soulève déjà la question de la proportion des *phrasèmes* ou *unités phraséologique*[4] dans le langage en général. La question est d'ailleurs controversée : selon Jackendoff (1995), le nombre d'expressions figées est comparable à celui des mots simples dans le dictionnaire, alors que Mel'čuk (1995) affirme que les expressions figées sont bien plus nombreuses que les mots simples. Tant la linguistique de corpus anglo-saxonne que la phraséologie ont donc abouti à relativiser l'importance à la fois du lexique et de la grammaire dans la langue, car le figement intervient à tous les niveaux.

Ce débat théorique n'est pas sans incidence sur la didactique des langues étrangères. Si, comme l'affirment les auteurs précités, la phraséologie représente au moins la moitié des structures de la langue, on voit mal comment l'apprentissage des langues étrangères pourrait être efficace sans tenir compte de cette réalité. La phraséologie (Burger 1998) insiste également sur les aspects culturels qui interviennent dans le figement de certaines structures. Il s'avère dès lors utile de diversifier les langues prises en compte pour étudier cette problématique.

Nous illustrerons notre propos par deux exemples en chinois officiel (mandarin). Notons tout d'abord que la linguistique traditionnelle a longtemps privilégié les exemples issus de langues européennes, principalement l'anglais, alors qu'un problème de taille se pose pour bien des langues, dont le chinois : il n'y a pas d'espace entre les mots, ce qui impose une première tâche de segmentation à la linguistique de corpus et à la linguistique informatique ; dans bien des cas, cette segmentation pourra se faire de plusieurs manières, ce qui relativise déjà la notion de *tokens* (chaînes de caractères séparées par un blanc) pour les langues européennes.

---

[3] La *délexicalisation* désigne ici le fait que les mots, au sein des unités phraséologiques, perdent largement leur sens individuel, au profit du sens global, souvent non-compositionnel, de l'unité phraséologique.

[4] Nous préférons l'appellation générale *unité phraséologique*, car *phrasème* est parfois utilisé (Burger 1998) au sens restreint des expressions idiomatiques.

Les routines communicatives dans les langues géographiquement éloignées de l'Europe illustrent bien l'enchevêtrement du lexique, de la grammaire, de la phraséologie et de la culture. Nous prendrons l'exemple de deux formules courantes en mandarin pour dire « bonjour, salut » :

(1) 你好(nǐhǎo, littéralement « toi bon »)

(2) 你吃了吗？((nǐchī le ma, littéralement « as-tu mangé ? »)

L'exemple (1) contient certes une structure de type syntaxique (bien que le verbe *être* ne soit pas exprimé), mais le choix de l'adjectif *bon* pour saluer quelqu'un est arbitraire, et en lien direct avec la culture. Il en va de même sous (2), qui comporte en outre deux particules : la particule d'achèvement 了(*le*) et la particule interrogative 吗 (*ma*), qui rappellent aussi l'interaction entre grammaire et phraséologie.

2. Les limites des corpus pour la phraséologie

La complexité, la diversité et la richesse culturelle des unités phraséologiques sont telles que l'apprenant en langue étrangère a souvent besoin d'exemples en contexte pour mieux les appréhender. Il se tourne alors tout naturellement vers les corpus, or c'est précisément ici que le bât blesse.

En effet, l'apport des corpus à cette problématique n'est pas toujours simple, car nombre de structures figées présentent une fréquence assez basse sur les corpus (Colson 2007, Moon 1998), et le langage figuré ne se distingue pas toujours aisément du sens propre.

Il est nécessaire de prendre un exemple concret pour mesurer la difficulté de cette problématique. *Le Grand Robert de la Langue française*[5] mentionne sous l'entrée *lac* l'expression *tomber dans le lac*, locution figurée et familière, qui signifie *échouer, n'avoir pas de suite*, par exemple pour un projet qui n'aboutit pas.

Si nous consultons un très vaste corpus de langue française, le frTenTen 2012[6], un corpus Web de 12 milliards de mots (*tokens*), nous constatons que la fréquence globale de *tomber dans le lac* est très moyenne (473 occurrences, soit 0.04 par million de mots), mais surtout que presque toutes les occurrences correspondent au sens littéral, ainsi que le montrent les phrases reprises sous (3) :

(3) # Corpus : French Web 2012 (frTenTen12)
Hits : 473
doc#9252658 les petits bouts, du coup désolée mon info < tombe dans le lac > ! Plouf ! J'ai trouvé un autre truc

[5] Edition payante en ligne, 2016, www.lerobert.com, consultée le 21/09/2016.
[6] www.sketchengine.co.uk, consulté le 21/09/2016.

doc#9379995 'eau glacée, il priait pour que personne < tombe dans le lac > avant les deux prochaines heures. Tout
doc#9408314 de panique il fuit, trébuche et finit par < tomber dans le lac >. À son réveil tout cela ne semble qu'un
doc#9416162 femme qu'il veut aider à grimper le sentier, < tombe dans le lac > qui est au-dessous ; plusieurs anneaux
doc#9434889 glissa sur une touffe d'herbe et manqua de < tomber dans le lac >. Manqua ? Quelqu'un l'avait retenu par
doc#9475080 Pierre rentre, trempé comme s'il était < tombé dans le lac >, et il déclare : « Maintenant il est sur

Il a fallu en réalité lire tous les exemples jusqu'au 240e (sur 473) pour enfin trouver une occurrence idiomatique de *tomber dans le lac*, reprise en tête de liste sous (3). Le corpus, malgré sa taille gigantesque, se révèle donc peu pratique pour distinguer le sens littéral de l'idiomaticité et pour rechercher des contextes nombreux d'unités phraséologiques.

Nous n'avons pas choisi cet exemple au hasard : les voies de la phraséologie sont très sinueuses et les emprunter présente bien des risques pour les apprenants, voire pour les locuteurs natifs. Ainsi, *tomber dans le lac* au sens figuré est bien entendu synonyme de *tomber à l'eau*. Mais l'expression *tomber à l'eau*, se demandera peut-être l'apprenant, tolère-t-elle le sens littéral ? Spontanément, un locuteur francophone utilisera plutôt *tomber dans l'eau* pour le sens littéral, mais que nous apprend le corpus ? Notons que, selon le *Grand Robert*, tant le sens littéral que le sens figuré sont possibles, sans autre précision sur leur fréquences respectives. Le même corpus (frTenTen 2012) nous révèle cette fois que l'expression *tomber à l'eau* est beaucoup plus courante au sens figuré, mais de nombreux exemples sont difficiles à interpréter à partir de la simple concordance, dont un court extrait est proposé sous (4) :

(4) # Corpus : French Web 2012 (frTenTen12)
Hits : 13768
doc#55146 été annoncée le weekend dernier est depuis < tombée à l'eau >. […] Les deux hommes, pourtant mariés
doc#57000 match se profiler… Evidemment ma théorie < tombe à l'eau > si l'attaque des Bears trouvent des brèches
doc#62492 mais il ne voulait pas abandonné sinon tout < tomberait à l'eau >, il le savait. Dans un premier temps sa
doc#64393 père par Matt et un point commun tout va < tomber à l' eau>. Cette année la liste des prix a
doc#65659 qu'il essayait d'être dur, mais cet effet < tombait à l'eau > avec ses joues, et l'embarra dans son ton
doc#77695 Myriam a déclaré en substance : Il ne faut pas < tomber à l'eau > ! Elle est froide, sans combinaison, cela
doc#81169 avait bien cru à ce moment qu'elle allait < tomber à l'eau >. En pénétrant pour la première fois dans

Une autre difficulté concerne la convivialité des outils. Une concordance brute est plutôt rébarbative, et l'utilisateur du Sketch Engine[7], l'outil le plus performant en matière de vastes corpus linguistiques, doit faire preuve d'un esprit porté sur l'informatique s'il ambitionne d'extraire du corpus des informations phraséologiques plus fines. L'exemple (5) est une requête CQL[8] (*Corpus Query Language*) que l'utilisateur doit mettre au point *via* le Sketch Engine, s'il souhaite obtenir des exemples de *laisser la proie pour l'ombre* à toutes les formes conjuguées, en tolérant jusqu'à deux mots entre *laisser* et *la* et en admettant un maximum de trois mots, quels qu'ils soient, entre *proie* et *ombre*, ce qui permet de rechercher les variantes de l'expression. S'il fallait une expression régulière plus complexe à partir d'un autre corpus, l'expression sous (6) pourrait être utilisée. Quelques exemples de la concordance obtenue par la requête CQL sur le corpus frTenTen 2012 du Sketch Engine (12 milliards de mots) sont présentés sous (7) :

(5) "laiss.*"[]{0,2 }"la"[]{0,3}"ombre"

(6) /laiss\S+\s(\S+\s)[0,2]la(\S+\s)[1,3]proie(\S+\s)[1,4]ombre\b/

(7) # Corpus : French Web 2012 (frTenTen12)
Hits : 111
#Query « laiss.* »[]{0,2 } »la »[]{0,3} »ombre » 111
doc#16194324 face des trous sur un méchant temps. On ne < laisse pas la proie pour l'ombre >, et on est pas assez con pour passer des
doc#16508185 jusqu'à ce qu'il s'effondre totalement pour < laisser place à la fameuse ombre >, qui le terrassait pour prendre contrôle
doc#16741338 puissants et l'univers si charmant. Je vous < laisse la part d'ombre > qui est en moi. De là où je suis, je devine
doc#17384312 regrettent leur vote et qui reconnaissent avoir < laissé la proie pour l'ombre >. Seul ce blog permet le débat car le maire

D'autres corpus Web accessibles en ligne permettent aux apprenants, dans une certaine mesure, de rassembler des informations phraséologiques pertinentes autour d'un mot ou d'un lemme. Citons, outre le Sketch Engine déjà mentionné, les corpus

[7] L'utilisation directe de corpus et de concordanciers par les apprenants en langue étrangère est contestable, même si cette pratique est largement répandue en linguistique de corpus. Dans l'enseignement universitaire, le Sketch Engine est en tout cas utilisé par de très nombreuses institutions, qui recommandent son utilisation aux étudiants en langues étrangères ou en traduction-interprétation.

[8] Le langage CQL utilisé par le Sketch Engine (www.sketchengine.co.uk) est une version simplifiée des « expressions régulières », un algorithme bien connu en informatique et utilisé notamment dans le langage Perl.

Aranea[9], les corpus Web de l'Université de Leeds[10], ou encore le portal Wortschatz de l'Université de Leipzig[11]. Tous offrent non seulement des contextes, mais également des scores statistiques pour les collocations, un terme bien connu en linguistique de corpus et en linguistique informatique ; nous renvoyons ici à la définition de Burger (1998) : une catégorie assez générale de combinaisons faiblement idiomatiques. L'exemple le plus souvent cité est celui des combinaisons de deux mots, et en particulier celles d'un substantif et d'un adjectif. Ces différents outils accessibles sur la Toile offrent à l'utilisateur la possibilité de sélectionner les collocations les plus significatives, sur la base de scores statistiques.

Voici précisément une nouvelle limite majeure du recours aux corpus en matière de phraséologie. L'information relative aux collocations est en effet cruciale dans la découverte des combinaisons de mots les plus utiles et de la phraséologie. Or, l'information livrée par les outils précités souffre d'une grande confusion, car le même outil livre pour un même corpus des résultats divergents, selon le score statistique utilisé.

Prenons l'exemple concret d'un apprenant avancé de langue anglaise, qui souhaite connaître les 10 collocations les plus courantes ayant pour base *criticism*. Sous (8) sont reprises les collocations les plus significatives pour un mot à gauche de *criticism*, sur la base du score statistique *logDice* (corpus enTenTen13 du Sketch Engine, 13 milliards de mots), tandis que sous (9) figurent les mêmes résultats avec le *t-score* et sous (10) avec le score *MI (Mutual Information)*. Tous ces scores sont notamment décrits dans Evert (2004) et Gries (2013).

(8) # Collocations
Corpus : English Web 2013 (enTenTen13)
Query : [lemma="criticism"] 431087

Cooccurrence count	Candidate count	logDice
constructive	12256	9.469
literary	5259	7.927
harsh	3596	7.321
textual	1661	6.825
constructive	898	6.070
widespread	1473	6.001
sharp	1786	5.884
drew	1411	5.859
scathing	747	5.767
deflect	721	5.695

(9) # Collocations
Corpus : English Web 2013 (enTenTen13)

[9] http://sketch.juls.savba.sk/aranea_about/.
[10] http://corpus.leeds.ac.uk/internet.html.
[11] http://corpora.informatik.uni-leipzig.de/fr.

# Query : [lemma="criticism"] 431087		
Cooccurrence count	Candidate count	T-score
constructive	12256	110.684
of	28436	10.512
the	33474	82.943
literary	5259	72.446
his	6468	69.481
any	5566	68.430
some	5029	62.375
harsh	3596	59.870
public	3512	57.300
much	3232	50.833

(10)

# Collocations		
# Corpus: English Web 2013 (enTenTen13)		
# Query: [lemma="criticism"] 431087		
Cooccurrence count	Candidate count	MI
unspokening	7	15.686
unspokened	6	15.686
constructive	4	15.364
Cruz-Moore's	4	14.879
Ernesti's	3	14.686
Brahmaguptas	3	14.686
Seervai's	5	14.686
Dzulkifli's	3	14.464
post-Hegelian	27	14.354
unspokens	11	14.338

Les limites de la linguistique de corpus pour une éventuelle didactique des collocations sont clairement illustrées par de tels résultats. Ainsi, seules trois des collocations les plus significatives selon le score *logDice* le sont également selon le *t-score*, alors que le score *MI* livre comme collocations les plus significatives des combinaisons aberrantes, notamment liées à des noms propres. Ainsi que l'a fait remarquer S. Gries (2013), plus de 50 années de recherches sur l'extraction automatique des collocations à partir des corpus n'ont vu aucun progrès sensible. De plus, la plupart de ces scores sont limités aux combinaisons de deux mots, les bigrammes, et ne peuvent être étendus aux structures plus longues. Enfin, les résultats très divergents obtenus selon le score utilisé incitent certains linguistes à travailler plutôt avec la fréquence brute des cooccurrences, qui livre toutefois également nombre de résultats non pertinents.

Vouloir rédiger un dictionnaire des collocations sur des bases objectives et quantifiables relève donc toujours de l'utopie. Aucun score statistique ne livre en effet des résultats fiables, qui puissent être reproduits sur d'autres corpus ou étayés par un autre score statistique, ou encore confirmés par des locuteurs natifs. Notons au passage que même l'avis des locuteurs natifs est hautement sujet à caution en la

matière, car ils n'ont pas toujours une perception claire des phrasèmes, ni des limites entre constructions libres et constructions figées.

3. Prise de conscience phraséologique et expérience sur corpus

La linguistique nous a habitués à des débats passionnés entre les écoles rivales, par exemple le courant générativiste (Chomsky 1957) et le courant de la linguistique de corpus (Sinclair 1991). Il est toutefois réconfortant de noter, depuis les années 1990, plusieurs points de convergence étonnants en matière d'unités polylexicales.

Ainsi, dès les années 1990, Michael Lewis (1993, 1997) insiste sur une *approche lexicale* dans laquelle les apprenants doivent développer une *prise de conscience* (*awareness*) des éléments préfabriqués de la langue, ou *chunks*. D'autres chercheurs de linguistique appliquée (Ellis 2005, 2008) ont également insisté sur l'importance de la manipulation de l'*input* linguistique par les apprenants, ce qui n'est pas incompatible avec une telle approche. La grammaire des constructions (Goldberg 1995, 2003, 2006 ; Croft 2001), elle aussi, souligne l'intérêt des expérimentations syntagmatiques par les apprenants. Enfin, les applications de la linguistique cognitive à l'apprentissage des langues étrangères (De Knop 2015, Roche & Suñer Muñoz 2016) proposent également des exercices de reconnaissance et d'utilisation créative des métaphores à partir de textes, dont beaucoup correspondent à des unités phraséologiques.

Malgré les nombreuses nuances entre ces diverses approches théoriques en linguistique appliquée, un point de convergence se dégage sur la nécessité de proposer aux apprenants diverses méthodologies qui éveillent une prise de conscience syntagmatique et phraséologique.

L'exemple le plus frappant est l'exercice du *chunking* (morcellement, ou quête des éléments préfabriqués), déjà proposé par Michael Lewis en 1993 : les apprenants sont invités à retrouver dans un texte toutes les combinaisons syntagmatiques fréquentes, en particulier les associations du lexique et de la grammaire, selon les principes déjà proposés par M.A.K. Halliday (1973)[12].

Il est louable de recommander des exercices basés sur la reconnaissance des *chunks* (éléments récurrents ou préfabriqués), fort proches des unités phraséologiques au sens large, mais nous avons vu au paragraphe 2 que les locuteurs natifs eux-mêmes se mettent rarement d'accord sur le nombre précis ou la taille des unités phraséologiques, et que les corpus présentent également de sérieuses limites.

[12] M.A.K. Halliday a introduit la notion de lexico-grammaire, qui souligne les nombreuses interactions entre le lexique et la grammaire, une idée également présente chez Sinclair (1991); à ne pas confondre avec la théorie française du lexique-grammaire (Gross 1968), davantage inspirée par la linguistique américaine (Harris 1964).

Nous proposons dès lors une méthodologie alternative, basée sur un projet de recherche en phraséologie informatique[13]. Dans un premier temps, nous avons assemblé de manière semi-automatisée une base de données phraséologiques (Colson 2016), accessibles en ligne via une interface graphique[14]. Le fonctionnement pédagogique de l'outil est illustré ci-dessous par quelques exemples.

Il peut être utile de montrer aux apprenants (étudiants de niveau intermédiaire ou avancé en langue étrangère ou en traduction-interprétation) que les structures syntagmatiques récurrentes, souvent constituées de phrasèmes au sens large (routines et formules, collocations, clichés, expressions idiomatiques), sont très présentes dans les articles de presse généraux ou semi-spécialisés qui abordent un domaine précis de l'actualité ou de la technique.

L'exemple (11) est extrait d'un texte du quotidien britannique *The Guardian*[15], relatif à l'évolution des prix du pétrole :

(11) At the heart of the current turmoil is a decision by Saudi Arabia and other leading voices in the Opec oil cartel to get drawn into a turf war with the new generation of US shale producers.
'Au cœur de l'agitation actuelle, nous trouvons une décision de l'Arabie Saoudite et d'autres voix dominantes au sein du cartel pétrolier de l'OPEP de livrer une guerre de la concurrence sans merci à la nouvelle génération de producteurs de pétrole de schiste américains.' (notre traduction)

Bien avisé sera le locuteur natif capable de démêler l'écheveau dans ce court extrait : la frontière entre le lexique, la grammaire et la phraséologie. Ainsi, *at the heart of* est une métaphore commune à bien des langues d'Europe (au cœur de), et a donné lieu à nombre d'unités phraséologiques, mais qu'en est-il de *current turmoil* (littéralement, 'la présente agitation') ? S'agit-il vraiment d'une collocation / d'un « chunk » / d'une structure fréquente et récurrente ? Gageons que tant le locuteur natif que l'apprenant pourront nourrir certains doutes à ce sujet. L'outil *IdiomSearch* apporte en tout cas les réponses suivantes :

(12) structures semi-figées : *at the heart of, and other leading, get drawn into, the new generation of*
structures figées : *current turmoil*
structures très figées : *turf war*

Notons que *turf war* est bien une unité hautement idiomatique et intraduisible en français (littéralement, 'guerre du gazon'), car il s'agit d'une lutte d'influence ou entre clans rivaux, proche de la

[13] Projet IdiomSearch, Université catholique de Louvain & J.-P. Colson 2016.
[14] http://idiomsearch.lsti.ucl.ac.be.
[15] http://www.theguardian.co.uk, consulté en ligne le 07/01/2016.

métaphore souvent utilisée en anglais et en français du jardin, de la pelouse, du gazon sur laquelle il ne faut pas empiéter (par exemple dans la locution *empiéter sur les plates-bandes de qqn*).

Notons toutefois que les structures extraites automatiquement par l'outil *IdiomSearch* proviennent de la présence attestée sur un corpus de référence de 200 millions de mots (*tokens*), ce qui n'offre pas une fiabilité absolue. Un locuteur natif pourra par exemple aussi considérer comme récurrentes ou phraséologiques les combinaisons suivantes : *a decision by, oil cartel, shale producers*. Au bout du compte, seuls quelques mots grammaticaux présents dans cette phrase ne sont pas liés à la phraséologie.

Le discours journalistique, et particulièrement les éditoriaux, sont très riches en phrasèmes au sens large (collocations, routines et formules rédactionnelles, clichés, expressions idiomatiques), et permettent dès lors aux apprenants de découvrir toute la diversité du phénomène. L'exemple (13) provient d'un éditorial sur le Brexit[16] :

(13) Why, when it comes down to it, do more people in this country not start from Delia Smith's wonderfully common sense assessment of the EU last weekend as "in essence … a group of democratic countries attempting to work alongside each other" ?
Pourquoi, au bout du compte, une majorité de citoyens de ce pays ne se réfèrent-ils pas davantage à cette magnifique vision pragmatique de l'UE livrée le week-end dernier par Delia Smith : « fondamentalement, … il s'agit d'un groupe de pays démocratiques qui tentent de collaborer en bonne entente » ?' (notre traduction)

L'outil *IdiomSearch* livre pour cette phrase les résultats suivants :

(14) structures semi-figées : *last weekend* ;
structures figées : *people in this country, assessment of the, in essence, a group of, work alongside each other* ;
structures très figées : *when it comes down to it, common sense, attempting to*.

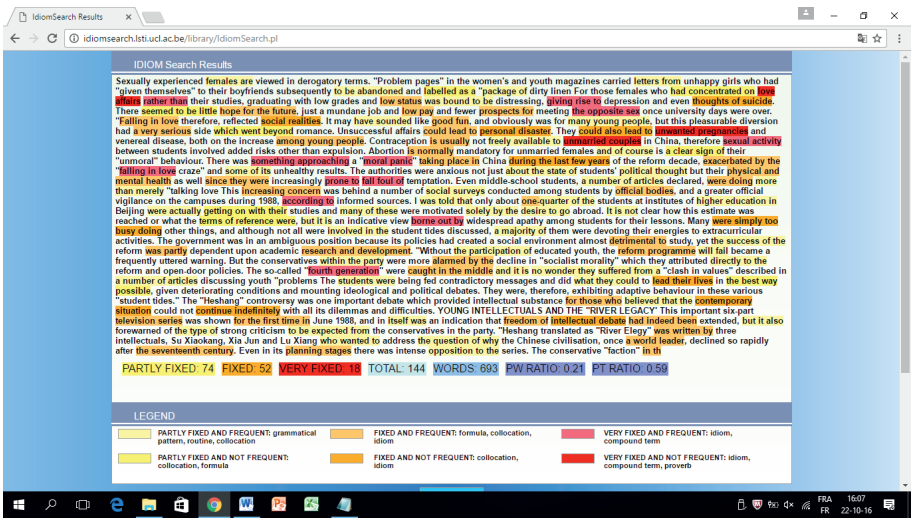
Notons à nouveau le nombre élevé de structures récurrentes à partir d'une simple phrase telle que (13), ce qui confirme l'omniprésence de la phraséologie au sens large. A force de manipuler de tels exemples concrets, glanés au hasard des lectures, les apprenants seront amenés progressivement à prendre conscience de l'existence et de l'usage de ces nombreux éléments préfabriqués du langage.

Les apprenants en langue étrangère de niveau intermédiaire ou avancé sont supposés améliorer leurs connaissances lexicales et phraséologiques par le biais de lectures variées. Ils éprouvent toutefois

[16] http://www.theguardian.co.uk, consulté en ligne le 03/06/2016.

des difficultés à identifier, au fil de leurs lectures, les constructions récurrentes et figées, qui pourraient précisément leur permettre de progresser dans l'acquisition des subtilités d'une langue étrangère. De ce point de vue, une double information est en réalité nécessaire : la fréquence relative des constructions, d'une part, et leur degré de figement ou d'idiomaticité, d'autre part.

Afin d'améliorer la convivialité de l'usage des corpus, l'outil IdiomSearch recourt à une palette de couleurs qui s'échelonnent du jaune pâle au rouge rubis, selon la fréquence et l'idiomaticité. La vision d'ensemble du texte est également importante car elle permet à l'utilisateur d'évaluer la présence globale de phrasèmes dans un texte, selon le nombre de parties colorées qui sont mises en évidence, ainsi que le montre en nuances de gris la capture d'écran ci-dessous :



4. Conclusion

Il y a loin de la coupe aux lèvres, comme dit l'expression, et vouloir enseigner la phraséologie ou améliorer ses compétences en la matière par le biais des corpus représente un défi pratique et technique : la clé réside dans de très vastes collections de textes, peu accessibles par les chercheurs et *a fortiori* par les apprenants ; la manipulation des outils existants est parfois laborieuse et l'informatique se mêle à l'exercice.

Une étape intermédiaire est sans doute nécessaire : créer des interfaces plus conviviales entre les ressources électroniques brutes et leur application didactique. Nous avons proposé une expérience en ce

sens, qui résulte de l'application d'algorithmes expérimentaux, encore à perfectionner. La gageure de la recherche en phraséologie consiste précisément à travailler à la fois sur la complexité des hypothèses théoriques et sur les données brutes qui permettent de les vérifier dans les corpus. Même si bien des aspects de la théorie linguistique qui sous-tend la phraséologie doivent encore être élucidés, une approche pratique, fondée sur des outils plus conviviaux, permet aux apprenants non seulement de travailler à partir de textes authentiques, mais aussi de découvrir de manière plus objective la présence des éléments préfabriqués du langage.

Références bibliographiques

Abercrombie, D. (1963), *Studies in phonetics and linguistics*, Oxford University Press, London.

Burger, H. (1998), *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Erich Schmidt Verlag, Berlin.

Chomsky, N. (1957), *Syntactic Structures*, Mouton, La Haye / Paris.

Colson, J.-P. (2007), "The World Wide Web as a corpus for set phrases", in Burger, H., Dobrovol'skij, D., Kühn, P. & Norrick, N. R. (éds), *Phraseologie / Phraseology. Ein internationales Handbuch der zeitgenössischen Forschung / An International Handbook of Contemporary Research*, volume 2, Walter de Gruyter, Berlin / New York, p. 1071-1077.

Colson, J.-P. (2016), "Set phrases around globalization: an experiment in corpus-based computational phraseology", in Alonso Almeida, F., Ortega Barrera, I., Quintana Toledo, E. & Sanchez Cuervo, M. E. (éds), *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics*, Cambridge Scholars Publishing, Newcastle, p.141-152.

Croft, W. (2001), *Radical construction grammar*, Oxford University Press, Oxford.

De Knop, S. (2015), "Conceptual tools for the description and the acquisition of the German posture verb *sitzen*", *Corpus Linguistics and Linguistic Theory*, 11, p. 127-160.

Ellis, N. (2005), "At the interface: dynamic interactions of explicit and implicit language knowledge", *Studies in Second Language Acquisition*, 27, p. 305-52.

Ellis, N., Simpson-Vlach, R., Maynard, C. (2008), "Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL", *TESOL Quarterly*, 42, p. 375-96.

Evert, S. (2004), *The statistics of word cooccurrences – word pairs and collocations*, Ph.D. thesis, University of Stuttgart, Stuttgart.

Geeraerts, D. (2002), "The interaction of metaphor and metonymy in composite expressions", in Dirven, R., Pörings, R. (éds), *Metaphor and metonymy in comparison and contrast*, Mouton de Gruyter, Berlin, p. 435-465.

Goldberg, A. (1995), *Constructions: A Construction Grammar Approach to Argument Structure*, University of Chicago Press, Chicago.

Goldberg, A. (2003), "Constructions. A New Theoretical Approach to Language", *Trends in Cognitive Sciences*, 10, p. 219-224.

Goldberg, A. (2006), *Constructions at Work: The Nature of Generalization in Language*, Oxford University Press, Oxford.

Gries, S. (2013), "50-something years of work on collocations. What is or should be next...", *International Journal of Corpus Linguistics*, 18, p. 137-165.
Gross, M. (1968), *Grammaire transformationnelle du français. Vol. 1, Syntaxe du verbe*, Larousse, Paris.
Halliday, M. A. K. (1973), *Explorations in the Functions of Language*, Edward Arnold, London.
Harris, Z. (1964), "Transformations in Linguistic Structure", *Proceedings of the American Philosophical Society*, 10, p. 418-122.
Huang, Y. (2012), *The Oxford Dictionary of Pragmatics*, Oxford University Press, Oxford.
Jackendoff, R. (1995), "The boundaries of the lexicon", in Everaert, M., van der Linden, E.-J., Schenk, A. & Schroeder, R. (éds), *Idioms: Structural and psychological perspectives*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, p. 133-165.
Lewis, M. (1993), *The Lexical Approach*, Language Teaching Publications, Hove.
Lewis, M. (éd.) (1997), *Implementing the Lexical Approach*, Language Teaching Publications, Hove.
McEnery, T., Xiao, R., Tone, Y. (2006), *Corpus-based language studies. An advanced resource book*, Routledge, London / New York.
Mel'čuk, I. (1995), "Phrasemes in language and phraseology in linguistics", in Everaert, M., van der Linden, E.-J., Schenk, A. & Schroeder, R. (éds), *Idioms: Structural and psychological perspectives*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, p. 167-232.
Moon, R. (1998), *Fixed Expressions and Idioms in English*, Clarendon Press, Oxford.
Peirsman, Y., Geeraerts, D., Speelman, D. (2015), "The corpus-based identification of cross-lectal synonyms in pluricentric languages", *International Journal of Corpus Linguistics*, 20, p. 55-81.
Roche, J., Suñer Muñoz, F. (2016), *Sprachenlernen und Kognition. Grundlagen einer kognitiven Sprachdidaktik*, Gunter Narr, Tübingen.
Sinclair, J. (1991), *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.
Storms, S., Speelman, D., Geeraerts, D., Storms, G. (2015), "Within-concept similarities in a taxonomy: a corpus linguistic approach", *Language and Cognition*, 7, p. 194-218.