Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés

Olivier Kraif

KRAIF (Olivier), « Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés », Cahiers de lexicologie, n° 108, 2016 - 1, Phraséologie et linguistique appliquée, p. 91-106
DOI : 10.15122/isbn.978-2-406-06281-3.p.0091

LE LEXICOSCOPE : UN OUTIL D'EXTRACTION DES SÉQUENCES PHRASÉOLOGIQUES BASÉ SUR DES CORPUS ARBORÉ

Kraif, Olivier

Univ. Grenoble Alpes, LIDILEM, F-38040, Grenoble

olivier.kraif@univ-grenoble-alpes.fr

Résumé

Cet article présente les fonctionnalités du Lexicoscope, une architecture dédiée à l'exploration de corpus arborés. Après l'examen de quelques outils similaires, nous montrons comment la caractérisation d'expressions complexes (correspondant à des arbres syntaxiques) dans la recherche de concordances et l'extraction de tableaux de cooccurrents, peut se révéler très utile pour l'étude des collocations et de la combinatoire lexico-syntaxique en général. Nous décrivons également un scénario d'utilisation permettant à l'utilisateur d'explorer finement les contextes des expressions cibles sans devoir s'initier au formalisme du langage de requête sous-jacent, grâce à une modalité originale d'interrogation basée sur l'exemple.

Mots clés : interrogation de corpus arborés, concordances, cooccurrents syntaxiques

Abstract

This article discusses the features of the Lexicoscope, an architecture dedicated to treebank exploration. After reviewing some similar tools, we show how the use of complex expressions (corresponding to syntactic trees) in concordancing as well as extracting word sketches, can be very useful for the study of collocations and phraseology. We also provide a scenario for users, allowing them to explore in detail the contexts of target expressions, without having to learn the query language of the underlying formalism, through an original feature of example-based syntactic querying.

Keywords : treebank querying, concordances, syntactic cooccurrence

1. Introduction

Depuis la campagne CONLL 2007 dédiée à l'évaluation des analyseurs syntaxiques en dépendances (Nivre et al., 2007), on sait que l'état de l'art permet

d'obtenir des résultats satisfaisants dans ce domaine, avec des scores de rattachement étiqueté (*labelled attachment score*) supérieurs à 85 % pour des langues comme l'anglais, l'italien ou le catalan. On constate par ailleurs que de plus en plus de corpus annotés syntaxiquement sont disponibles au téléchargement et/ou interrogeables via des interfaces web.

Afin de montrer l'intérêt de ces ressources pour la recherche en linguistique de corpus – en particulier en ce qui concerne l'étude de la phraséologie et de ses prolongements – nous présentons dans cet article une architecture dédiée à l'exploration de corpus arborés, baptisée le Lexicoscope. Cette architecture, initialement développée pour interroger des corpus multilingues de grande dimension (le corpus Emolex, Diwersy et al. 2014), est spécialement conçue pour permettre un accès rapide aux cooccurrents syntaxiques des unités étudiées, sans toutefois requérir la maîtrise d'un formalisme complexe pour la formulation des requêtes.

Après avoir examiné quelques plate-formes similaires conçues pour la recherche d'expressions à travers des corpus arborés, telles que Scientext ou le Sketch Engine, nous verrons comment le Lexicoscope réunit des fonctionnalités complémentaires de concordance, de recherche d'expressions complexes et de tableaux de cooccurrents – la richesse des outils sous-jacents restant toutefois accessible pour l'utilisateur non spécialisé grâce à des principes ergonomiques originaux tels que la reformulation de requête basée sur l'exemple. Nous finirons par la présentation d'un scénario d'utilisation de l'interface pour l'extraction d'expression complexes, testée auprès d'étudiants de master en Industries de la langue.

2. Exploration de corpus arborés pour l'étude de la phraséologie

De très nombreux outils permettent d'interroger des corpus afin d'extraire des concordances autour d'une expression pivot (ou mot pôle). Des applications telles que WordSmith, AntConc[1], TXM (Heiden, 2010), Nooj (Silberztein, 2015), ou paraConc[2] (pour des corpus parallèles), donnent accès à des fonctionnalités de concordance ainsi qu'à des outils de textométrie (liste de fréquences, de cooccurrence, etc.), voire pour certains à des modules de traitements de la langue. À ces applications s'ajoutent les outils d'exploration de corpus en ligne : pour tous les grands corpus de référence comme le *British National Corpus* pour l'anglais britannique, le *Corpus Of Contemporary American English* pour l'anglais américain, ou *Frantext* pour le français, le concordancier apparaît comme la principale modalité

[1] cf. http://www.laurenceanthony.net/software/antconc/, février 2016.

[2] cf. http://www.athel.com/para.html, février 2016.

d’exploration, les sorties au format KWIC[3] étant le mode de présentation à la fois naturel et standardisé pour ce type de recherche. Grâce à ces outils, on peut parcourir le texte sur un mode hypertextuel, la concordance établissant un lien entre des contextes éloignés qui partagent un rapport analogique selon l’axe paradigmatique. Ce mode d’appréhension non linéaire de la textualité n’est pas neuf, puisqu’il remonte aux traditions médiévales d’exégèse de la bible, notamment avec les concordances d'Hughes de Saint-Cher, dès le XIIIe siècle.

Dans l’interrogation des grands corpus en ligne, on peut en général s’appuyer sur la lemmatisation et l’étiquetage morphosyntaxique, pour effectuer sa recherche en fonction des parties du discours. Ainsi, l’expression de recherche peut porter sur des lemmes, des catégories morphosyntaxiques, des traits. La figure 1 montre l’exemple d’une sortie KWIC obtenue dans l’interrogation du corpus COCA, pour lemme *search* pris en tant que nom. Comme on le voit, la requête peut s’appuyer sur l’étiquette N pour désambiguïser le lemme, mais les étiquettes apparaissent également dans les sorties sous forme de couleurs associées aux formes cooccurrentes. On peut ainsi distinguer, d’un seul coup d’œil, tous les adjectifs (en vert à l’écran) ou tous les noms (en bleu à l’écran).

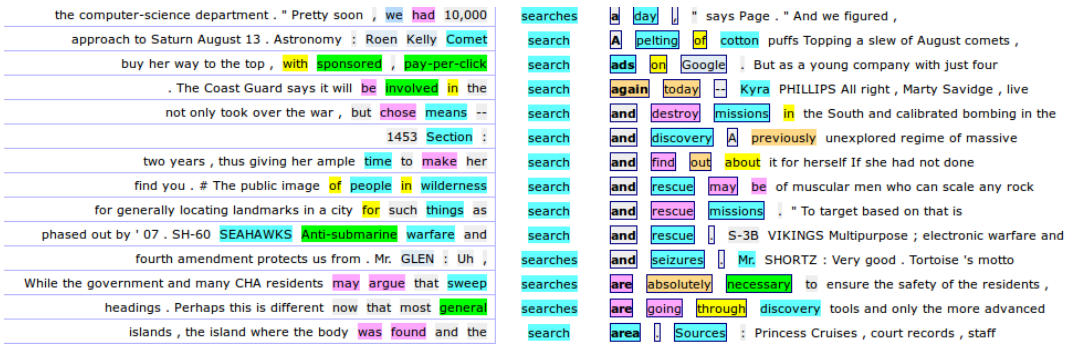


Figure 1 : Résultats pour la requête `[search].[N*]` (corpus COCA)

Le recours à la morphosyntaxe dans ce type de requête peut être très utile pour l’étude de la phraséologie. Par exemple, dans l’exemple de la figure 1, on voit que l’expression *search and rescue* (litt. « recherche et sauvetage ») est récurrente : il s’agit d’un composé adjectival terminologique s’appliquant à un certain type d’opération de secours. Pour mieux cerner la distribution de l’unité, on peut rechercher tous les noms qu’elle qualifie, avec une requête du type `search and rescue [N]` : on trouve des noms tels que *mission*, *operation*, *effort*, *team*, *pilot*, *dog*, *volonteer*. L’expression *search and rescue effort* est intéressante d’un point de

[3] KWIC pour *Key Word In Context*, ce format permet d’afficher les expressions pivots dans une colonne centrale, entourées de leurs contextes gauche et droite.

vue phraséologique, car non traduisible mot-à-mot en français, *effort* étant ici équivalent à « opération ». Par ailleurs, l'occurrence de *search and destroy missions* suggère un pattern plus général *search and* + *V*, qui pourrait s'appliquer à d'autres types d'opération. La requête `search and [V] mission` permet de répondre à la question : on ne trouve que *search and destroy*, qui s'utilise dans un contexte militaire et/ou sécuritaire.

Pour aller plus loin dans ce type d'exploration, il paraît naturel de s'appuyer sur les cooccurrents syntaxiques, en recherchant ces derniers non pas dans une fenêtre fixe à gauche ou à droite de l'expression pivot, mais en ciblant directement les mots qui lui sont syntaxiquement liés. Dans le cas présent, on pourrait par exemple rechercher les verbes dont l'expression *search and rescue mission* est l'objet. Comme le montre l'exemple de la figure 2, la simple cooccurrence de surface des verbes, dans une fenêtre de 4 mots à gauche 4 mots à droite, produit beaucoup de bruit – ici seuls 4 verbes sur 12 correspondent à ce qu'on cherche (*deploy*, *complete*, *conduct*, *launch*).

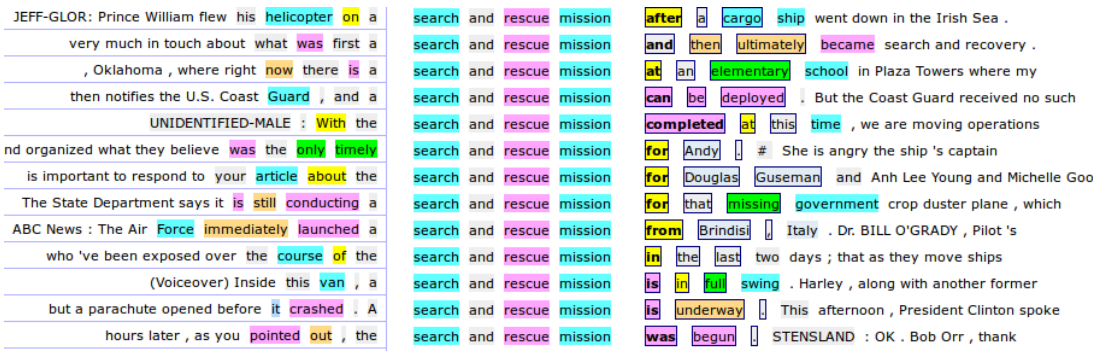


Figure 2 : Résultats pour la requête `search and rescue mission` (corpus COCA)

Par ailleurs ce type de recherche génère aussi du silence. Certaines cooccurrences ne sont pas détectées, car elles dépassent la fenêtre fixée *a priori*. À la sixième ligne de la figure 2 : « and **organized** what they believe was the only timely *search and rescue mission* for Andy » le verbe est beaucoup trop loin de l'expression pivot pour être détecté.

Cependant, à l'heure actuelle, il existe peu d'interfaces d'interrogation conçues pour s'appuyer sur le niveau syntaxique. Parmi les applications, TigerSearch (König & Wolfgang, 2003) permet de mettre en œuvre un formalisme riche pour interroger des arbres syntaxiques exprimant des constituants immédiats. On peut ainsi constituer des requêtes pour définir des relations de dominance (gouverneur>dépendant) et de précédence (précédent.successeur), combinées à des

traits morphosyntaxiques (étiquettes de constituants, parties du discours, etc.). Par exemple, la requête `#np:[cat="NP"] & #np > [pos="ADJA"] & #np > [pos="NN"]` permet de définir un syntagme nominal nommé #np, étiqueté par NP, et dominant un adjectif d'étiquette ADJA ainsi qu'un nom d'étiquette NN. Un mode d'édition graphique est également proposé pour élaborer visuellement l'arbre syntaxique correspondant.

Suivant des principes similaires, mais pour une annotation en dépendances plutôt qu'en constituants, l'interface web ANNIS[4], accessible en ligne, permet d'interroger un corpus de textes latins et grecs classiques. Les requêtes peuvent être élaborées selon un formalisme mêlant contraintes de traits et relations. Les résultats sont affichés sous forme de concordances assorties, de manière optionnelle, des différentes couches d'annotation en traits et en relations (cf. figure 3).

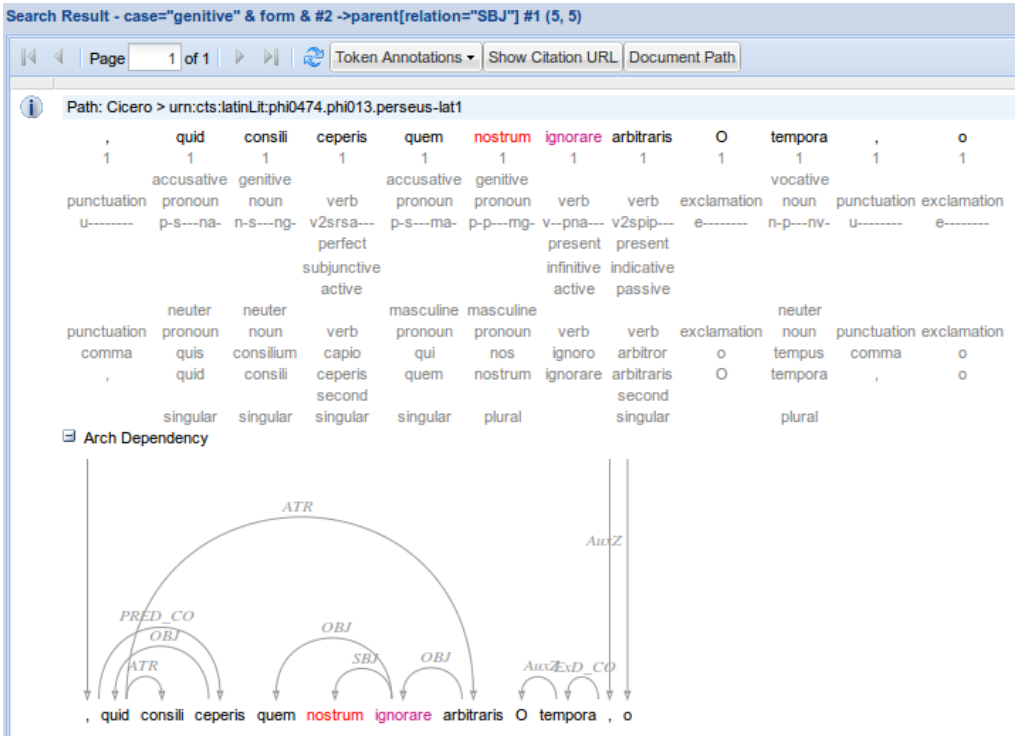


Figure 3 : Résultats pour la requête `case="genitive" & form & #2 ->parent[relation="SBJ"] #1` (corpus ANNIS/CICERO)

Ce type d'outils s'adresse cependant à un public spécialisé, capable de maîtriser à la fois le métalangage des étiquettes associé à l'analyse en constituants ou en dépendances et un formalisme de requête symbolique à la syntaxe assez complexe.

[4] cf. http://annis.perseus.tufts.edu/, février 2016.

Parmi les applications accessibles sur le web, on trouve quelques interfaces permettant l'exploration de corpus arborés selon des modalités plus simples et plus intuitives : nous citerons principalement ScienQuest (Falaise et al., 2011) et le Sketch Engine (Kilgarriff et al., 2004), ce dernier outil étant plus spécifiquement ciblé sur l'étude des collocations, et donc de la phraséologie.

L'interface de ScienQuest[5] a été conçue sur la base du langage de requête de ConcQuest (décrit plus loin) permettant de définir des expressions complexes combinant des contraintes sur différents niveaux : forme de surface, lemme, partie du discours, traits. Le caractère novateur de cette interface réside dans la mise en œuvre d'une assistance graphique simplifiée pour la construction de requêtes complexes. Par un système de boîtes représentant les différents mots composant l'expression, l'utilisateur peut formuler ses contraintes sans se soucier de la syntaxe du langage sous-jacent, ni du codage des étiquettes. Par exemple, si l'on cherche tous les verbes ayant pour objet direct la forme *étude*, on pourra construire la requête représentée figure 4. Ce système de construction graphique joue le double rôle d'assistant et de guide pour l'apprentissage du formalisme, l'expression résultante étant accessible dans un onglet « avancé » à mesure que la requête est construite graphiquement. L'utilisation directe du formalisme de requête est possible dans le mode d'interrogation avancée : par ce biais, l'utilisateur aguerri peut accéder à toute la puissance expressive du langage pour des recherches plus ciblées.

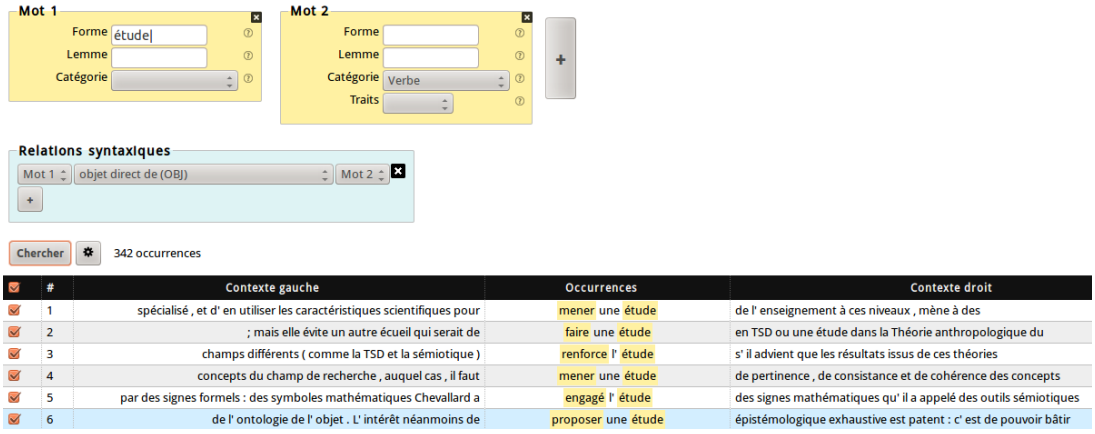
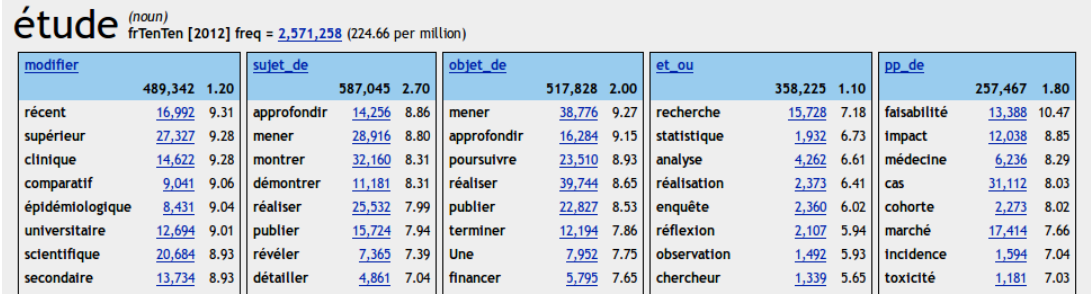


Figure 4 : Mise en œuvre d'une requête portant sur *étude* pris en position de COD (interface ScienQuest).

Quant au Sketch Engine, son intérêt ne se situe pas dans la formulation d'expressions complexes pour les pivots de concordance, mais dans les *word sketches*, des tableaux regroupant les principaux cooccurrents syntaxiques (les cooccurrents reliés par une relation de dépendance) classés en fonction des

[5] cf. http://corpora.aiakide.net/scientext18/, février 2016.

principales relations – objet, sujet, modifieurs, ... – donnant ainsi une vision synthétique de la combinatoire du mot pivot, comme le montre la figure 5.



étude *(noun)*
frTenTen [2012] freq = 2,571,258 (224.66 per million)

modifier			sujet_de			objet_de			et_ou			pp_de		
	489,342	1.20		587,045	2.70		517,828	2.00		358,225	1.10		257,467	1.80
récent	16,992	9.31	approfondir	14,256	8.86	mener	38,776	9.27	recherche	15,728	7.18	faisabilité	13,388	10.47
supérieur	27,327	9.28	mener	28,916	8.80	approfondir	16,284	9.15	statistique	1,932	6.73	impact	12,038	8.85
clinique	14,622	9.28	montrer	32,160	8.31	poursuivre	23,510	8.93	analyse	4,262	6.61	médecine	6,236	8.29
comparatif	9,041	9.06	démontrer	11,181	8.31	réaliser	39,744	8.65	réalisation	2,373	6.41	cas	31,112	8.03
épidémiologique	8,431	9.04	réaliser	25,532	7.99	publier	22,827	8.53	enquête	2,360	6.02	cohorte	2,273	8.02
universitaire	12,694	9.01	publier	15,724	7.94	terminer	12,194	7.86	réflexion	2,107	5.94	marché	17,414	7.66
scientifique	20,684	8.93	révéler	7,365	7.39	Une	7,952	7.75	observation	1,492	5.93	incidence	1,594	7.04
secondaire	13,734	8.93	détailler	4,861	7.04	financer	5,795	7.65	chercheur	1,339	5.65	toxicité	1,181	7.03

Figure 5 : Extrait d'un *word sketch* pour le nom *étude*

Par ailleurs, des *word sketches* différentiels permettent de comparer les profils combinatoires de deux mots, en affichant dans un même tableau les cooccurrents communs et les cooccurrents spécifiques.

3. Le Lexicoscope : un outil dédié à l'étude de la combinatoire lexico-syntaxique

Le Lexicoscope (Kraif, Diwersy, 2012) a été initialement développé pour l'interrogation des corpus arborés du projet Emolex : il s'agit d'une extension de l'interface EmoConc accessible au sein de la plateforme EmoBase (Diwersy et al., 2014). Dans sa conception, le Lexicoscope hérite à la fois des principes de ScienQuest, pour la définition d'expression complexes intégrant éventuellement des contraintes syntaxiques, et du Sketch Engine, par la présentation des cooccurrents syntaxiques de l'expression pivot. Au niveau des résultats de requête, le Lexicoscope permet donc d'accéder aussi bien à des concordances – éventuellement bilingues lorsque le corpus d'étude sélectionné est parallèle – qu'à des tableaux de cooccurrents. Plutôt que des *word sketches*, nous nommons ces derniers des *lexicogrammes* (le terme est emprunté à Tournier & Heiden, 1998). Tout comme dans le Sketch Engine, il est possible de comparer les cooccurrents de différents pivots. Pour ce faire, nous proposons deux types d'affichage : tableau croisé (à partir de 2 pivots) ou analyse multivariée (pour au moins 4 pivots). Si l'affichage sous forme de tableau croisé est moins synthétique, au plan visuel, que celui d'un *word sketch* différentiel pour la comparaison de deux pivots, les représentations sous forme d'analyse factorielle des correspondances (AFC), de classification hiérarchique ou d'échelonnement multidimensionnel permettent de comparer les profils d'un grand nombre de pivots. L'AFC, notamment, permet d'identifier quels sont les cooccurrents principaux qui caractérisent les différentes zones de regroupement des pivots à travers leur projection sur un plan. Dans la figure 6, par

exemple, la comparaison des lexicogrammes de plusieurs noms d'affect en position d'objet montre une répartition en trois groupes, en fonction des verbes de manifestation associés : 1/ des sentiments négatifs exprimant plutôt la distance ; 2/ des sentiments interpersonnels positifs, que l'on exprime directement à leur objet ; 3/ des sentiments interpersonnels assez intenses qui s'expriment sous la forme de la confidence, par l'aveu ou du dévoilement.

1/ *montrer / afficher / manifester + dédain / détachement / indifférence*

2/ exprimer / signifier / témoigner + respect / sympathie / reconnaissance

3/ avouer / (ne pas) cacher / (ne pas) dissimuler + mépris / fascination / admiration

Ces distributions, par les regroupements qu'elles opèrent, donnent un éclairage à la fois sur les nuances sémantiques associées à ces noms et sur les verbes de manifestation associés (intensité, attitude de l'expérienceur vis-à-vis de son affect, relations de synonymie) : dans ce corpus à dominante journalistique, si le *dédain* « s'affiche », sans être verbalisé, le *mépris*, plus intense et aussi plus compromettant, s'exprime plus souvent par un euphémisme et « ne se cache pas » – quand la *fascination*, qui révèle peut-être une forme de faiblesse ou d'infériorité, se confie sous la forme de l'aveu.

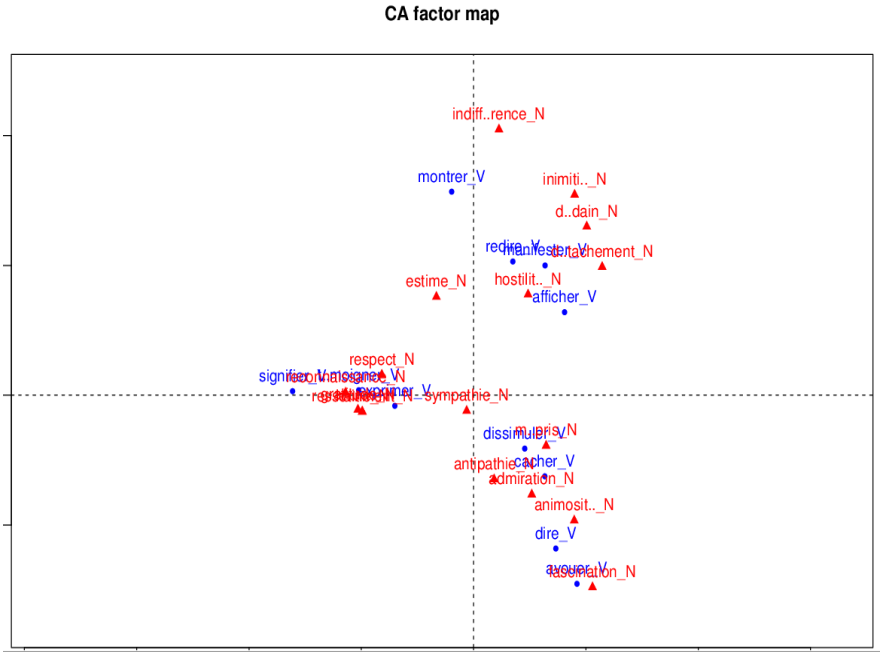


Figure 6 : Comparaison par AFC des lexicogrammes de plusieurs noms d'affect (corpus Emolex)

Avec le Lexicoscope, l'utilisateur a la possibilité, en amont, de préciser le sous-ensemble des relations de dépendance caractérisant l'espace de cooccurrence retenu (p.ex. verbe-objet, sujet-verbe, etc.). Le lexicogramme extrait regroupe ensuite en un seul tableau tous les cooccurrents trouvés, sans différencier les relations comme dans le Sketch Engine. La figure 7 montre un extrait de lexicogramme pour le lemme *étude* pris dans diverses relations exprimant qu'il est objet direct (~OBJ), objet profond dans une tournure passive (~DEEPOBJ) ou modifié par un participe passé correspondant à un passif réduit (NMOD), par exemple *l'étude menée*.

Show 25 entries Search:

l1	l2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likelihood
étude_.*	mener_VERB	NMOD ~OBJ ~DEEPOBJ	76	673	932	529095	10	497,6299	1
étude_.*	réaliser_VERB	NMOD ~OBJ ~DEEPOBJ	51	673	1416	529095	9	248,0645	2
étude_.*	conduire_VERB	~DEEPOBJ NMOD	17	673	861	529095	8	62,1044	3
étude_.*	consacrer_VERB	NMOD ~OBJ	10	673	605	529095	6	33,0990	4
étude_.*	compléter_VERB	~OBJ	8	673	337	529095	5	31,9392	5
étude_.*	publier_VERB	NMOD ~DEEPOBJ ~OBJ	9	673	571	529095	5	28,9818	6

Figure 7 : extrait d'un lexicogramme pour le lemme étude (corpus Scientext)

Un des principaux avantages du Lexicoscope tient dans la possibilité de définir un « pivot complexe », c'est-à-dire comportant plusieurs mots et correspondant éventuellement à une structure arborée. Cette fonctionnalité repose sur la définition préalable d'un langage de requête riche permettant de mêler différents niveaux de contraintes.

3.1. *Langage de requête*

Comme nous l'avions montré pour l'élaboration du langage de requête de ConcQuest (Kraif, 2008), il est important qu'un tel langage d'interrogation puisse permettre une progressivité dans la complexité des requêtes : l'utilisateur novice doit pouvoir élaborer dans un premier temps des requêtes simples, par mots-clés, comme dans toutes les interfaces grand public – moteurs de recherche classiques ou moteurs de concordances tels que Linguee[6] ou WebCorp[7].

Le formalisme initialement développé pour ConcQuest est conçu pour intégrer un passage progressif des requêtes simples vers les requêtes avancées. Les mots graphiques, simplement séparés par un espace sont acceptés tels quels, et définissent

[6] http://www.linguee.fr/, février 2016.

[7] http://www.webcorp.org.uk/live/, février 2016.

des formes de surface. Lorsque l'utilisateur a besoin de rechercher un lemme, plutôt qu'une forme fléchie, il peut préfixer le lemme par le caractère % :

`%avoir peur` → permet de trouver *avoir peur*, *avait peur*, etc...

Pour l'ajout de contraintes morphosyntaxiques, l'utilisateur doit utiliser des chevrons et lister ces contraintes sous la forme d'une liste `attribut=valeur`, séparée par des virgules :

```
<l=avoir,f=.*infi.*,c=V> <l=peur>
```

On voit par ailleurs qu'il est possible d'utiliser des expressions régulières dans la définition d'une valeur. Ces expressions régulières peuvent intervenir à deux niveaux, dans la définition des valeurs d'attribut, mais aussi au niveau des tokens eux-mêmes. Par exemple `<l=avoir,c=V> <>? <l=peur>` permet de considérer l'insertion d'un token quelconque entre *avoir* et *peur*. Les expressions régulières présentent l'avantage d'être un langage très puissant en termes d'expressivité, tout en étant progressif en termes de complexité : un utilisateur ne connaissant pas toutes les finesses du langage peut néanmoins écrire des expressions très simples en se limitant aux opérateurs de base, tels que ? (pour un token facultatif) et | (la disjonction, pour exprimer des tokens alternatifs).

Pour les sorties au format KWIC, il faut définir quel va être le pivot de la sortie (la colonne centrale). Pour ce faire, on attribue des identifiants #1, #2, … etc. aux différents tokens de la requête : `<l=avoir,c=V,#1> <l=peur,#2>`. L'identifiant #1 correspond au pivot central, et les autres tokens identifiés apparaissent en surbrillance dans les sorties.

On peut rechercher plusieurs expressions dans la même phrase, en les connectant avec l'opérateur `&&`. Par exemple, si l'on cherche des occurrences de *avoir peur* dans une phrase où le mot *courage* apparaît également, on écrira :

```
<l=avoir,c=V><l=peur,#1>&&<l=courage>
```

Dans ce cas, aucune contrainte séquentielle n'est imposée entre les termes de la requête connectés par `&&` : *avoir peur* peut apparaître avant ou après *courage*. En revanche, *avoir* doit apparaître avant *peur*.

Si l'on s'intéresse à toutes les réalisations syntaxiques de la collocation *avoir peur*, on voit bien qu'il faut s'affranchir de la contrainte séquentielle entre *avoir* et *peur* et chercher la cooccurrence dans une fenêtre élargie (p.ex. pour l'énoncé *la peur que j'ai eue !*). On écrira donc `<l=avoir,c=V,#2>&&<l=peur,#1>`, mais en ajoutant une contrainte d'ordre syntaxique, exprimant le fait qu'il y a une relation de dépendance directe entre les deux. Les relations de dépendances sont exprimées entre parenthèse sous la forme (REL,ID1,ID2), et séparées du reste de la requête par deux fois deux points : `<l=avoir,c=V,#2>&&<l=peur,#1>::(OBJ,2,1)`

On peut ici aussi utiliser des expressions régulières si l'on ne veut pas qualifier précisément la relation : `<l=avoir,c=V,#2>&<l=peur,#1>::(.*,2,1)`

Notons qu'avec l'expression `.*`, la hiérarchie entre 2 et 1 n'est pas définie : il se peut que 2 gouverne 1 ou l'inverse. Dans le Lexicoscope, les relations préfixée par le caractère ~ vont du dépendant vers le gouverneur. Ainsi, si l'on écrit `(~.*,2,1)` on impose que 1 soit le gouverneur.

3.2. *Reformulation de requêtes pour les expressions polylexicales*

Avec des requêtes par mots-clés, c'est au système de rechercher les résultats s'approchant le plus de la requête initiale, triés par ordre de pertinence. Dans un second temps, le système doit pouvoir permettre à l'utilisateur d'affiner sa recherche. Pour ce faire, on propose traditionnellement deux types de fonctionnalités :

- l'expansion de requête : il s'agit de proposer, en fonction de ce qui a été trouvé dans le corpus, une reformulation de la requête susceptible d'élargir la recherche, en proposant des expressions voisines plus productives que la requête initiale.

- la recherche avancée : on propose à l'utilisateur de raffiner ses critères de recherche en cochant certaines options, voire en mettant en œuvre un formalisme plus complexe et plus riche, tel que les expressions régulières, XPATH, CQP, etc.

Mais l'observation des usages nous indique que peu d'utilisateurs recourent au mode de recherche avancé. Comme le notent avec justesse Augustinus et al. (2012), ces systèmes pèchent par manque d'ergonomie et de standardisation[8]. Pour remédier à ce problème, les auteurs ont proposé un ingénieux système de requête « basée sur l'exemple » : l'utilisateur écrit d'abord sa requête en langage naturel, après quoi celle-ci est parsée afin de produire un arbre syntaxique, à partir duquel une requête XPATH est générée pour interroger le corpus arboré, formaté en XML. L'idée s'apparente au mode d'expansion de requête que nous avions proposé en 2008 (Bouallegue, 2008) : partir d'un exemple pour ensuite le généraliser, afin de récupérer toutes les constructions analogues à l'exemple initial.

Le Lexicoscope reprend cette idée, mais de façon plus simple, car son implémentation ne nécessite aucun parseur. L'exemple tapé par l'utilisateur est tout simplement recherché en surface et « à plat », c'est-à-dire dans l'ordre où les formes sont données. Pour ce faire, la requête exemple est généralisée de façon rudimentaire : une expression régulière est générée afin d'autoriser des insertions de

[8] "A major obstacle is the limited user-friendliness of the query languages and search tools. That problem is closely related to another issue of treebank mining: the lack of standardisation in both tree-banks and query languages." Augustinus et al. (2012:3162)

0, 1 ou 2 tokens entre les formes recherchées. Par exemple, si l'utilisateur formule la requête "*entre dans une colère noire*", on génère l'expression : `<w=entre,#1><>?<>?<w=dans,#2><>?<>?<w=une,#3><>?<>?<w=colère,#4><>?<>?<w=noire,#5>`. Cette expression est ensuite recherchée dans le corpus. Les premières occurrences trouvées fournissent alors des arbres (éventuellement différents), qui permettent de reformuler la requête avec les relations syntaxiques sous-jacentes et les lemmes correspondants. L'utilisateur n'est pas contraint à connaître de formalisme ni le métalangage des étiquettes : il doit seulement sélectionner l'arbre qui correspond à sa recherche. Dans le cas présent, l'arbre obtenu est le suivant :

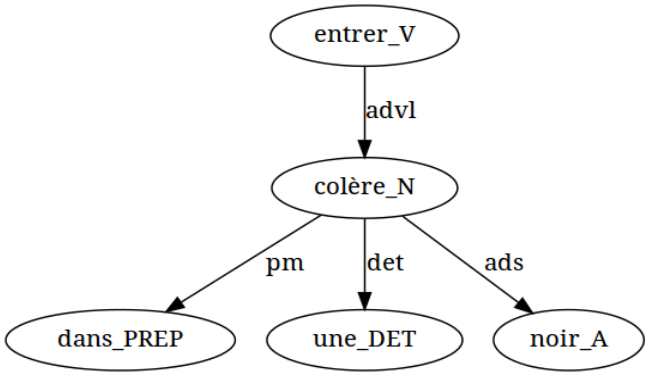


Figure 8 : Arbre extrait pour la reformulation de la requête "entre dans une colère noire".

La requête correspondante est générée automatiquement dans l'interface avancée, de sorte que l'utilisateur puisse la modifier :

`<l=entrer,c=V,#1>&&<l=dans,c=PREP,#2>&&<l=une,c=DET,#3>&&<l=colère,c=N,#4>&&<l=noir,c=A,#5>::(ads,4,5)(advl,1,4)(det,4,3)(pm,4,2)`

Par exemple, si l'utilisateur supprime le lemme `l=noir` pour le token #5, il obtiendra tous les autres adjectifs susceptibles d'occuper la même position : *violente*, *terrible*, etc.

Si le corpus sélectionné par l'utilisateur comporte des systèmes d'annotation syntaxique différents, les différents arbres trouvés sont proposés à l'utilisateur, qui doit alors être conscient que sa requête est adaptée pour un sous-ensemble du corpus choisi. Par exemple, avec la requête "*se mettre en colère*", on trouve les deux arbres de la figure 9, correspondant à deux parseurs différents (Connexor et XIP).

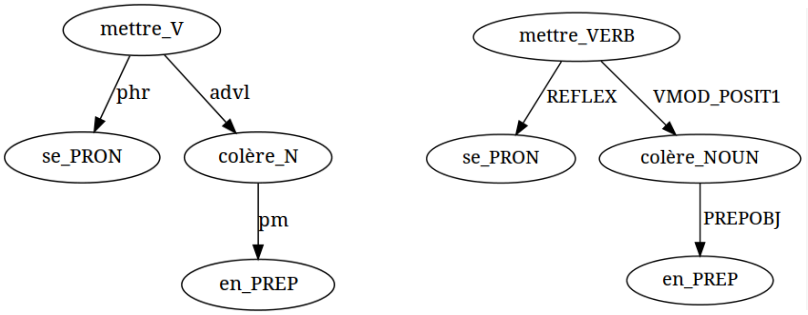


Figure 9 : Deux arbres extraits pour la reformulation de la requête `se mettre en colère`.

3.3. *Extraction d'arbres lexicosyntaxiques récurrents (ALR)*

La possibilité d'obtenir les lexicogrammes pour des pivots complexes permet de mettre en œuvre une technique d'extraction automatique des arbres lexico-syntaxiques récurrents (ALR) : partant d'une expression de départ (par exemple *colère*), on lance l'extraction de son lexicogramme, et l'on trouve des collocatifs fréquents (par exemple *mettre* ou *entrer*). On réitère ensuite récursivement pour le pivot augmenté de ses collocatifs (ici *mettre+colère* et *entrer+colère*), jusqu'à obtenir des arbres formant des expressions récurrentes complètes (*se mettre en colère*, ou *entrer dans une violente colère*). Le début de ce processus est représenté sur la figure 10

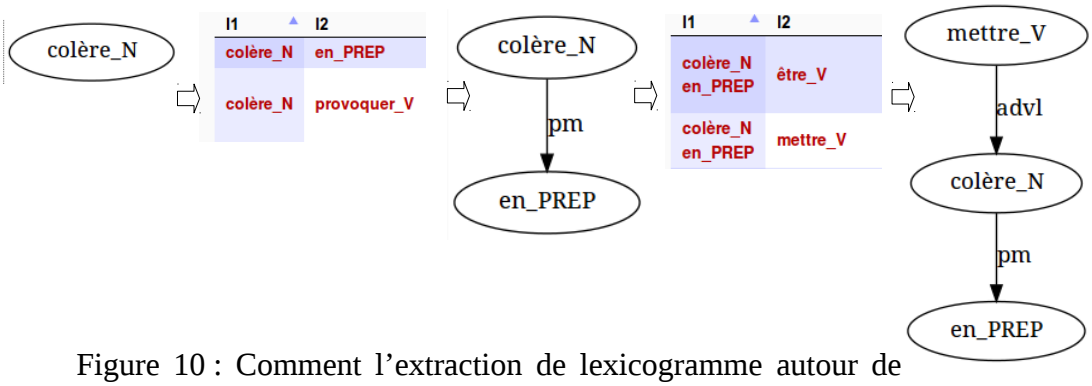


Figure 10 : Comment l'extraction de lexicogramme autour de pivots complexes permet d'extraire automatiquement des ALR : exemple d'extraction automatique de *mettre en colère.*

4. Exemple de scénario d'utilisation

Nous avons développé, dans le cadre d'un cours du master Industries de la langue de l'université Stendhal Grenoble 3, un scénario d'utilisation du Lexicoscope, dans le but de guider les étudiants dans l'exploration de la combinatoire lexicale.

Le corpus de l'étude est constitué d'articles scientifiques collectés dans le cadre du projet Scientext (environ 5 millions de mots répartis entre 10 disciplines de SHS). L'étude porte sur l'étude des collocations pour des noms relatifs au lexique transdisciplinaire, tels que *résultat*, *question*, *hypothèse*, *expérience*, *conclusion*, *recherche*, *étude*, *discussion*, *théorie*. L'objectif de l'étude est de caractériser les principales collocations liées à ces entrées, en décrivant finement leur combinatoire lexico-syntaxique en rapport avec le genre des textes (article scientifique) et les domaines (disciplines des SHS). Prenons l'exemple du nom *analyse*. Une première extraction de lexicogramme indique les principaux collocatifs suivants (pour les verbes et les adjectifs) :

verbes : *réaliser, montrer, effectuer, mener, permettre, révéler, proposer, détailler, centrer, confirmer, conduire ;*

adjectifs : *statistique, fine, détaillée, quantitative, comparative, qualitative, approfondie ;*

Dans un deuxième temps, pour chacune des collocations trouvées, il est possible d'approfondir l'étude en identifiant comment celles-ci se réalisent en contexte : détermination, genre et nombre, temps et voix pour les verbes, etc. Examinons le cas de : *réaliser + analyse*. Si l'on s'intéresse à la diathèse, l'utilisateur pourra taper "`analyse réalisée`" et "`réalisé analyse`". Le système lui proposera alors les arbres de la figure 9, correspondant respectivement au passif réduit (*Ces observations vont être confirmées par les* ***analyses réalisées*** *au niveau des groupes [...]*) au passif (*Une* ***analyse*** *a été* ***réalisée*** *[...]*), et à la voix active (*Nous avons* ***réalisé*** *une* ***analyse*** *[…]*).

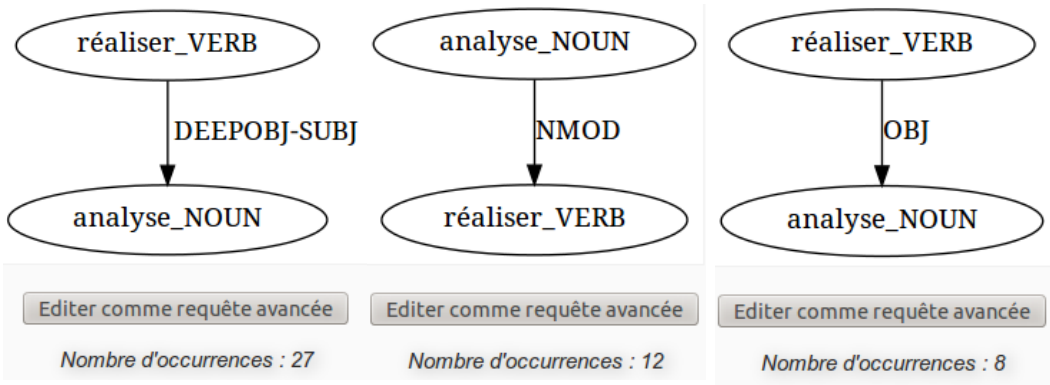


Figure 11 : Copies d'écran – différents arbres extraits pour la collocation *réaliser + analyse*.

Les statistiques d'occurrences montrent de façon très claire une préférence marquée pour la voix passive, complète (27) ou réduite (12). Un lexicographe désirant donner un exemple représentatif des réalisations de cette collocation, pour

ce genre de texte, choisira donc plutôt une attestation telle que "*Une analyse a été réalisée...*", et indiquera que cette tournure est saillante dans l'usage. Par ailleurs, la répartition des occurrences montre que la collocation est massivement utilisée en psychologie (42 occurrences sur 59) et dans une moindre mesure en sciences de l'éducation (6 occurrences sur 59), les 8 autres disciplines en faisant un usage assez marginal (entre 0 et 4 occurrences). Pour aller plus loin, on peut extraire les ALR autour de la collocation *réaliser+analyse*. On trouve alors des collocations récursives à caractère terminologique ressortissant au domaine des statistiques et de l'analyse de données : *réaliser + analyse de variance*, *réaliser + analyse factorielle*, *réaliser + analyse confirmatoire*, *réaliser + analyse de régression.* Sans maîtrise des étiquettes ni du formalisme de requête sous-jacent, il est donc possible, avec ce type d'outil, de tirer le meilleur parti d'une annotation syntaxique des corpus pour effectuer une étude fine de la phraséologie étendue.

5. Conclusion et perspectives

Le Lexicoscope, à travers ses versions successives, a été utilisé dans plusieurs études ayant trait à la phraséologie : les projets Emolex et Termith, et bientôt le projet PhraseoRom qui s'intéresse à la comparaison des unités phraséologiques à travers différents sous-genre littéraires. Par la mise en œuvre de la recherche d'expressions complexes, correspondant à des arbres syntaxiques, nous avons montré que l'extraction de concordances et de lexicogrammes (des tableaux de cooccurrents) devenait plus puissante et plus riche pour l'étude de la phraséologie, notamment par la possibilité d'extraire des arbres lexico-syntaxiques récurrents de manière automatisée. Bien que la reformulation de requête basée sur l'exemple permette de simplifier considérablement l'appréhension de l'outil, des progrès importants restent à faire en ce qui concerne l'ergonomie et la présentation des résultats. Notamment, nous prévoyons de rendre modifiables, par un mode d'édition graphique, les arbres renvoyés par les exemples trouvés dans le corpus. De la sorte, les utilisateurs pourront d'autant mieux intervenir sur la construction de leur requête. Par ailleurs, nous prévoyons de travailler sur un mode d'affichage synthétique de la combinatoire liée à un pivot complexe, en automatisant le parcours que nous avons présenté dans le scénario d'utilisation (partie 4). Ainsi, pour une expression phraséologique donnée, on pourra répondre rapidement aux questions suivantes : quelles sont ses réalisations morphosyntaxiques les plus fréquentes ? des traits liés au nombre, à la détermination, au temps, au mode, à la voix, etc. sont-ils statistiquement saillants ? d'autres collocations y sont-elles liées ?

6. Références bibliographiques

BOUALLEGUE Mohamed (2008) : *Expansion de requête pour le repérage d'occurrences de constructions polylexicales exprimées sous leurs formes canoniques,* Mémoire de Master, sous la dir. de O. Kraif, Université Stendhal Grenoble 3.

DIWERSY Sascha, GOOSSENS Vannina, GRUTSCHUS Anke, KERN Beate, KRAIF Olivier, MELNIKOVA Elena et NOVAKOVA Iva (2014) : « Traitement des lexies d'émotion dans les corpus et les applications d'EmoBase » , *Revue Corpus* No 13/2014, p. 269-293.

FALAISE Achille, TUTIN Agnès et KRAIF Olivier (2011) : « Définition et conception d'une interface pour l'exploitation de corpus arborés pour non-informaticiens : la plateforme ScienQuest du projet Scientext », *Revue TAL*, Volume 52 – n° 3/2011, p. 103-128.

HEIDEN Serge (2010) : « The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme », in *24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japon, ENS-Lyon, p. 10.

KILGARRIFF Adam, RYCHLY Pavel, SMRZ Pavel et TUGWELL David (2004) : « The Sketch Engine », *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France, p. 105-116.

KÖNIG Esther et LEZIUS Wolfgang (2003) : *The TIGER language - A Description Language for Syntax Graphs, Formal Definition. Technical report IMS*, Universität Stuttgart, Germany.

NIVRE Joakim, HALL Johan, KÜBLER Sandra, MCDONALD Ryan, NILSSON Jens, RIEDEL Sebastian et YURET Deniz (2007) : « The CoNLL 2007 shared task on dependency parsing », in *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, Prague, June 2007, p. 915-932.

SILBERZTEIN Max (2015) : *La formalisation des langues : l'approche de NooJ*, Paris, ISTE Éditions.

TOURNIER Maurice et HEIDEN Serge (1998) : « Lexicométrie textuelle, sens et stratégie discursive », in *Proc. of I Simposio Internacional de Análisis del Discurso*, Madrid, p. 2287-2300.