Le *Lexicoscope* : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques

Olivier Kraif [1], *Sascha Diwersy* [2]

(1) LIDILEM, Université Stendhal Grenoble 3, BP 25, 38040 Grenoble Cedex

(2) Université de Cologne

olivier.kraif@u-grenoble3.fr, sascha.diwersy@uni-koeln.de

RÉSUMÉ

Dans le cadre du projet franco-allemand Emolex, dédié à l'étude contrastive de la combinatoire du lexique des émotions en 5 langues, nous avons développé des outils et des méthodes permettant l'extraction, la visualisation et la comparaison de profils combinatoires pour des expressions simples et complexes. Nous présentons ici l'architecture d'ensemble de la plate-forme, conçue pour effectuer des extractions sur des corpus de grandes dimensions (de l'ordre de la centaine de millions de mots) avec des temps de réponse réduits (le corpus étant interrogeable en ligne[1]). Nous décrivons comment nous avons introduit la notion de pivots complexes, afin de permettre aux utilisateurs de raffiner progressivement leurs requêtes pour caractériser des constructions lexico-syntaxiques élaborées. Enfin, nous donnons les premiers résultats d'un module d'extraction automatique d'expressions polylexicales récurrentes.

ABSTRACT

The Lexicoscope : an integrated tool for combinatoric profiles observation and lexico-syntactic constructs extraction.

The German-French research project Emolex whose aim is the contrastive study of the combinatorial behaviour of emotion lexemes in 5 languages has led to the development of methods and tools to extract, display and compare the combinatorial profiles of simple and complex expressions. In this paper, we present the overall architecture of the query platform which has been conceived to ensure efficient processing of huge annotated text corpora (consisting of several hundred millions of word tokens) accessible through a web-based interface. We put forward the concept of “complex query nodes” introduced to enable users to carry out progressively elaborated extractions of lexical-syntactic patterns. We finally give primary results of an automated method for the retrieval of recurrent multi-word expressions, which takes advantage of the complex query nodes implementation.

MOTS-CLÉS : collocations, cooccurrences, profil combinatoire, expressions polylexicales, lexique des émotions.

KEYWORDS : collocations, combinatorial profiles, multi-word expressions.

[1] L'accès au corpus sera rendu public, moyennant authentification, d'ici quelques mois.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 399–406,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Cette communication présente des travaux réalisés dans le cadre du projet Emolex, projet franco-allemand cofinancé par l'ANR et la DFG. Dans le cadre de cette recherche, nous avons rassemblé des corpus massifs comportant plusieurs centaines de millions de mots pour 5 langues différentes (l'allemand, le français, l'anglais, l'espagnol et le russe). L'objectif du projet est d'analyser, dans une perspective formulée par Sinclair (2004) ou encore Hoey (2005) et d'un point de vue contrastif, les valeurs sémantiques et les rôles discursifs à partir de la combinatoire du lexique des émotions, afin d'élaborer une cartographie permettant de mieux structurer ce champ lexical, avec des applications en lexicographie mais aussi en didactique des langues et traductologie. Cette étude porte plus précisément sur le développement d'une approche automatisée permettant de guider l'observation linguistique par l'extraction de cooccurrences autour d'un pivot.

2 Un modèle de cooccurrence flexible

Pour caractériser le profil combinatoire d'une entrée, nous reprenons le concept de *lexicogramme*, introduit par Maurice Tournier et repris dans le logiciel WebLex (Heiden, Tournier 1998) : il s'agit d'établir, pour un pivot donné, la liste de ses cooccurrents les plus fréquents, à gauche et à droite, en faisant l'extraction des fréquences de cooccurrence et en calculant des mesures d'association statistiques (telles que rapport de vraisemblance ou t-score). Pour construire ces lexicogrammes, nous proposons un modèle de cooccurrence flexible permettant à l'utilisateur de définir lui-même les *unités de cooccurrences* : formes, lemmes, catégories morphosyntaxiques, traits additionnels (p.ex. sémantiques), relations syntaxiques (dans le cas des *colligations*) ou des combinaisons de ces informations. La possibilité de faire intervenir des combinaisons de ses traits nous semble importante pour permettre à l'utilisateur d'ajuster la focale de ses observations en allant du général au particulier (ou vice-versa), de préciser des contraintes pour désambiguïser certains contextes, et de combiner les aspects lexicaux et syntaxiques dans ses observations. Par ailleurs nous proposons également une caractérisation flexible de *l'espace de cooccurrence,* qui conditionne les points de rencontre entre pivot et collocatifs, ainsi que la manière de les dénombrer. On peut par exemple définir la cooccurrence à l'intérieur d'un empan de largeur fixe, éventuellement différente à droite et à gauche du pivot. Mais on peut aussi rechercher la *cooccurrence syntaxique*, à l'instar de Kilgariff et Tugwell (2001) ou Charest et al. (2010), mise en jeu lorsqu'une relation fonctionnelle (du type sujet, complément d'objet, modifieur, etc.) a été identifiée entre deux unités. Evert (2007), signale l'intérêt de ce type de cooccurrence en terme de bruit et de silence : "(...) unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time introduces less “noise” than textual cooccurrence". Pour la cooccurrence syntaxique, nous exploitons les relations de dépendances obtenues grâce à différents analyseurs : XIP pour l'anglais (Aït-Mokhtar et al. 2001), Connexor pour l'allemand, le français et l'espagnol (Tapanainen & Järvinen 1997), DeSR pour le russe (Attardi et al. 2007), basé sur un modèle stochastique créé à partir du corpus arboré SyntagRus (Nivre *et al.,* 2008). Un post-traitement a permis d'harmoniser et de standardiser l'annotation des relations de dépendance entre les

langues (l'annotation de Connexor ayant servi de référence). Nous avons par la suite complété ces relations pour obtenir des dépendances plus pertinentes sur le plan sémantique (p. ex. sujet profond dans les constructions passives, etc.).

Avec le modèle de cooccurrence ainsi défini, on peut viser des aspects très génériques de la combinatoire (par exemple : quels sont les principaux collocatifs de la forme *surprise* toutes relations confondues) ou beaucoup plus spécifiques et circonscrits (par exemple : quels sont les principaux collocatifs verbaux à l'imparfait du nom lemmatisé *surprise* en tant qu'objet direct). Le tableau 1 montre un tel lexicogramme :

l1	l2	f	f1	f2	loglike
surprise_N	créer_V	614	2098	21658	4548,43
surprise_N	réserver_V	230	2098	2869	2143,50
surprise_N	avoir_V	484	2098	423602	627,50
surprise_N	constituer_V	94	2098	13778	406,80
surprise_N	éviter_V	43	2098	16296	109,30
surprise_N	manifester_V	22	2098	2424	106,62
surprise_N	causer_V	19	2098	2210	90,06
surprise_N	ménager_V	15	2098	1495	75,58
surprise_N	exprimer_V	23	2098	6186	72,54
surprise_N	provoquer_V	23	2098	10551	50,61
surprise_N	feindre_V	9	2098	676	50,31

TABLEAU 1 : extrait du lexicogramme pour le nom lemmatisé *surprise* pris en tant qu'objet direct (f = fréquence de cooccurrence, f1 = fréquence de l1, f2 = fréquence de l2)

3 Visualisations comparatives

A partir de ces lexicogrammes, nous offrons différentes modalités d'exploration :

– pour l'analyse linguistique, le "retour au texte" est indispensable : un simple clic sur un collocatif permet de retrouver, sous forme de concordance, tous les contextes de cooccurrence avec le pivot.

– pour comparer de manière synthétique divers profils combinatoires, nous proposons d'identifier les lexicogrammes à des points dans un espace vectoriel, en ne retenant que la mesure jugée la plus pertinente (fréquence, loglike, t-score, etc.). Il est dès lors possible d'utiliser des méthodes d'analyse de données pour visualiser les similarités entre pivots : analyse factorielle des correspondances (AFC), échelonnement multidimensionnel (MDS) ou classification hiérarchique ascendante (hClust). La figure 1 montre ces sorties pour des unités du domaine sémantique de la 'colère' (obtenues grâce aux modules du projet 'GNU R'). La classification, réalisée pour la relation "objet", indique une hiérarchisation assez bien corrélée à l'intensité du sentiment. Quant à la 'factor map', réalisée pour des relations quelconques

concernant des collocatifs adjectivaux, elle permet de distinguer trois groupes : *révolte*, *indignation* - souvent lié à la sphère publique et politique ; *fureur*, *rage*, *colère* - lié à l'expression ponctuelle et plus ou moins intense de l'affect ; enfin *énervement*, *irritation*, *exaspération* - qui concernent plutôt des états émotionnels précurseurs de cette manifestation. Ces cas montrent de façon assez éclairante le lien entre les valeurs sémantiques et la combinatoire lexico-syntaxique.

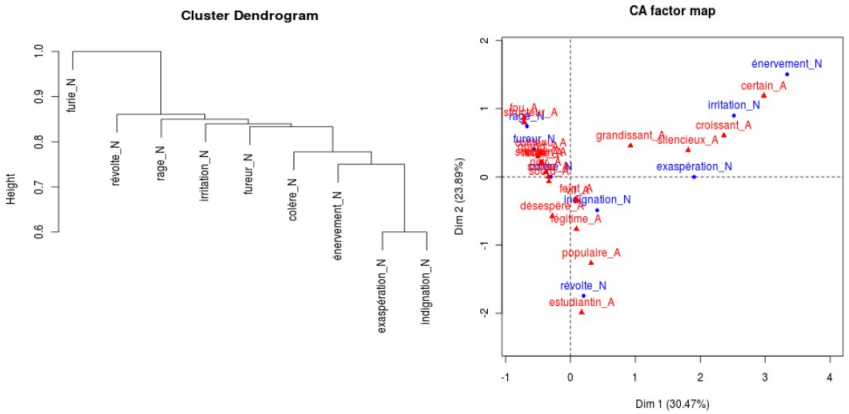

FIGURE 1 : Classification hiérarchique et AFC (domaine sémantique de la 'colère')

4 Architecture logicielle

Comment répondre rapidement à une requête d'utilisateur lorsqu'on interroge des corpus contenant des centaines de millions d'occurrences ? La réponse est simple, a priori : grâce à une indexation préalable des unités et des cooccurrences. Mais la difficulté de notre système tient au fait que ni les unités, ni l'espace de cooccurrence ne sont définis à l'avance : on peut interroger des lemmes, des formes, des combinaisons lemmes-catégories, et toute combinaison de forme, lemme, catégorie et traits (ces derniers pouvant être caractérisés par des expressions régulières). En outre, l'espace de cooccurrence est établi dynamiquement, au moment de la requête, par des expressions régulières définissant l'ensemble des relations à prendre en compte.

Pour répondre à la double exigence de flexibilité et d'efficacité, nous avons élaboré une indexation multi-niveaux, sous la forme de hachages de hachages sérialisés : chaque forme pointe vers l'ensemble des lemmes correspondants ; chaque lemme pointe vers l'ensemble de ses catégories possibles (dans le corpus) ; chaque lemme-catégorie pointe vers l'ensemble des traits associés (dans le corpus) ; chaque lemme-catégorie-traits pointe vers l'ensemble des relations associées ; chaque lemme-catégorie-traits-relation pointe vers un ensemble de paires (collocatif,fréquence). Les expressions régulières liées au contraintes portant sur les catégories, traits et relations sont appliquées lors du parcours de l'index. Les ensembles de catégories, traits et relations étant réduits (et fermés) et le temps de recherche dans le hachage étant en O(1), la succession de ces recherches n'est

pas très couteuse.

En ce qui concerne l'implémentation, nous avons opté pour le langage Perl, pour son traitement très efficace des expressions régulières. Pour les index, nous avons testé deux systèmes de bases de données réputés pour leur efficacité : BerkeleyDB 5.1.25[2] et KyotoCabinet 1.2.48[3]. Les résultats du tableau 3 montrent que le système qui est apparu le plus efficace pour nos requêtes était celui de KyotoCabinet::BTree.

Corpus presse (fr) 2007-2008 (87 807 463 tokens)	Taille des index	**test 2 1 pivot**	**test 2 5 pivots**	**test 3 24 pivots**
BerkeleyDB:Hash	1800 Mo	125 s. / 1,3 s.	275 s. / 206 s.	892 s. / 766 s.
KyotoCabinet::Hash	1200 Mo	50 s. / 1 s.	376 s. / 180 s.	749 s. / 702 s.
KyotoCabinet::Btree	955 Mo	76 s. / 1.5 s.	247 s./ 231 s.	416 s. / 315 s.

TABLEAU 2 : comparaison des tailles et des temps de réponse pour différents types de DBM (le 2ème temps est obtenu lorsqu'une requête est immédiatement réitérée).

Ces temps sont donnés à titre de comparaison : ils ont été obtenus sur un PC ancien et assez lent. Sur notre matériel actuel (Intel Core2 Quad CPU Q9550 2.83GHz, avec 4Go de RAM) nous obtenons des temps environ 4 fois supérieurs. La différence importante entre le 1er et le 2ème temps indique que ce sont les accès disques qui pénalisent les traitements, car lorsque la DBM est en cache, la réponse est presque instantanée. En utilisant un disque SSD ultra-rapide, nous prévoyons d'améliorer les temps de réponse de manière drastique.

5 Prise en compte des pivots multimots

L'aspect exclusivement binaire des relations de dépendance directe peut aboutir à un rétrécissement du contexte des observations et faire manquer des phénomènes intéressants sur le plan phraséologique. Ces limitations empêchent notamment l'extraction automatique de séquences polylexicales à valeur d'unité minimale de sens (les « meaning units » selon Sinclair 2004), qui peuvent présenter une variabilité considérable sur le plan de l'expression.

Cependant, en ce qui concerne les « collocations lexicales », Tutin (2008) affirme que la plupart d'entre elles ont une structure binaire, même pour celles qui s'étendent à plus de deux éléments, car elles correspondent sémantiquement à une structure prédicat-argument : "*Collocations can be considered as predicate-argument structures, and as such, are prototypically binary associations, where the predicate is the collocate and the argument is the base. Most ternary (and over) collocations are merged collocations (collocational clusters) or recursive collocations.*"

Et en effet, de nombreux travaux dédiés à l'extraction de collocations étendues à plus de deux mots se basent en fait sur des modèles binaires, appliqués à deux éléments composés : collocation d'arbres syntaxiques (Charest et al., 2010), construction itérative

[2]http://www.oracle.com/technetwork/database/berkeleydb/overview/index.html

[3]http://fallabs.com/kyotocabinet/

de cooccurrence multimots à partir de cooccurrences binaires (Seretan et al., 2003), ou encore calcul de mesure d'association multimots en combinant des mesures à deux termes.

De la même manière, il est possible d'étendre notre architecture pour le calcul des lexicogrammes d'un pivot donné, en la généralisant à des configurations plus complexes : la solution consiste à définir le pivot non plus seulement à partir d'une forme prise isolément, mais comme *une forme associée à un certain contexte lexico-syntaxique*. Une fois déterminé ce contexte, il est possible de calculer le tableau de contingence comme précédemment, le pivot et son contexte formant en quelque sorte une nouvelle unité pour laquelle il est possible de calculer à la fois les fréquences de cooccurrence (en se basant sur les relations du pivot) et la fréquence marginale dans le corpus.

Pour l'écriture des contextes, nous utilisons le formalisme de méta-expressions régulières proposé par Kraif (2008). Par exemple, pour rechercher le pattern V+ DET(poss.) + admiration_N + POUR, nous définissons le contexte suivant :

pivot : #1= *admiration#N*
contexte : <l=son> <#1> && <#2> && <pour,#3>::(.*,#1,#2)(.*,#2,#3)

Le calcul est seulement un peu plus long à mettre en œuvre, car les pivots multimots n'étant pas connus a priori, il n'est pas possible de les indexer tels quels. Seuls les tokens (formes ou lemmes) composant le contexte, ainsi que les relations de dépendances entre deux tokens définis, sont indexés, ce qui permet de réduire significativement l'ensemble des phrases à analyser. Pour des expressions comportant plusieurs relations, comme c'est l'intersection des phrases indexées pour chaque relation qui est retenue, la recherche est plus rapide : en d'autres termes, plus un pivot complexe est long, plus sa recherche est rapide. Dans le tableau 3 ci-dessous, on constate que pour le contexte donné en exemple, la mesure du log-likelihood fait clairement ressortir les verbes *cacher* et *dissimuler*, qui correspondent tous deux à la même construction stéréotypée : *X ne pas cacher/dissimuler son admiration pour Y*.

l1	l2	f	f1	f2	N	loglike
admiration_N	cacher_V	4	14	527	544994	38,83
admiration_N	dissimuler_V	2	14	107	544994	22,70
admiration_N	proclamer_V	2	14	176	544994	20,70
admiration_N	exprimer_V	2	14	642	544994	15,53
admiration_N	redire_V	1	14	76	544994	10,57
admiration_N	manifester_V	1	14	193	544994	8,70
admiration_N	confier_V	1	14	1319	544994	4,91

TABLEAU 3 - extrait de lexicogramme pour le pivot multimot *son admiration pour* pris en tant qu'objet direct

Ainsi conçue, l'extraction des lexicogrammes pour les pivots multimots se veut surtout être un outil d'observation permettant aux utilisateurs, par complexification progressive, de mieux préciser le contexte des phénomènes qui les intéressent (comme ici en précisant la détermination ou la structure prépositionnelle).

Cette approche qui va du simple vers le complexe peut néanmoins, d'une certaine

manière, s'automatiser. Partant d'un pivot simple, on peut retenir ses collocatifs les plus saillants pour former de nouveaux pivots multimots. Et l'on peut réitérer l'opération de manière récursive sur les nouveaux pivots, jusqu'à une taille limite fixée arbitrairement. Nous avons implémenté ce processus jusqu'à une taille maximale de 5 mots, en ne retenant, à chaque itération, que les candidats à l'extension qui cooccurrent au moins 3 fois et pour lesquelles la valeur de loglike sont supérieure à 10. Ne sont retenus que les pivots multimots maximaux (de 5 mots) ou qui ne peuvent être étendus par un pivot multimot plus long.

Dans l'exemple ci-dessous, pour mieux cibler l'extraction autour du nom *admiration*, nous avons imposé que le premier collocatif soit issu de la relation d'objet direct (on trouve donc, pour commencer, un verbe). Voici les résultats obtenus, sans filtrage, pour les 3 verbes les plus saillants.

1 : précision_N qui_PRON forcer_V la_DET admiration_N
2 : précision_N forcer_V la_DET admiration_N
3 : vouer_V une_DET admiration_N sans_PREP borne_N
4 : vouer_V une_DET profond_A admiration_N
5 : vouer_V une_DET grand_A admiration_N
6 : il_PRON vouer_V une_DET grand_A admiration_N
7 : qui_PRON vouer_V une_DET admiration_N
8 : qui_PRON pas_ADV cacher_V son_PRON admiration_N
9 : ne_ADV pas_ADV cacher_V son_PRON admiration_N
10 : avoir_V cacher_V son_PRON admiration_N
11 : il_PRON pas_ADV cacher_V son_PRON admiration_N
12 : qui_PRON ne_ADV cacher_V pas_ADV admiration_N

Comme souvent dans les extractions d'expressions multimots, on trouve un ensemble d'expressions de natures diverses (collocations simples, collocations récursives, locutions, etc.), avec notamment des fragments incomplets d'expressions plus larges (cf. exemple. 2) ou des expressions qui agrègent des éléments de contexte non pertinent (cf. exemple 10, avec *avoir*). On obtient cependant, et ceci de façon assez précise, des constructions récurrentes et stéréotypées caractéristiques de la combinatoire du nom *admiration* pris en tant qu'objet.

6 Conclusion

Nous avons présenté un nouvel outil d'exploration de la combinatoire lexico-syntaxique, que nous avons baptisé le *lexicoscope*. Cet outil s'appuie sur un modèle de cooccurrence flexible permettant à l'utilisateur de définir lui même les unités qui l'intéressent (en combinant forme, lemme, catégorie et traits) ainsi que l'espace de cooccurrence visé (en précisant les relations de dépendance concernées). Le lexicoscope permet en outre d'effectuer des comparaisons des profils combinatoires, synthétisés sous la forme de lexicogrammes, et propose en sortie des visualisations du type AFC, MDS ou hClust.

Enfin, pour permettre à l'utilisateur de ne pas se limiter aux seules dépendances directes autour d'un pivot, nous avons ajouté la possibilité de définir des pivots multimots avec leurs contextes syntaxiques. Ce nouvel outil est actuellement à l'essai, dans le cadre des

observations contrastives effectuées pour le projet Emolex. L'interface sera accessible pour le grand public d'ici quelque mois (mais les corpus, qui sont soumis à des restrictions de droits d'auteur, ne pourront être diffusés dans leur intégralité). D'ici là, nous pourrons effectuer une analyse plus précise des possibilités offertes par le lexicoscope pour la comparaison des profils combinatoires de différents pivots, et en dégager une méthodologie d'observation adaptée.

7 Références

Aït-Mokhtar, S., Chanod, J.-P., Roux C. (2002) "Robustness beyond Shallowness: Incremental Deep Parsing", *Natural Language Engineering*, 8 :121-144.

Attardi, G., Dell'Orletta, F., Simi, M., Chanev, A., Ciaramita, M. (2007) "Multilingual Dependency Parsing and Domain Adaptation using DeSR", In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague.

Charest,S. Brunelle E., Fontaine J. (2010) Au-delà de la paire de mots : extraction de cooccurrences syntaxiques multilexémiques, *Actes de TALN 2010,* Montréal, juillet 2010

Evert, Stefan (2007). Corpora and collocations. in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Heiden S., Tournier M. (1998) Lexicométrie textuelle, sens et stratégie discursive, actes *I Simposio Internacional de Análisis del Discurso*, Madrid.

Hoey, M. (2005) : *Lexical Priming: A New Theory of Words and Language*, London, Routledge.

Kilgariff A.,Tugwell D. (2001) WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography, *Proc ACL workshop on COLLOCATION Computational Extraction Analysis and Exploitation*, Toulouse July 2001.

Kraif, O. (2008) Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, *JADT 2008*, PUL, 625-634, vol. 2.

Nivre, J., Boguslavsky, I. M., Iomdin, L. L. (2008) "Parsing the SYNTAGRUS Treebank of Russian", *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), Manchester, August 2008, p. 641–648.

Seretan V., Nerima L., Wehrli E. (2003). Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. *Proceedings of the Fourth International Conference on Recent Advances in NLP*, (RANLP-2003), 424–431.

Sinclair, John McH. (2004) *Trust the text : language, corpus and discourse*, London, Routledge.

Tapanainen, P., Järvinen, T. (1997) "A non-projective dependency parser", In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, p. 64-74.

Tutin A. (2008), For an extended definition of lexical collocations, *Proceedings of Euralex*, Barcelone 15-19 juillet 2008, Université Pompeu Fabra.