SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

Extraction de pivots complexes pour l'exploration de la combinatoire du lexique : une étude dans le champ des noms d'affect

Kraif, Olivier[1], & Tutin, Agnès[1], & Diwersy, Sascha[2]

1 Univ. Grenoble Alpes, LIDILEM, F-38040 Grenoble
{Olivier.Kraif et Agnes.Tutin}@u-grenoble3.fr

2 Romanisches Seminar - Universität zu Köln
sascha.diwersy@uni-koeln.de

Résumé

Cet article porte sur le développement d'une nouvelle approche pour l'exploration de la combinatoire lexico-syntaxique, en vue de la caractérisation des valeurs sémantiques des unités étudiées. Cette approche a été mise en œuvre à travers le développement d'un outil nommé EmoConc, permettant de d'étudier la combinatoire des pivots (ou mots pôles) visés à travers l'extraction de *lexicogrammes*, des matrices de cooccurents syntaxiques enregistrant les principaux collocatifs du pivot pour un ensemble de relations syntaxiques données. Nous montrons, à travers une étude dans le champ des noms d'affect, l'intérêt qu'il y a à considérer non pas des pivots isolés, mais des pivots *complexes*, associés à un sous-arbre syntaxique précisant leur environnement. Enfin, nous décrivons une méthode d'extraction automatique d'expressions polylexicales dérivée de cette notion de pivot complexe.

1 Introduction

Cette étude a été réalisée grâce à des corpus et des outils développés dans le cadre du projet franco-allemand Emolex, dont l'objectif est d'analyser d'un point de vue contrastif les valeurs sémantiques, les rôles discursifs et 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 en traductologie.

Nous présentons ici une approche visant à caractériser et catégoriser les collocatifs verbaux d'une certaine classe de noms (ici des noms d'affect) pris au travers d'une relation syntaxique préalablement fixée (ici la relation verbe - complément d'objet). Nous faisons l'hypothèse que les propriétés sémantiques des unités sont reflétées par leurs propriétés combinatoires, et notamment par les constructions préfabriquées typiques de ces unités, qui traduisent ce que Sinclair (1991) appelle le « principe de l'idiome » (par opposition au « principe du libre choix »).

Pour étudier le *profil combinatoire* (au sens de Blumenthal, 2006) des unités, nous avons élaboré un outil interrogeable en ligne (Kraif, Diwersy, 2012), permettant, pour un pivot donné, d'extraire l'ensemble de ses cooccurrents avec les valeurs de son tableau de contingence, ce que Tournier et Heiden (1998) nomment son lexicogramme. Mais à la différence de ces derniers, nous ne retenons ici que les cooccurrences syntaxiques (par exemple un verbe et son objet direct), et non les cooccurrences de surface : les cooccurrences syntaxiques présentent en effet l'intérêt de réduire à la fois le bruit et le silence (Evert, 2008 ; Seretan 2010), les cooccurrents pertinents pouvant se situer à une distance arbitraire dans la phrase, au delà d'une fenêtre dont la largeur est fixée à priori.

La première partie de cet article est consacrée à la présentation de l'outil et de ses fonctionnalités. Dans un second temps, nous présentons une étude de cas, autour des verbes liés à la verbalisation des émotions (*hurler sa joie*, *confier sa honte, ...*), destinée à explorer les potentialités de l'extraction des

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801362

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

lexicogrammes. Nous montrons ainsi que la prise en compte de « pivots complexes », permettant de définir un ensemble de contraintes lexicales et syntaxiques autour du mot pivot placé au centre d'une requête (et donc des collocatifs) permet de mieux circonscrire, sur un plan sémantique, le champ des pivots étudiés. Nous examinons enfin une série de constructions plus générales issues de notre méthode d'extraction d'expressions polylexicales, basée sur l'expansion itérative des pivots complexes, et observons dans quelle mesure ces configurations permettent de mieux circonscrire les propriétés sémantiques des unités lexicales. Nous concluons sur des prolongements possibles de notre méthodologie.

2 EmoConc : un outil flexible pour l'observation de la combinatoire

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). Notons que pour nous le terme *pivot* (synonyme de l'expression *mot pôle* utilisé en textométrie) ne désigne pas une classe d'unité lexicales spécifiques, mais uniquement l'entrée lexicale - quelle qu'elle soit - placée au centre de l'observation, autour de laquelle on extrait par exemple des concordances ou des lexicogrammes.

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 ces 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 cotextes, et de combiner les aspects lexicaux et syntaxiques dans ses observations. Nous proposons ainsi trois modèles de cooccurrence :

- *cooccurrence lexico-syntaxique* : à l'instar du modèle mis en œuvre dans le Sketch Engine de Kilgariff et Tugwell (2001), ou des travaux de Charest et al. (2010) pour le dictionnaire Antidote RX, il s'agit de s'appuyer sur les relations fonctionnelles (du type sujet, complément d'objet, modifieur, etc.) identifiées entre deux unités. Evert (2008), signale l'intérêt de ce type de cooccurrence en termes 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*".
- *colligation syntaxique* : cette fois-ci, on s'intéresse à la récurrence d'un pivot pris dans une relation de dépendance donnée - par exemple le nom *admiration* pris en tant qu'objet direct.
- *colligation textuelle* : afin d'intégrer une dimension discursive, EmoConc permet d'extraire les statistiques d'occurrence d'un certain pivot par rapport à ses positions (début, milieu, fin) dans les textes - ou au sein des paragraphes - et de mesurer une éventuelle attirance de ce pivot par rapport à telle ou telle position (Novakova, Sorba, 2013).

Pour la cooccurrence lexico-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 (Attardi et al. 2007) pour le russe, basé sur un modèle stochastique créé à partir du corpus arboré SyntagRus (Nivre *et al.,* 2008).

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 pour un sous-ensemble de relations de dépendance précises (p.ex. épithète antéposée et postposée, modifieurs du nom, ...), et calculer les mesures d'associations spécifiques à cet espace.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

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 *souci* toutes relations confondues ?) ou beaucoup plus spécifiques et circonscrits (par exemple : quels sont les principaux collocatifs verbaux du nom *souci* au pluriel en tant qu'objet direct ?). Le tableau 1 montre l'extraction d'un lexicogramme pour ce dernier cas (le corpus compte environ 125 millions de mots, dont 110 millions issus d'articles de presse et 15 millions issus de textes littéraires contemporains[1]).

l1	**l2**	**f**	**f1**	**f2**	**loglike**
souci_N_msc:pl	avoir_V	243	475	423602	720,1006
souci_N_msc:pl	causer_V	24	475	2210	195,9398
souci_N_msc:pl	poser_V	28	475	15537	129,7117
souci_N_msc:pl	connaître_V	28	475	35189	86,5004
souci_N_msc:pl	oublier_V	20	475	13273	85,4552
souci_N_msc:pl	régler_V	11	475	4353	57,7889
souci_N_msc:pl	partager_V	9	475	8543	32,1681
souci_N_msc:pl	confier_V	7	475	10563	19,0419
...	...	...	...	...	...

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

2.1 Comparaison des lexicogrammes

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 (CAH). La figure 1 montre ces sorties pour des unités du domaine sémantique de la 'colère' (obtenues grâce à des modules du projet 'GNU R'[2]). La classification hiérarchique, réalisée pour la relation « "objet »", indique une hiérarchisation assez bien corrélée à l'intensité du sentiment. Quant à l'AFC, réalisée pour des relations quelconques concernant des collocatifs adjectivaux, elle permet de distinguer trois groupes :
    - *révolte*, *indignation* associés à des adjectifs relationnels tels que *populaire*, *estudiantin* qui suggèrent une manifestation collective liée à la sphère publique et politique. Ces noms sont par ailleurs en relation avec l'adjectif évaluatif *légitime* qui caractérise le bien fondé de cette émotion en tant qu'expression publique.
    - *fureur*, *rage*, *colère* - sémantiquement liés à l'expression ponctuelle et plus ou moins intense de l'affect. Étonnamment, les adjectifs qui caractérisent le mieux ces manifestations « "explosives »" de l'émotion expriment souvent le contrôle et la retenue : *rentré*, *froid*, *contenu, sourd* pour *rage* et *colère*, ce trait sémantique pouvant peut-être expliquer également le collocatif *noir*, que l'on trouve pour *colère* et *fureur*. Le contrôle étant le corrélat antagoniste de la perte de contrôle, on trouve également les adjectifs exprimant la violence et l'intensité : *fou* et *destructeur* pour *rage*, *grand* et *meurtrier* pour *fureur*, *gros*,

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

immense, *violent*, *vif* et *terrible* pour *colère*. Enfin, la colère seule se caractérise par des adjectifs évaluatifs qui la rapproche du précédent groupe : *sain*, *légitime*, *juste*.

- enfin, *énervement*, *irritation*, *exaspération* - qui concernent plutôt des états émotionnels précurseurs de cette manifestation. Ce dernier groupe est plus spécifiquement associé à des adjectifs évoquant le caractère intermédiaire ou modéré de l'affect, comme *certain* ou *réel* (utilisé dans un sens concessif) et l'aspect progressif, avec *croissant*, *grandissant*.

Ces exemples illustrent de façon éclairante le lien entre les valeurs sémantiques et la combinatoire lexico-syntaxique, et montrent comment une représentation géométrique de cette combinatoire peut servir de guide à une analyse plus fine sur le plan linguistique, dans une perspective heuristique.

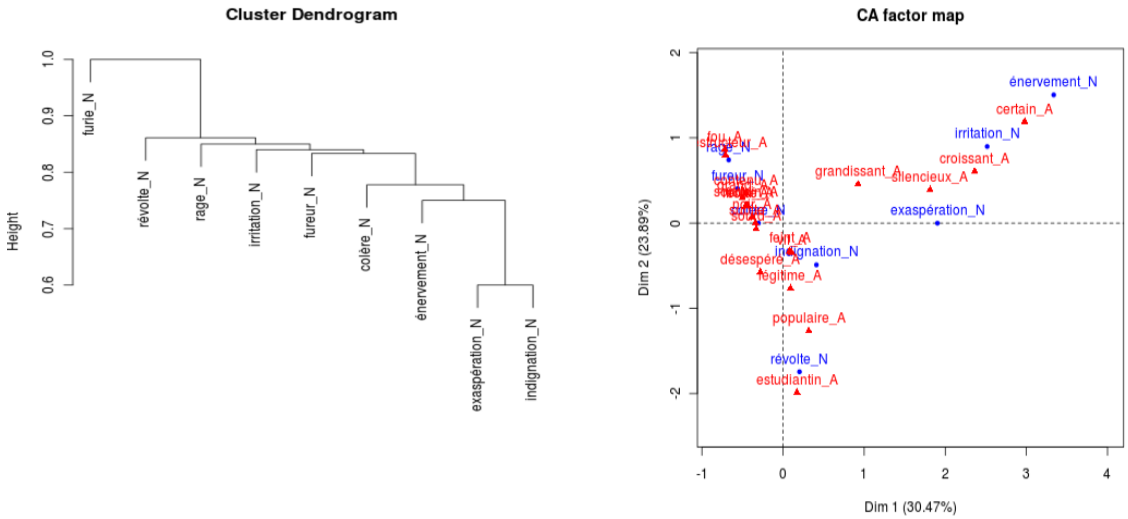

Figure 1 : Classification hiérarchique et AFC (domaine sémantique de la 'colère')

2.2 Prise en compte des *pivots complexes*

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 « units of meaning » 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."*

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 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*

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

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 *avouer* + DET(poss.) + N, nous définissons le contexte suivant :

pivot : #1= avouer_V
contexte : <c=N,#2> && <l=son,#3>::(obj,1,2)(det,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 dépendance 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 deux expressions récurrentes : *avouer son impuissance* et *avouer son admiration.*

l1	l2	f	f1	f2	loglike
avouer_V	impuissance_N	10	226	2868	142,0125
avouer_V	admiration_N	9	226	4016	119,8055
avouer_V	crime_N	6	226	26464	52,3355
avouer_V	peur_N	6	226	28357	51,5103
avouer_V	faute_N	5	226	15441	47,1415
avouer_V	goût_N	5	226	25267	42,2369
avouer_V	participation_N	5	226	28769	40,9463

Tableau 2 - extrait de lexicogramme pour le pivot complexe *avouer son + N*

Ainsi conçue, l'extraction des lexicogrammes pour les pivots complexes 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). Par exemple, le corpus nous permet de constater que dans la plupart des cas, l'expression *avouer son admiration* attend la réalisation d'un troisième actant, le plus souvent introduit par la préposition *pour*.

2.3 Extraction automatique d'expressions polylexicales

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 complexes. Et l'on peut réitérer l'opération de manière récursive sur les nouveaux pivots, jusqu'à une taille limite fixée arbitrairement. La figure 2 montre comment un sous-arbre récurrent a été extrait pour identifier, de façon totalement automatique, l'expression *vouer une admiration sans borne*.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

Figure 2 - Extraction itérative d'une expression complexe (*vouer une admiration sans borne*)

Nous avons effectué une telle extraction pour le pivot *colère* pris en tant qu'objet direct, en ne retenant que les collocatifs obtenant au moins 5 pour le loglike, et d'une fréquence de cooccurrence au moins égale à 3. On obtient la liste des expressions ci-dessous (partiellement lemmatisées et regroupées), qui constitue un « instantané » assez riche illustrant la combinatoire du pivot étudié :

provoquer la/une colère
provoquer la colère des syndicats/du président/du gouvernement
l'annonce avait provoqué colère
susciter la colère d'une partie
susciter la colère des associations
pour exprimer leur/sa colère
exprimer sa/leur/une colère
avoir exprimé hier colère
déclencher la colère
piquer une/des colère/s
apaiser la colère
tenter d'apaiser la colère
pour apaiser la colère
calmer la colère
tenter de calmer la colère
pour calmer la colère
calmer sa colère
attiser la colère
laisser éclater sa colère
manifester sa/leur colère
pour manifester leur colère
venir manifester leur colère
ne pas cacher sa/leur colère
crier sa/leur colère
ravaler sa colère
ruminer sa colère
contenir sa/la colère
avoir du mal à contenir colère
déchaîner la colère

Tableau 3 - Liste des expressions polylexicales extraites pour *colère* (obj)

3 Les pivots complexes pour caractériser le sens des noms d'affect

Après ce tour d'horizon des fonctionnalités de notre outil, nous proposons d'illustrer notre méthodologie d'observation par des observations sémantiques associées à des patrons lexico-syntaxiques liés à des pivots complexes.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

3.1 Utilisation des pivots complexes pour caractériser les noms d'affect : l'exemple des collocatifs de verbalisation

Note objectif est ici d'observer dans quelle mesure l'utilisation de pivots complexes permet d'affiner l'étude sémantique des noms d'affect. Plusieurs études ont montré que les différentes dimensions sémantiques dégagées par les collocatifs (Buvet *et al.* 2005 *;* Tutin *et al.* 2006) permettent de caractériser cette classe sémantique : collocatifs exprimant différents types de marques aspectuelles (*un accès de colère*), la manifestation extérieure et physique (*bondir de joie, rayonner de bonheur*), le contrôle (*dompter sa peur, dissimuler sa joie*), la causativité (*susciter* l'admiration), la verbalisation (*hurler sa joie, confier son chagrin*).

Parmi ces dimensions, la verbalisation a retenu notre attention pour cette petite étude de cas. Elle est principalement associée à deux fonctions sémantiques : l'accent peut être mis sur l'expression du sentiment (*crier, hurler sa joie/sa colère*) ou sur la fonction communicative (*avouer/confier son chagrin/peine*). Nous avons observé dans les corpus que les collocatifs apparaissaient dans des environnements lexico-syntaxiques plus contraints qu'une simple association verbe-nom : d'une part, le nom d'émotion est généralement introduit par un déterminant possessif coréférentiel au sujet ; d'autre part, le destinataire de la verbalisation est souvent mentionné (*il lui confia sa joie ; il avoua à Marie sa déception*). Notre objectif a été alors de déterminer dans quelle mesure ces pivots complexes permettent de mieux caractériser la classe des noms d'affect.

Notre expérimentation a consisté à comparer les classes distributionnelles extraites par trois types de configurations :

1. les verbes de verbalisation associés à un complément direct nominal

2. les verbes de verbalisation associés à un complément direct nominal déterminé par un possessif

3. les verbes de verbalisation associés 1) à un complément direct nominal déterminé par un possessif, 2) à un complément datif

La figure 3 ci-dessous résume ces configurations lexico-syntaxiques.

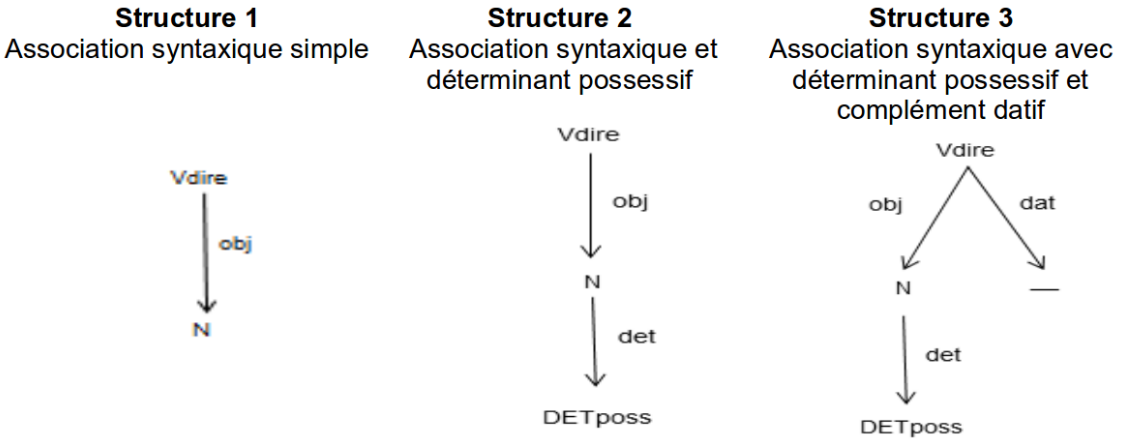

Figure 3 – Configurations syntaxiques explorées

Sept verbes de verbalisation ont été explorés (*avouer, clamer, confier, crier, dire, exprimer, hurler*) , sur la partie française du corpus Emobase (cf. *supra*). Seules les associations apparaissant au moins 5 fois avec une mesure de log-like de 10,83 ont été sélectionnées. Nous avons observé dans quelle mesure la configuration permettait d'isoler les noms relevant de notre classe sémantique. Par exemple, pour

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

l'extraction effectuée avec le verbe *avouer* (Cf. tableau 4), seule l'association avec *admiration* a été retenue.

Collocatif de verbalisation	Pivot
avouer	meurtre
avouer	crime
avouer	infanticide
avouer	vérité
avouer	admiration
avouer	ignorance
avouer	vol

Tableau 4 – Associations avec le verbe *avouer*

- Associations avec les pivots simples (V –obj ->N) : *confier ... crainte*

Les associations extraites avec les pivots simples apparaissent dans le tableau 5 ci-dessous. On observe une grande variabilité entre verbes de verbalisation. La dimension verbalisation n'apparaît spécifique de notre champ qu'avec un sous-ensemble de verbes (*exprimer, hurler*). Le verbe *dire* quant à lui apparaît souvent employé comme introducteur de discours direct dans cette configuration et n'est guère employé en cooccurrence avec des noms d'affect.

Collocatifs de verbalisation	Nombre total de noms associés	Noms d'affect	% de noms d'affect
avouer	35	13	37
clamer	10	2	20
confier	143	5	3,5
crier	37	11	29,5
dire	45	3	6,5
exprimer	135	76	56
hurler	9	5	55,5
Total	414	115	27,5

Tableau 5 : Extractions avec des noms d'affect avec les pivots simples

- Associations avec les pivots complexes à possessif (V -obj->N<-Det-Det_poss) : *hurler sa joie*

Les extractions exploitant la contrainte du déterminant possessif donnent de meilleurs résultats (Cf. tableau 6). Plus de la moitié des configurations observées concernent un nom d'affect. La structure est extrêmement caractéristique pour certains verbes comme *hurler* ou *crier* qui semblent se spécialiser dans cet emploi sémantique avec cette structure comme on le verra plus bas avec des configurations comme *ne pas cacher* Det_poss. La configuration permet également de filtrer des verbes comme *dire* ou *confier* pour lesquels la simple association syntaxique n'apparaît pas suffisante. Enfin, plus intéressant encore, la méthode opère un meilleur rappel des noms d'affect puisqu'elle extrait davantage de noms de ce type que les associations syntaxiques simples.

- Les associations avec les pivots complexes (possessifs et datifs) : N/Pro <-dat- V -obj-> N ->det-Det_poss. *Lui confier sa joie*

Nous avons encore affiné la configuration en ajoutant une contrainte sur un deuxième complément au datif (information présente dans les annotations de sorties de l'analyseur Connexor) (Tableau 7). Ici, la structure apparaît finalement trop contrainte. Elle n'extrait au final que 12 noms d'affect, soit 10 fois

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

moins que la configuration précédente. Pour 4 verbes sur 7, on n'observe aucune extraction. Par ailleurs, elle s'avère moins caractéristique des noms d'affect que la structure précédente tout en étant toutefois un peu plus caractéristique que la structure à simple association syntaxique.

Collocatifs de verbalisation	**Nombre total de noms associés**	**Noms d'affect**	**% de noms d'affect**
avouer	16	9	53
clamer	6	2	33
confier	40	12	30
crier	10	8	80
dire	57	29	51
exprimer	97	64	66
hurler	6	5	83
Total	232	125	55,5

Tableau 6 : Extractions avec des noms d'affect avec les déterminants possessifs

Collocatifs de verbalisation	**Nombre total de noms associés**	**noms d'affect**	**% de noms d'affect**
avouer	0	0	0
clamer	0	0	0
confier	16	4	25
crier	0	0	0
dire	14	6	42
exprimer	3	2	66,5
hurler	0	0	0
Total	33	12	36,5

Tableau 7 : Extractions avec des noms d'affect avec les déterminants possessifs et le datif

- Synthèse

Au terme de cette petite expérimentation (qu'il faudrait bien entendu étendre sur d'autres dimensions sémantiques), plusieurs éléments se dégagent. Ajouter une contrainte spécifique sur le type de détermination apparaît tout à fait pertinent. Cela accroît sensiblement d'une part le rappel de la classe sémantique visée, et d'autre part, la précision de l'extraction effectuée. En revanche, des contraintes trop strictes apparaissent contre-productives, en tout cas sur le corpus considéré, dans la mesure où le rappel des expressions se révèle ici très fortement affecté. Il faudrait encore établir si l'utilisation d'un très grand corpus (d'un milliard de mots, par exemple) confirme les résultats observés.

Ces résultats nous semblent avoir une incidence importante sur les études de sémantique utilisant les critères distributionnels ou les outils comme le SketchEngine qui n'utilisent à notre connaissance que la relation lexico-syntaxique. S'il est avéré que l'utilisation de relations lexico-syntaxiques apparaissait plus efficace que le simple environnement distributionnel pour effectuer des regroupements sémantiques pertinents (Grefenstette 1996), il convient peut-être maintenant d'aller plus loin en exploitant des pivots complexes d'une granularité adaptée, comme le type de détermination, ainsi que cela a été observé dans notre petite étude.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

3.2 Des pivots complexes aux constructions polylexicales

Nous avons par ailleurs effectué plusieurs types de sondage avec notre méthode sur le corpus des noms d'affect. En effectuant l'extraction automatique des expressions polylexicales sur une cinquantaine de noms d'affect, nous avons noté que certaines expressions correspondaient à des schémas génériques très répandus pour l'ensemble de ces noms.

Par exemple, sur nos 39 pivots ayant suffisamment d'occurrences dans le corpus pour avoir permis d'extraire des expressions polylexicales, 15 ont été identifiés dans la construction **ne pas cacher + Det_poss + N** (avec les seuils de significativité que nous avons imposés i.e. un nombre d'occurrences supérieur ou égal à 3 et un loglike supérieur à 5).

Cette construction semble donc assez générale dans ce champ sémantique. Si réciproquement, en partant de cette construction prise comme pivot complexe, on cherche tous les collocatifs nominaux en position d'objet direct, dans la même démarche que celle effectuée plus haut, alors on trouve non seulement une grande variété de noms d'affect, mais ces noms sont presque *tous* des noms d'affect (nous avons souligné les deux seuls intrus) :

inquiétude, satisfaction, déception, admiration, ambition, joie, intention, agacement, scepticisme,sympathie, amertume, volonté, préférence, colère, intérêt, pessimisme, embarras, hostilité, irritation, enthousiasme, désir, exaspération, fierté, mécontentement, impatience, émotion, étonnement, souhait, soulagement, mépris, aversion, crainte, désarroi, jubilation, perplexité, plaisir, bonheur, réticence, préoccupation, envie, réserve, goût, doute, espoir, jeu

On a donc trouvé une *construction*, dont les unités prises isolément ont peu à voir avec le sémantisme des affects, mais dont la cooccurrence avec les noms d'affect montre une grande spécialisation sémantique. Ce type de cooccurrence évoque ce que Stefanowitsch & Gries (2003) nomment des *collostructions*.

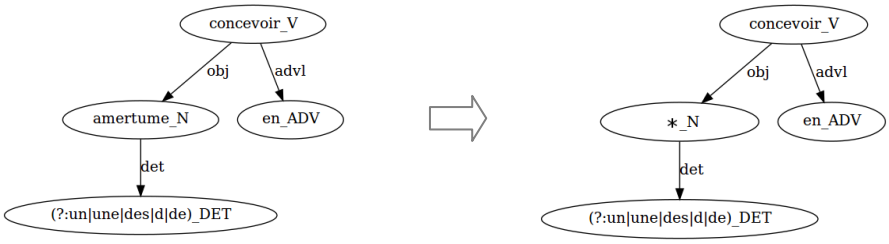

Figure 4 - Généralisation d'une expression polylexicale

Pour vérifier si d'autres constructions pouvaient déboucher sur ce type de paradigme, nous avons opéré une généralisation à partir des expressions polylexicales issues de l'extraction automatique, en recherchant tous les noms apparaissant dans le même contexte. Par exemple, à partir de l'expression *en concevoir une amertume*, correspondant au sous-arbre de la figure 4, on considère le pivot complexe obtenu en substituant *amertume* par un nom quelconque et on cherche tous les collocatifs nominaux qui entrent en cooccurrence avec ce pivot.

Pour cet exemple précis, on obtient un paradigme assez restreint, et assez homogène sur le plan sémantique :

***en concevoir une/un + N** : [amertume, chagrin, déception]*

D'autres expressions issues de nos extraction ont permis d'obtenir des résultats comparables :

pour éviter une/un nouveau + N : *[désillusion,déconvenue,dérapage, crise]*

exprimer son/sa + N à l'égard : *[déception, défiance]*

pour calmer le/la + N : *[colère, jeu, esprit, grogne, tension, surchauffe, ardeur, inquiétude, mécontentement, fronde, ire, impatience, douleur, crainte, prix, monde, crise]*

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

ne pas cacher son/sa + N de voir : *[déception, satisfaction, souhait, espoir]*

exprimer son/sa + N de voir : *[souhait, déception, satisfaction, désir]*

laisser éclater sa + N : *[joie, colère]*

Toutes ces constructions affichent une nette attirance pour des noms d'affect (nous avons souligné les intrus) et présentent chacune des traits sémantiques particuliers rendus manifestes par ces différents paradigmes (aspect ponctuel, polarité, attente, etc.).

Enfin, on trouve également des expressions très récurrentes mais caractéristiques d'un nom en particulier, et typique du genre textuel. Par exemple toutes les occurrences ci-dessous ont été identifiées sur la partie journalistique de notre corpus (le nombre d'occurrences figure entre parenthèse) :

provoquer la colère des syndicats (16)
provoquer la colère de l'opposition (5)
provoquer la colère des habitants (4)
provoquer la colère d'une partie (4)
provoquer la colère des salariés (4)
provoquer la colère des autorités (4)
provoquer la colère du gouvernement (4)
provoquer la colère du président (4)

Certaines expressions stéréotypées affichent encore un degré de spécialisation supérieur, comme **pour ne pas connaître une désillusion**, qui n'apparaît dans notre corpus que dans les articles sportifs :

Mais une fois l'ennui d'un match à zéro essai digéré, il faut reconnaître que, face à ces Argentins toujours aussi pénibles et embrouilleurs, il valait mieux remiser ambitions offensives et grain de folie ***pour ne pas connaître une*** *nouvelle et cruelle* ***désillusion****.*

Finalement, le Stade Rennais n'aurait- il pas intérêt à gérer au mieux sa fin de saison tout en s'activant à préparer la suivante ***pour ne pas connaître une*** *nouvelle grande* ***désillusion****?*

Mais les hommes d'Oswald Tanchot devront se montrer prudents ***pour ne pas connaître une*** *réelle* ***désillusion****, devant une équipe qui a, selon le coach vitréen,« un fort potentiel offensif.*

4 Conclusion et perspectives

Les pivots complexes, dans notre système, permettent de définir tout un environnement lexico-syntaxique sous la forme d'un sous-arbre de dépendance. Les observations effectuées sur un corpus de grande dimension (plus de 120 millions de mots) montrent que la possibilité d'étudier et de manipuler des pivots complexes dans un système de concordance ouvre des perspectives intéressantes pour l'étude de la combinatoire :

- d'une part, la simple caractérisation de l'environnement syntaxique et fonctionnel (déterminants, adverbes, prépositions, etc.) de certaines paires de cooccurrents permet de mieux cibler les emplois des mots pleins : c'est ce que nous avons montré avec les verbes de verbalisation des émotions, dont les occurrences sont fortement marquées par la présence du déterminant possessif devant le nom.

- d'autre part, la possibilité d'extraire automatiquement des expressions polylexicales, en s'appuyant sur l'expansion itérative des pivots complexes, permet d'identifier à la fois des constructions génériques mais caractéristiques d'un certain champ sémantique, telle que *ne pas cacher son + N*, et des expressions pré-construites, stéréotypiques dans un certain genre de texte (*pour ne pas connaître une désillusion*).

Afin d'aller plus loin dans l'extraction de ce type d'expression préfabriquées, nous prévoyons d'améliorer nos outils, en nous intéressant à l'extraction de sous-arbres plus abstraits : par exemple, dans les occurrences de *pour ne pas connaître une désillusion*, on remarque que *désillusion* est toujours modifié par un adjectif. Il faudrait que notre système puisse identifier l'expression complète, malgré ses parties

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801362

variables : *pour ne pas connaître une ADJ désillusion.* Nous chercherons ainsi à repérer des patterns récurrents plus abstraits, mais très productifs, et relevant également de ce que Sinclair appelle le « principe de l'idiome ».

Enfin, une piste de recherche intéressante est la caractérisation du genre textuel à partir des expressions polylexicales automatiquement extraites. Il semblerait en effet que les formules routinisées qui émergent de nos extractions puissent constituer un marqueur fiable de l'inscription d'un texte dans un genre textuel précis : nous chercherons dans de futures études à vérifier cette hypothèse.

5 Références bibliographiques

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.

Buvet, P.A., Girardin, Ch., Gross G., Groud C. (2005). « Les prédicats d'<affect> », *Lidil*, 32 | 2005, p. 123-143.

Charest,S. Brunelle E., Fontaine J. (2010). Au-delà de la paire de mots : extraction de cooccurrences syntaxiques multilexémiques, *Actes de TALN 2010,* Montréal, 19-23 juillet 2010.

Evert, Stefan (2008). Corpora and collocations. in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter : Berlin.

Grefenstette, G. (1996). Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. In Boguraev, B. & Pustejovsky, J. (eds). *Corpus Processing for Lexical Acquisition*, 205-216. Cambridge, Massachusset : MIT Press.

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.

Husson, F., Josse, J., Lê, S. (2008). "FactoMineR: An R Package for Multivariate Analysis", *Journal of Statistical Software*, 25(1): 1-18.

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.

Kraif, O., Diwersy, S. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques, *Actes de la conférence TALN 2012*, Grenoble, p. 399-406

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), p. 424–431.

Seretan ,V. (2010). *Syntax-based collocation extraction*. Springer.

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

Stefanowitsch, A. & Gries S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2:209-43.

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.

Tutin, A., Novakova, I., Grossmann, F., Cavalla, C. (2006). « Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires », *Langue française* 2/2006, n° 150, p. 32-49.

---

[1] Le corpus Emobase est accessible sur le site : http://emolex.u-grenoble3.fr/emoBase.

[2] Nous avons mis en oeuvre l'AFC au moyen du package *FactoMineR* (Husson, Josse & Lê 2008 ; http://factominer.free.fr/index_fr.html), et le CAH au moyen du module *hclust* qui fait partie du package *stats* de R.