Extraction d'expressions polylexicales sur corpus arboré

Julien Corman

Extraction d'expressions polylexicales sur corpus arboré

Nom : CORMAN
Prénom : Julien

UFR Langage, lettres et arts du spectacle, information et communication

---

Mémoire de master 2 recherche - 30 crédits

Spécialité : Industries de la Langue

Sous la direction d'Agnès Tutin et Olivier Kraif

Année universitaire 2011-2012

Extraction d'expressions polylexicales sur corpus arboré

Nom : CORMAN
Prénom : Julien

UFR Langage, lettres et arts du spectacle, information et communication

---

Mémoire de master 2 recherche - 30 crédits

Spécialité : Industries de la Langue

Sous la direction d'Agnès Tutin et Olivier Kraif

Année universitaire 2011-2012

Remerciements

Merci à mes encadrants et relecteurs, *pour leurs conseils avisés*, et aux évaluateurs, pour *s'être prêtés au jeu*.

Table des matières

INTRODUCTION

LE LIDILEM

Le Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles (LIDILEM) a été créé en 1987, et est rattaché à l'Université Stendhal-Grenoble 3. Ses activités s'organisent principalement autour de trois axes de recherche :

- description linguistique, corpus et TAL,
- sociolinguistique et acquisition du langage,
- didactique des langues et recherches en ingénierie éducative.

Ce travail s'inscrit dans le premier de ces trois axes, qui s'intéresse depuis plusieurs années à la phraséologie, tant pour les aspects théoriques que pour les aspects applicatifs (en particulier informatiques et didactiques). La question de la phraséologie a principalement été abordée à travers la modélisation des collocations (Tutin et Grossmann, 2002 ; Grossmann et Tutin, 2003 ; Tutin, 2010), des expérimentations monolingues et multilingues sur les écrits scientifiques dans le cadre du projet ANR Scientext (Kraif et Tutin 2011), et l'extraction (Kraif et Diwersy 2012) et la modélisation des expressions polylexicales dans le cadre du projet ANR Emolex. Le travail développé ici tire profit de corpus analysés syntaxiquement, et prolonge certains travaux réalisés au LIDILEM, basés sur des techniques plus classiques, comme les automates à états finis, et sur des associations essentiellement binaires pour l'extraction de collocations.

OBJECTIFS

Ce mémoire s'intéresse aux mots et expressions composés, comme *disque dur* ou *rendre un service*, à leur recensement à partir de larges corpus de textes numérisés, et à leur prise en compte par des applications manipulant du texte libre, comme les moteurs de recherche, ou les traducteurs automatiques. La non reconnaissance de ces mots et expressions est régulièrement pointée comme un problème central en traitement automatique des langues (TAL). L'enjeu pour l'analyse, la traduction, voire la génération de texte peut être abordé principalement sous deux angles. Ils correspondent intuitivement à deux situations auxquelles un individu apprenant une langue étrangère, ou bien découvrant un domaine technique ou scientifique, a déjà été confronté.

La première situation est le contresens ou l'incompréhension face à une expression dont la signification n'est pas littérale : il est par exemple assez rare dans un texte que *jeter l'éponge* désigne un lancer d'ustensile ménager, ou que *sac d'os* désigne réellement un sac rempli d'os. Pour un système devant interpréter le sens d'une phrase, ou extraire des informations d'un texte, ou encore classer des textes selon leurs contenus, il est donc essentiel de savoir que *se rendre compte, porter ses fruits, trou noir* ou *au quart de tour* ne doivent pas être analysés littéralement.

La deuxième situation, plus courante encore, peut être paraphrasée par « on ne dit pas X mais Y », alors qu'aucune règle claire ne permet d'expliquer cette préférence. On ne dit pas **essai d'assassinat* mais *tentative d'assassinat.* On ne dit pas *faire* ni **donner un service,* mais *rendre un service* (on dit en revanche *faire une faveur*). On ne dit pas **chemin ferré* ni **chemin ferroviaire* mais *chemin de fer,* alors qu'on dit *voie ferrée* ou *voie ferroviaire* mais pas **voie de fer.* Et on préférera dire *en bonne voie/chemin* plutôt que **en bonne route,* mais *se mettre en chemin/route* plutôt que **se mettre en voie.* Quelques minutes passées à chercher des cas similaires dans sa langue maternelle suffisent en général à mesurer l'ampleur du phénomène. En traduction automatique, ne pas prendre en compte ces contraintes conduit à la production de tournures difficilement compréhensibles, ou peu crédibles, comme **faire une décision* ou **lui dire un sermon.* Mais ces associations privilégiées sont aussi une aide précieuse en analyse pour la désambiguïsation, un autre problème central en TAL.

Plus concrètement, un lexique enrichi de mots et expressions composés, s'il est utilisé dans un cadre applicatif, nous semble devoir posséder au minimum les deux qualités suivantes.

- Il doit permettre la reconnaissance des mot ou expressions composés présents dans des textes similaires à ceux du corpus ayant servi à l'extraction, y compris lorsque ces expressions sont sujettes à des variations de forme, d'ordre des mots ou de distance entre les mots (par exemple *il lui a fait tout un cirque* et *le cirque incroyable qu'elle leur a fait*).
- Il doit distinguer les mots ou expressions composés dont le sens justifie une entrée distincte, et ceux devant être rattachés à un ou plusieurs de leurs mot(s) et expression(s) inclus(es). Si l'on se représente le lexique d'une application comme un dictionnaire, il nous semble par exemple peu intéressant, du point de vue du sens, de rattacher l'expression *passer un savon* à l'entrée du mot *passer,* ni à celle du mot *savon.* Un lien étymologique et métaphorique peut bien entendu être retrouvé, mais, dans une perspective applicative, ce rapprochement serait surtout source d'erreur. De même, il nous semble en pratique plus pertinent de réserver une entrée autonome à *abominable homme des neiges* ou *ne pas avoir froid aux yeux.* Ce n'est pas nécessaire en revanche pour *prendre des proportions alarmantes,* ou *affligeante banalité,* qui sont relativement transparents*,* ni pour *jouer à guichets fermés,* à condition que l'on ait déjà identifié *à guichets fermés* comme une expression autonome.

Or ces deux conditions sont loin d'être remplies dans la plupart des systèmes actuels. Seule une petite proportion de ces mots et expressions est prise en compte, et leur description est souvent insuffisante. De plus leur recensement n'est que semi-automatisé, nécessitant un lourd travail manuel de validation et d'encodage, travail réalisé soit par des linguistes ou lexicographes, soit par des spécialistes d'un domaine (droit, médecine, …) pour les nombreuses expressions à valeur terminologique. Ils sont aidés en amont d'outils d'extraction automatique, mais cette extraction est encore très imparfaite. Le travail présenté ici se veut donc une étape supplémentaire vers l'automatisation de ce processus d'extraction et d'encodage.

Une des limites actuelles de ces outils est la difficulté à extraire des constructions de plus de deux mots pleins, comme *passer un coup de fil, avoir le vent en poupe* ou *faire partie des heureux élus,* en particulier lorsque les mots ne sont pas nécessairement contigus, ou que leur ordre peut varier. Nous avons pour cela développé une méthode d'extraction basée sur l'analyse syntaxique préalable du corpus utilisé, qui permet de traiter efficacement ces phénomènes. Une approche spécifique a également été adoptée pour les constructions imbriquées, afin par exemple d'extraire *coup d'œil* et *jeter un coup d'œil*, mais pas **jeter un coup.*

Par ailleurs, une attention particulière a été portée à la caractérisation des expressions extraites, en termes notamment d'ordre et de distance observés entre les mots, flexions privilégiées, présence ou absence de déterminants, etc. Cette description fine facilite la reconnaissance de

ces expressions sans ambiguïté dans d'autres textes que ceux ayant servi à l'extraction, et sans que ceux-ci aient nécessairement fait l'objet d'une analyse syntaxique. Il est ainsi possible de distinguer *se rendre compte* et *rendre des comptes*, ou *tenir la jambe* et *tenir par les jambes*, mais aussi d'identifier deux occurrences de la même expression dans *il entretenait depuis peu une relation amoureuse* et *la relation amoureuse qu'ils entretiennent*.

La question de l'organisation des expressions extraites en fonction de leurs sens (entre elles, et par rapport aux mots qu'elles incluent) n'a en revanche pas été directement abordée. Elle constitue un des prolongements évidents de ce travail, et l'observation des sorties laisse entrevoir plusieurs pistes allant dans cette direction, dont une sera présentée à la fin de ce document.

PLAN

La première partie propose tout d'abord un bref aperçu de la notion linguistique de *polylexicalité* (les « mots et expressions composés » ci-dessus), les propriétés remarquables de ces constructions, et quelques enjeux afférents (lexicographiques, didactiques et surtout applicatifs). Sont ensuite abordées les principales techniques développées depuis une vingtaine d'années pour l'extraction automatique de mots ou expressions composés. Ces techniques sont assez hétérogènes, au regard notamment des propriétés discriminantes choisies, reflétant directement l'absence de consensus linguistique autour de la notion de polylexicalité.

La seconde partie, plus algorithmique, décrit l'outil développé. Les trois phases de l'extraction y sont présentées dans l'ordre chronologique des traitements : tout d'abord l'extraction des constructions candidates, sur une base syntaxique et non linéaire, puis leur filtrage, avec notamment un traitement spécifique des constructions imbriquées, et enfin une phase plus originale de caractérisation morphosyntaxique des constructions retenues.

La troisième partie propose une première évaluation de cette méthode, en comparant les sorties à celles d'un outil d'extraction basé sur des séquences de mots récurrentes. Une analyse qualitative plus fine des expressions extraites permet ensuite de proposer quelques pistes d'utilisation et d'approfondissement.

1 L'EXTRACTION AUTOMATIQUE D'EXPRESSIONS POYLEXICALES

1.1 ASPECTS LINGUISTIQUES

1.1.1 Les expressions polylexicales

Le terme générique *expression polylexicale* (*EPL* dans ce qui suit), et son équivalent anglais *multi-word expression*, se sont récemment imposés dans le cadre TAL pour désigner l'ensemble des constructions relevant du domaine linguistique de la *phraséologie :* collocations, idiomes, locutions, constructions à verbe support... La gamme des phénomènes est variée, et, malgré un relatif consensus autour de certaines catégories, il n'existe pas de typologie faisant l'unanimité, ni même de délimitation claire de l'ensemble des constructions concernées.

Une définition en intension est également délicate, ce qui complexifie les traitements automatiques, qui ne peuvent s'appuyer sur une description ou un formalisme aboutis. Une hypothèse de travail minimale est malgré tout nécessaire ici, afin d'orienter le travail d'extraction, mais également d'évaluer la méthode appliquée sans tautologie.

1.1.1.1 Les principaux courants de la phraséologie

Granger et Paquot (2008) distinguent deux traditions phraséologiques. L'approche dite classique est issue de Bailly, puis des travaux de Vinogradov et Amosova, une de ses variantes les plus influentes étant formalisée dans le cadre de la Théorie Sens-Texte (TST), dont le traitement des phrasèmes est présenté dans (Mel'čuk, 2011). Le champs de la phraséologie est celui des « constructions multilexémiques non libres » (Mel'čuk, *Ibid*), par opposition donc aux associations de mots dites « libres » (hors contraintes morphosyntaxiques ou sémantiques traditionnelles), comme *acheter un gâteau* ou *une chemise usée.* Ces travaux sont largement guidés par des considérations typologiques : selon les auteurs, différents critères (et tests correspondants) permettent de distinguer différentes catégories de phrasèmes, soit de façon catégorielle, soit sur un ou des continuums. Parmi les critères de classification les plus employés, on peut citer la non-compositionnalité sémantique (abordée en 1.1.2.1), ou encore la distinction entre unités syntagmatiques et unités pragmatiques (ces dernières étant équivalentes à une phrase, comme *De rien* ou *Tel est pris qui croyait prendre*). La catégorie des collocations binaires, constituées d'une base et d'un collocatif contraint, et dont le sens est transparent, comme *commettre un crime* (et non **faire un crime*) a notamment fait l'objet dans le cadre de la TST d'une formalisation sémantique poussée, via une trentaine de *fonctions lexicales* syntagmatiques.

L'autre grande tradition phraséologique, plus récente, est l'approche dite « néo-firthienne », qui s'est surtout développée suite aux travaux de Sinclair, et attribue un rôle central au travail sur corpus. La notion de phraséologie y est essentielle (*the idiom principle*), mais dans une acception plus large, incluant des contraintes de co-restriction non seulement léxèmiques, mais également morphologiques, syntaxiques ou sémantiques. Parmi les types de constructions qui ont ainsi enrichi le champ phraséologique, on peut citer les patrons lexico-syntaxiques (comme *tous* N *confondus*), voire lexico-sémantiques (comme *faire la* [adjectif à connotation négative] *expérience de*), qui ont donné lieu à la publication d'un volume dédié du *Cobuild Dictionnary* (décrit dans Hunston et Francis, 2000).

1.1.1.2 Hypothèse de travail par défaut

Une première hypothèse de travail, qui nous semble raisonnable dans le cadre d'une tâche d'extraction automatique, consiste à adopter comme critère de validation la co-restriction léxèmique, selon l'acception classique, mais sans adopter de typologie *a priori*, suivant en cela l'approche néo-firthienne. Est alors admis comme EPL tout ensemble de lemmes cooccurrents, et liés entre eux par une ou plusieurs préférences proprement lexicales, en plus des restrictions morphosyntaxiques régulières (qui interdisent par exemple une séquence comme **très viande cuits*), ou de cohérence sémantique (qui interdisent par exemple une séquence comme **une équation de fierté pâle*). C'est par exemple le cas d'associations de mots dont le sens est transparent, mais consacrées par l'usage. Il est difficile de substituer *juste* à *équitable* dans *commerce équitable* ou *essai* à *tentative* dans *tentative d'assassinat.* C'est aussi le cas d'expressions dont le sens est plus métaphorique. Ainsi dans *jeter l'éponge, jeter* ne peut pas être remplacé par *lancer,* ni *éponge* par *serpillière,* sans altérer le sens du propos, et cette préférence est presque impossible à formaliser en terme de traits sémantiques, l'explication étant ici étymologique. Entre ces deux extrêmes, on trouve quantité d'expressions dont une partie seulement des mots pleins ont un sens métaphorique, et qui obéissent au même principe de préférence lexicale : *découvrir le pot aux roses* et non **trouver le pot aux roses* ni **découvrir le pot aux begonias*, *une politesse exquise* et non *une politesse délicieuse*...

Cette hypothèse autorise cependant, pour une EPL extraite sur la base de préférences lexicales, la mise en évidence (manuelle ou automatique) de préférences morphosyntaxiques complémentaires, comme des flexions et fonctions syntaxiques privilégiées, ou encore, pour une expression à base verbale, la non passivation ou la non relativisation. La partie 2.4 décrit un traitement automatique allant en ce sens.

1.1.2 Propriétés remarquables

Voici quelques propriétés remarquables associées à de nombreuses EPL, quoique non systématiques. Elles ont été employées en linguistique pour distinguer différentes catégories de phrasèmes. Ce sont toutes des déclinaisons de la notion de *figement*, dont la corrélation avec la phraséologie est généralement admise, mais qui reste trop vague pour être opératoire. La liste qui suit n'est pas exhaustive, mais les propriétés choisies expliquent en grande partie les enjeux pratiques liés à l'extraction d'EPL, qui sont détaillés en 1.2. Ces propriétés ont par ailleurs été exploitées par certains travaux d'extraction en tant que critères discriminants, les techniques correspondantes étant présentées en 1.3.2.

1.1.2.1 Non compositionnalité sémantique

La non compositionnalité sémantique désigne le fait que le sens d'une expression ne soit pas totalement déductible de celui de ses parties. Elle est admise comme un des traits saillants de nombreuses EPL, dont des noms composés (*fer de lance, coup de grâce, ...*) ou adverbes complexes (*au quart de tour, à toute allure...*), voire même une condition nécessaire (mais non suffisante) pour les idiomes (*avoir le vent en poupe, jeter l'éponge, casser les pieds ...*). Ce trait reste cependant largement subjectif : des expressions aussi diverses que *salle de bain, donner des sueurs froides, mémoire vive* ou *à guichets fermés* pourront être perçues comme non compositionnelles par certains individus, et compositionnelles par d'autres, en fonction notamment de leur âge, de leur connaissance du domaine ou de leur rapport à la langue (native ou seconde). Par ailleurs, de très nombreuses constructions pouvant être assimilées à des EPL (dont les collocations au sens de Mel'čuk) sont manifestement compositionnelles, comme *d'une affligeante banalité, un port altier* ou *faire un petit signe amical.*

1.1.2.2 Non substituabilité paradigmatique

La substitution d'un mot par un de ses (quasi-)synonymes, voire hyperonymes ou co-hyponymes, abondamment pratiquée en phraséologie, est probablement un des tests intuitifs les plus opératoires (divers exemples de substitutions impossibles, comme **essai d'assassinat,* ont déjà été donnés en 1.1.1.2). Mel'čuk (2011) utilise par exemple ce test comme preuve d'une restriction de sélection lexicale. Pour les mots polysémiques, on utilise évidemment les synonymes du sens actualisé : par exemple, dans *en tirer les conséquences* on essaiera de substituer *retirer* ou *extraire* à *tirer,* mais pas *traîner*. Cette distinction entre synonymes est cependant souvent difficile à établir en pratique (par exemple décider si *prendre* fait partie ou pas de l'ensemble formé par *retirer, extraire,* et *tirer*). Surtout, il n'existe pas toujours de synonyme suffisamment proche du sens actualisé : pour *équation différentielle* par exemple, il est difficile de trouver un synonyme d'*équation* ou de *différentiel* qui puisse être utilisé.

Plusieurs autres raisons rendent difficile une application rigoureuse de ce test. Tout d'abord, un mot peut être considéré comme contraint soit lorsqu'aucun de ses synonymes ne peut lui être substitué, soit lorsqu'au moins un de ses synonymes ne peut lui être substitué. Dans le premier cas, il faut tester tous les synonymes de *tentative* (*essai, effort, velléité, ...*), et pas uniquement *essai*, avant de pouvoir conclure que *tentative d'assassinat* est une EPL. Mais il faut alors admettre par exemple que *voie* dans *en (si) bonne voie* n'est pas contraint, puisqu'il peut être remplacé par *chemin.* Dans le second cas au contraire, il suffit de montrer que *voie* ne peut pas être remplacé par *route* pour prouver qu'il est contraint. D'après Cowie (1998), cette seconde hypothèse n'est pas incompatible avec l'approche phraséologique classique : Vinogradov notamment admettait comme phrasèmes des constructions autorisant quelques substitutions paradigmatiques (*rompre/briser la glace*, mais pas **casser la glace*). Cette substituabilité restreinte est cependant difficilement modélisable de façon discrète, et une approche statistique, comme celle présentée en 1.3.2.1, devient alors nécessaire. Elle permet en outre de tenir compte de la fréquence (ou rareté) des mots que l'on cherche à substituer.

Un problème de circularité peut également se présenter, selon la façon dont est défini un (quasi-)synonyme. Par exemple *couler* est considéré comme un synonyme de *vivre* dans le *Trésor de la Langue Française Informatisé* (TLFI), mais cette synonymie est elle-même due à leur substituabilité dans des constructions que beaucoup assimileraient à des EPL, comme *vivre*/*couler des jours heureux*/*paisibles.* Le test s'en trouve alors invalidé.

Enfin un test discret peine à rendre compte de préférences de sélection lexicale marquées, mais qui n'excluent pas totalement la substitution par un synonyme. On trouve par exemple sur le Web des occurrences de *politesse délicieuse*.

1.1.2.3 Comportement morphosyntaxique idiosyncrasique

Cette propriété est mise en avant dans Sag et al. (2002) comme un des traits caractéristiques de nombreuses EPL, et un champ d'investigation important pour le TAL. Les travaux sur corpus de Moon notamment (dans Cowie, 1998) ont mis en évidence un figement morphosyntaxique plus fréquent parmi les EPL que parmi des combinaisons de mots non contraintes lexicalement. *Porter ses fruits* par exemple n'est ni passivable, ni transformable en relative, tandis que le verbe n'admet que la troisième personne, et l'objet uniquement le pluriel, ce dernier ne pouvant prendre de modifieur.

Cette propriété est cependant loin d'être systématique, quantité d'EPL (comme *rendre un service*) étant régulières. Surtout, elle recouvre des phénomènes très hétérogènes : irrégularité de la structure syntaxique (*à la va vite*), flexions interdites, alternances interdites, modifieur interdit, etc. Enfin c'est une propriété partagée par de nombreux mots simples (*fiançailles* par exemple n'admet que le pluriel).

1.2 ENJEUX

1.2.1 Enjeux lexicographiques

1.2.1.1 Dictionnaires de langue

La lexicographie moderne a été largement renouvelée grâce au travail sur corpus, dont le TLF est un des premiers aboutissements marquants. Plus récemment, les travaux de l'école néo-firthienne ont conduit à la publication du *COBUILD dictionary* (ou encore plus récemment du *MacMillan english dictionary*), mais ont également influencé la constitution de dictionnaires préexistants (*Oxford, Longman*, …).

La disponibilité de corpus numérisés et d'outils élaborés pour les interroger est un des facteurs essentiels de ce renouvellement des pratiques. Des techniques d'extractions d'unités polylexicales proches de celle présentée ici sont notablement utilisées en amont du travail des lexicographes. Ainsi le *Sketch Engine* (Kilgariff, 2004) se base sur des cooccurrences syntaxiques remarquables pour proposer un premier aperçu du profil combinatoire d'un mot, et est utilisé pour l'enrichissement de plusieurs dictionnaires anglophones. Charest et al. (2007 et 2010) ont utilisé une technique similaire pour la constitution d'*Antidote RX*, un dictionnaire des cooccurrences du français.

Cette approche lexicographique reste cependant largement centrée sur le mot (et son contexte), à l'exception notable des patrons lexico-syntaxiques (Hunston et Francis, 2000), présentés en 1.1.1.1. Les EPL sont typiquement associées à un ou plusieurs des mots simples qui les constituent, et font encore rarement l'objet d'entrées distinctes, avec leurs propres contraintes morphosyntaxiques ou sémantiques, leurs propres collocatifs (*coup d'oeil/jeter un coup d'oeil, bout du tunnel*/*voir le bout du tunnel,* …), actants typiques (*suivre son cours*/*l'enquête suit son cours*,...), etc. Les rapports de synonymie entre EPL, ou entre EPL et mots simples, sont en revanche progressivement intégrés aux dictionnaires (*mémoire vive/mémoire système/mémoire volatile/RAM, les mains dans les poches/les doigts dans le nez, coup d'œil/regard, jouer à guichets fermés/faire salle comble*,...).

Mel'čuk (2011) propose lui un traitement distinct pour les phrasèmes non compositionnels, comme *casser les pieds* ou *un château en Espagne*, qui font l'objet dans le *Dictionnaire explicatif et combinatoire* d'une entrée propre (et sont décrits de façon aussi exhaustive que les mots simples), et les phrasèmes compositionnels, comme *proprement scandaleux,* rattachés à la base dont ils dépendent (ici à *scandaleux*). Ces principes n'ont malheureusement pas (encore) donné lieu à la publication de dictionnaires à large couverture. Surtout, aucune technique à

notre connaissance n'a encore été mise au point pour distinguer (semi)-automatiquement phrasèmes compositionnels et non compositionnels. Or c'est un enjeu évident pour l'analyse automatique (1.2.3.2).

1.2.1.2 Terminologies et ontologies

Plus encore que pour la langue dite générale, le vocabulaire des langues dites de spécialité s'enrichit très largement par composition lexicale (*trou noir, onde de choc...*), en particulier le vocabulaire professionnel ou des sciences appliquées (*assurance vie, circuit intégré, taux de change, tête de lecture...*), y compris en cas de traduction (*disque dur, carte mère...*), ces EPL venant elles-mêmes régulièrement enrichir la langue dite générale.

La constitution de terminologies est avant tout guidée par des besoins applicatifs, notamment pour des systèmes spécialisés basés sur l'analyse de texte brut, comme la veille informationnelle ou l'extraction d'information. Sont conjointement utilisés des listes de termes préexistantes, et des patrons morphosyntaxiques d'extraction (du type *N de N Adj*), généralement à base nominale (1.3.1.2).

Les ontologies permettent en outre d'organiser les concepts correspondant aux termes validés, prototypiquement en classes et sous-classes hiérarchisées. Elles autorisent ainsi des raisonnements sur les instances de ces concepts, nécessaires par exemple dans le cadre du dialogue homme-machine, mais sont aussi utilisées en recherche et en extraction d'information, ou pour structurer des métadonnées dans le cadre du Web sémantique. La plupart d'entre elles servent à modéliser un domaine précis, comme la *Gene Ontology* pour le domaine génétique, bien qu'il existe quelques thésauri « généralistes ». Enfin *Wordnet* (Fellbaum et Miller) constitue un cas particulier, entre ressource lexicale structurée et ontologie. Dans tous les cas, un important travail manuel de validation, spécification, et organisation des termes est nécessaire pour que ces ressources soient exploitables.

1.2.2 Enjeux didactiques

La polylexicalité est fréquemment pointée comme une des grandes difficultés de l'enseignement d'une langue étrangère, faute parfois de ressources lexicographiques adéquates. En situation de compréhension, la reconnaissance d'EPL non compositionelles sémantiquement est évidemment essentielle pour éviter les contresens. Le sens de certaines d'entre elles peut cependant parfois être inféré (*the power behind the throne*/*l'éminence grise*), éventuellement à partir du contexte. Une même EPL pourra alors être perçue comme (métaphoriquement) compositionnelle par un locuteur non natif, et non compositionnelle par un locuteur natif (*suivre son cours*, *salle de bain, pousser le bouchon trop loin, trouver un écho favorable, peser ses mots* ...).

Mais la maîtrise des associations de mots privilégiées d'une langue est surtout un enjeu en situation d'énonciation. Le choix du collocatif juste en particulier est une compétence difficile à acquérir, notamment lorsque le sens de l'expression est transparent (*une haute opinion* et non **une grande opinion, un visage impassible* et non **un visage imperturbable, …*). On parle alors d'*idiomaticité* du discours produit, quoique l'acception soit assez différente de celle d'*idiome* dans l'approche phraséologique classique (il y désigne un phrasème non compositionnel sémantiquement, généralement à base verbale). Cette idiomaticité est également sujette à des variations dialectales.

1.2.3 Enjeux applicatifs

On présente régulièrement le traitement adéquat des EPL comme une marge de progression importante pour le TAL. Elles sont en réalité déjà présentes dans les lexiques de nombreuses applications, mais avec une modélisation parfois insuffisante, et moyennant un travail de validation manuelle plus ou moins important et/ou le recours à des listes préexistantes de termes ou expressions.

1.2.3.1 Fréquence

Les estimations sur la fréquence des EPL en corpus sont très variables, selon l'approche linguistique adoptée, ou la méthode utilisée pour les identifier. Comme le notent Granger et Paquot (2008), l'école néo-firthienne a élargi la gamme des phénomènes concernés à de nombreuses contraintes syntagmatiques considérées jusque là comme extérieures au domaine de la phraséologie (les *colligations* de Hoey, qui désignent des préférences en terme de fonction syntaxique, en sont un bon exemple).

En langue dite générale, on estime que le nombre d'unités polylexicales justifiant sémantiquement une entrée propre, comme les idiomes (*casser les pieds*), noms composés (*pomme de terre*) ou adverbes composés (*à toute allure*) est très supérieur au nombre de mots simples. Maurice Gross (d'après Green et al., 2011) a par exemple évalué à plus de 5 000 le nombre d'adverbes composés du français, contre 1 500 adverbes simples. Cette disproportion est évidemment beaucoup plus marquée si l'on ajoute les collocations.

1.2.3.2 Analyse automatique

La reconnaissance d'un nombre plus ou moins important d'EPL, via un lexique et/ou des patrons, est une condition indispensable à la plupart des applications basées sur l'analyse automatique de texte brut : recherche d'information, veille, indexation, classification, extraction, résumé… Même dans le cadre d'une analyse de surface, la reconnaissance d'entités nommées

polylexicales constitue un minimum. Plus généralement, pour l'analyse du sens d'un texte, un traitement adéquat des EPL non compositionnelles (*essuyer les plâtres, jeter un coup d'œil, un dos d'âne...*) ou partiellement compositionnelles (*filer le parfait amour, avoir une peur bleue,...*) permet d'éviter un nombre important de contresens.

L'intérêt des EPL compositionnelles peut sembler moins évident. Pour l'analyse du sens d'un énoncé, il est *a priori* inutile de savoir que *verre* a pour collocatif *briser* plutôt que *casser* (il suffit de connaître le sens de *briser* et celui de *verre*). Cette information peut constituer un indice de désambiguïsation sémantique (*briser* employé avec *verre* n'a pas le même sens que dans *briser tous ses espoirs* ou *briser une grève,* ou encore *blessé* avec *grièvement* a son sens physique premier, et non son sens psychologique), mais il s'agit ici de cas classiques de désambiguïsation sur critères distributionnels, comme pour *introduire*/*entrer dans*/*écrire une pièce*, qui ne relèvent pas spécifiquement du domaine de la phraséologie.

En revanche les EPL compositionnelles comme non-compositionnelles sont pertinentes pour l'analyse syntaxique, plus en amont des traitements. La reconnaissance d'une EPL grâce à un patron lexicalisé (syntaxique ou simplement linéaire, selon les parseurs), est un indice fort de désambiguïsation. Par exemple l'identification de l'EPL *pomme de terre* permet, dans *pomme de terre cuite*, de rattacher *cuite* à *pomme* et non à *terre,* contrairement à *assiette de terre cuite.* Des cas de figure plus complexes sont bien entendu possibles, comme dans la phrase suivante :

> *Le Crédit lyonnais (LCL)* ***réunit*** *un comité central d'entreprise (CCE)* ***en séance extraordinaire*** *le 1er juin dans le cadre d'une « préinformation » sur un « plan de sauvegarde de l'emploi ».*

L'identification de l'EPL *réunir en séance extraordinaire* permet non seulement de rattacher *séance* à *réunit* et non à *comité* ou à *entreprise*, mais offre aussi un « noyau dur » non ambigu de relations et étiquettes syntaxiques pour l'analyse du reste de la phrase.

Enfin les EPL « agrammaticales » (*à la va vite, à bras le corps, un en avant,...*) que Sag et al. (2002) nomment *fixed expressions,* gagnent à être identifiées très en amont des traitements, comme le sont déjà par exemple les locutions prépositionnelles les plus courantes (*à partir de, avant de...*). Elles sont parfois exocentriques, c'est-à-dire que la catégorie du syntagme ne peut être déduite de celle de sa tête (*en veux-tu en voilà* par exemple se comporte parfois comme un adjectif). Un cas plus complexe est celui des EPL exocentriques, mais dont la structure syntaxique interne est régulière. C'est le cas en français de nombreux adverbes complexes à base nominale (*les yeux fermés, la bouche pleine...*). Un mode de description spécifique peut alors être adopté, comme celui utilisé pour le *French Treebank*, décrit entre autres dans Green et al. (2011).

1.2.3.3 Traduction

L'analyse ou la traduction « mot à mot » d'EPL non compositionnelles sémantiquement (c'est-à-dire dont le sens du tout est difficilement déductible de celui des parties, comme *bande dessinée* ou *porter ses fruits*) est un des écueils les plus célèbres de la traduction automatique, de par son côté ludique (*it's raining cats and dogs*/**il pleut des chats et des chiens, to spill the beans/*renverser les haricots, ...*). Mais le phénomène le plus délicat à traiter est probablement celui des EPL compositionnelles, beaucoup plus nombreuses, et surtout très difficiles à inventorier, qui peuvent conduire à la production d'*anti-collocations*, définies par Pearce (2001) comme une combinaison de mots qu'un locuteur natif n'utiliserait pas, par exemple **faire un service*, ou **demander une question*. Elle ne nuisent pas toujours à la compréhension de l'énoncé, mais au minimum à la fluidité de la lecture.

1.2.3.4 Génération

C'est ici la non substituabilité paradigmatique (1.1.2.2) qui oblige à prendre en compte la polylexicalité, afin d'éviter la génération d'anti-collocations, notamment lorsque la base d'une EPL n'admet pas les mêmes collocatifs que ses synonymes, antonymes ou hypo/hyperonymes (selon l'organisation du lexique utilisé) : **dire un sermon, *essuyer une victoire,...* Le comportement morphosyntaxique irrégulier d'une EPL peut lui aussi être source d'erreurs de génération : au même titre que les mots simples (mais de façon beaucoup plus fréquente), les EPL font l'objet de restrictions d'emploi, comme la non passivation (**les pieds lui ont été cassés par,...*).

1.3 TECHNIQUES D'EXTRACTION

Seretan (2008 et 2011) propose un aperçu assez complet des expériences menées depuis plus d'une vingtaine d'années sur l'extraction automatique d'EPL, bien que ses propres recherches portent plus spécifiquement sur l'extraction de collocations. Les techniques se sont affinées, bénéficiant des travaux antérieurs, mais également de la fiabilité des prétraitements linguistiques appliqués aux corpus. La tâche est cependant loin d'être résolue. Nous proposons ici un aperçu plus synthétique des méthodes existantes, en les distinguant sur deux axes : la définition des constructions candidates, elle-même conditionnée par le prétraitement du corpus, et la ou les propriété(s) discriminante(s) utilisée(s) pour distinguer une EPL d'une construction candidate quelconque.

1.3.1 Constructions candidates

1.3.1.1 Unités lexicales

La lemmatisation du corpus via un étiquetage morphosyntaxique désambiguïsé est aujourd'hui une étape préalable très largement répandue. Elle correspond au traitement lexicographique habituel (*jettent l'éponge*, *a jeté l'éponge* et *jeter l'éponge* sont normalement considérés comme trois occurrences de la même expression), quoiqu'elle puisse conduire très ponctuellement à des regroupements abusifs *(faire une avance* et *faire des avances*). Elle permet aussi de limiter l'éparpillement des occurrences (une des principales difficultés de l'extraction d'EPL, présentée en 1.4.1.1). Certains travaux utilisent cependant volontairement les formes de surface : c'est notamment le cas des expériences menées par Dias et al. (2000a) sur des corpus anglais, français et portugais.

1.3.1.2 Cooccurrence linéaire ou syntaxique

Les travaux de Church et Hanks (1991) utilisaient (entre autres) comme critère de coocurrence de deux mots leur apparition conjointe dans une fenêtre linéaire de 5 tokens. Cette approche permet d'extraire des couples *a priori* éligibles, comme *honorary doctor*, mais aussi d'autres comme *doctor* et *hospital,* qui peuvent être significatifs sémantiquement, mais échappent au champs phraséologique classique, de par l'absence de relation syntaxique directe. Une première solution, utilisée par exemple par Smadja (1993) ou Dias et al. (2000b) consiste à observer les variations de distance entre mots cooccurrents : s'il apparaissent suffisamment régulièrement à

une distance fixe (par exemple à une distance de +2 tokens), cette cooccurrence est supposée motivée syntaxiquement. Une limite pratique de ces approches est l'importante combinatoire générée (1.4.1.2), en particulier si l'on cherche à extraire des EPL de plus de deux mots.

Une alternative moins coûteuse consiste à n'accepter comme candidats que des séquences de mots contigus, aux prix cependant d'une perte en rappel. Cette approche est encore très utilisée dans le cadre de l'extraction de terminologies, quoique sous une forme plus élaborée (on peut par exemple en trouver un aperçu dans Aubin et Hamon, 2006). Les terminologies sont essentiellement composées de syntagmes à base nominale assez rigides (**trou très noir*...), et obéissant souvent à des patrons morphosyntaxiques réguliers (N Prep N, N N, ...). Les patrons les plus productifs peuvent alors être utilisés en tant que critères d'extraction des candidats, moyennant un *chunking* préalable (délimitation des syntagmes nominaux maximaux). Afin d'améliorer le rappel, une certaine souplesse peut être introduite (épithète optionnelle...), ainsi que des traitements particuliers pour les syntagmes nominaux imbriqués (comme *émissions de gaz à effet de serre*).

Le recours à une véritable analyse syntaxique s'est fait plus tardivement, et de façon progressive, du *chunking* à l'analyse en dépendances ou par constituants, parallèlement aux progrès des analyseurs. Une description détaillée de cette évolution peut être trouvée dans Seretan (2008 et 2011). Une construction candidate devient alors un ensemble de mots explicitement liés par des relations syntaxiques (objet, épithète, …).

Enfin Green et al. (2011) ont récemment combiné patrons et analyse syntaxique dans le cadre d'une extraction par apprentissage supervisé, à l'aide d'un parseur statistique, à partir du *French Treebank.* Les EPL du corpus d'apprentissage sont annotées syntaxiquement en suivant le Lexique-Grammaire de Maurice Gross : une EPL a une étiquette spécifique en tant que syntagme, mais est constituée d'une séquence plate (éventuellement disjointe) de catégories. Par exemple le syntagme *tirer la sonnette d'alarme* y a pour structure syntaxique le syntagme plat [$_{\text{MWE\_V}}$ V Det N Prep N].

1.3.1.3 Filtres catégoriels

Lors de cette phase d'extraction des constructions candidates, la très grande majorité des travaux appliquent également des filtres catégoriels prédéfinis, du type Adj+N ou V+N. C'est évidemment le cas des approches basées sur des patrons morphosyntaxiques, mais aussi de la plupart des travaux recourant à la syntaxe (parfois sans requérir d'ordre d'apparition des catégories), à l'exception notable de l'expérience décrite dans (Martens et Vandeghinste, 2011). L'utilisation de ces filtres à des fins typologiques est discutable (on leur préfère généralement les propriétés présentées en 1.1.2), mais ils permettent en revanche une réduction importante du

bruit sur les sorties, via l'exclusion par exemple de mots appartenant à des classes fermées (déterminants, prépositions,...). Ces filtres peuvent en outre être interprétés de façon plate ou récursive. Seretan (2008) remarque cependant que les filtres catégoriels utilisés pour l'extraction de collocations varient largement d'une méthode à l'autre, et que les choix en la matière semblent assez arbitraires.

1.3.2 Critère(s) d'identification

Les trois propriétés remarquables décrites en 1.1.2 ont déjà été ponctuellement utilisées afin d'extraire des EPL sur corpus. Cependant le critère le plus fréquemment appliqué reste de très loin la cooccurrence significative des mots, quelle que soit la façon dont sont définis un mot (forme de surface ou lemme-catégorie) et une cooccurrence (linéaire contiguë, linéaire non contiguë ou syntaxique).

1.3.2.1 Non substituabilité paradigmatique

Les méthodes basées sur la non substituabilité paradigmatique peuvent s'appuyer soit sur une base de synonymes extraite du corpus étudié (Lin, 1999 ; van de Cruys et Villada Moiron, 2007), soit sur une ressource lexicale externe, comme *Wordnet* pour Pearce (2001), ou la base de synonymes extraite par Lin pour Fazly (2007). Il s'agit dans tous les cas d'identifier une préférence marquée, au sein d'une construction, pour un terme plutôt que pour un de ses quasi-synonymes, voire hyperonymes. Par exemple, *cadavre exquis* peut être considéré comme une association caractéristique de par l'absence (ou rareté) dans le corpus de **cadavre délicieux* et **corps exquis* (alors que *corps* et *délicieux* sont respectivement plus fréquents que *cadavre* et *exquis*).

La question de la fiabilité des tests de substitution paradigmatique a déjà été abordée en 1.1.2.2. Les approches statistiques offrent cependant plusieurs avantages sur un test intuitif discret. Elles permettent tout d'abord de modéliser des préférences (pour un des synonymes) plutôt que la simple non substituabilité. Cette préférence peut elle-même être pondérée par la fréquence marginale du lemme (toutes choses égales par ailleurs, une préférence pour le verbe *procurer* aura plus de poids qu'une préférence pour le verbe *donner*). Enfin cette mesure peut encore être affinée, afin de modéliser une préférence pour quelques synonymes, autrement dit une substituabilité restreinte, comme *briser/rompre la glace,* mais pas **casser la glace* (elle se rapproche alors d'une mesure d'entropie, mais prenant en compte les fréquences marginales des synonymes et la taille du synset).

Cependant la fiabilité et la couverture de la ressource lexicale utilisée limitent l'efficacité de cette technique, comme le note Baron (2007) à propos de *Wordnet*. Les bases de synonymes extraites automatiquement peuvent quant à elles justement souffrir de la polylexicalité, si elles sont construites sur des critères distributionnels : il est par exemple possible que *fil* et *téléphone* soient reconnus comme des synonymes, car apparaissant tous deux dans *passer*/*donner*/*recevoir un coup de fil*/*téléphone*, ce qui invalide en retour (du moins partiellement) le test de non substituabilité dans ces EPL. Enfin une autre limite pratique est la nécessaire constitution préalable de synsets (ou cliques, ou équivalents), et surtout la sélection automatique du bon synset lors des tests de substitution (essayer de substituer *chambre* à *pièce* dans *écrire une pièce* est par exemple inapproprié).

1.3.2.2 Non-compositionnalité traductionnelle

Cette propriété a été exploitée en tant que critère d'extraction d'EPL, via l'exploitation de corpus alignés. La méthode consiste à identifier des EPL non compositionnelles à partir d'écarts de traduction : si la traduction d'un ensemble de mots cooccurrents diffère de la combinaison de leurs traductions prototypiques respectives (dans le corpus), alors cet ensemble est considéré comme non compositionnel. Par exemple, *not playing with a full deck* ne se traduit pas par *ne joue pas avec une main complète (*ni *avec un pont complet, ...*). Melamed (1997) procède par comparaison de modèles de traduction, en cherchant des constructions ayant une traduction propre, mais récurrente. Par exemple l'EPL ci-dessus sera extraite si elle est régulièrement traduite par (*avoir*) *une case en moins*. Villada Moiron et Tiedemann (2006) supposent au contraire que la traduction d'une EPL non compositionnelle est plus volontiers sujette à variations qu'une construction dont le sens est littéral.

Cette approche présente cependant une limite importante : les résultats de l'extraction sont fortement dépendants des langues utilisées (et des choix de traduction). De très nombreuses EPL sont par exemple traduites littéralement ou quasi littéralement du français vers l'anglais (*prendre son temps, faire sens, le bénéfice du doute, prendre un risque, mettre tous ses œufs dans le même panier, le jeu n'en vaut pas la chandelle, un parachute doré, être dans le rouge* …).

1.3.2.3 Comportement morphologique ou flexionnel idiosyncrasique

Comme il a déjà été mentionné en 1.1.2.3, cette propriété est difficilement exploitable seule pour deux raisons : elle est loin d'être systématique pour une EPL, ni même propre aux EPL, ce qui nuit à la qualité de l'extraction, et elle recouvre des phénomènes très hétérogènes, ce qui la rend difficile à modéliser sans apprentissage supervisé.

Evert, Heid et Spranger (2004), montrent ainsi qu'il est possible d'extraire des couples *Adj+Nom* remarquables en allemand à partir de simples préférences flexionnelles, mais préconisent d'utiliser cet indice soit en complément d'un autre critère d'extraction, soit après extraction, à des fins de caractérisation. Fazly (2007) évalue le degré de figement morphosyntaxique de constructions à verbe support, en mesurant leur préférence pour des patrons obtenus en combinant trois critères (passivation/non passivation, type de déterminant et nombre). Cette mesure est utilisée pour identifier des EPL, mais en complément d'une mesure de cooccurrence plus classique (l'information mutuelle spécifique).

1.3.2.4 Cooccurrence remarquable

La cooccurrence significative des mots de l'expression reste de très loin le critère le plus largement utilisé pour l'extraction automatique ou semi-automatique d'EPL, en particulier de collocations. Même lorsqu'une des trois propriétés qui viennent d'être mentionnées sont utilisées, c'est en complément d'un indice de cooccurrence. La corrélation entre préférence lexicale (1.1.1.2) et cooccurrence remarquable semble *a priori* assez évidente, mais la définition d'une cooccurrence remarquable l'est beaucoup moins.

Un premier indice très simple est la fréquence relative d'un type de construction candidate (par exemple *disque dur*) parmi toutes les occurrences de constructions candidates (par exemple toutes les séquences *N Adj*, ou toutes les dépendances *N→Adj,* ou toutes les dépendances simples...). Cet indice est souvent reconnu comme étonnamment fiable (Manning et Schütze,1999), mais tend évidemment à surévaluer des associations peu significatives de mots très courants, comme *autre idée* ou *dernier jour,* qui auront probablement une fréquence plus élevée que *port altier* ou *affligeante banalité.*

L'approche la plus courante consiste alors à évaluer l'association mutuelle entre deux ou plusieurs mots, en prenant en compte à la fois leurs cooccurrences et leurs fréquences marginales respectives, afin de valoriser les cooccurrences de mots rares par ailleurs. De très nombreuses mesures ont été proposées pour cette tâche. On distingue parfois les mesures basées sur un test d'hypothèse (comme le *t-score*) des mesures issues de la théorie de l'information (comme l'information mutuelle spécifique). Le *log-likelihood ratio* proposé par Dunning (1993) est souvent considéré comme une des mesures d'association les plus robustes. Une étude détaillée des propriétés respectives de ces différentes mesures peut être trouvée dans Evert (2005). Pecina et Schlesinger (2006) ont recensé 82 mesures utilisées dans le cadre de l'extraction d'EPL, et en ont utilisé plusieurs de façon simultanée, pondérées par apprentissage

supervisé. Evert et Krenn (2001) ont cependant montré que la fiabilité d'une mesure d'association était variable selon le prétraitement des données, mais également selon le type d'expression visé.

La plupart de ces mesures d'association sont binaires. La cooccurrence significative de trois mots ou plus se calcule alors généralement par récursivité, entre un couple validé pris comme un tout et un troisième mot, ou bien elle peut être calculée en mesurant, pour chacun des mots d'une construction candidate, son score d'association binaire avec le reste de la construction pris comme un tout (Dias et al., 2000b).

1.4 LIMITES

Outre les problèmes de définition évoqués en 1.1, voici quelques-unes des limites pratiques actuelles de l'extraction automatique d'EPL.

1.4.1 Limites calculatoires

1.4.1.1 Dispersion des occurrences

Il n'existe pas à notre connaissance d'estimation fiable de la fréquence des EPL en corpus, ni de leur proportion parmi les unités (mono et poly)lexicales d'une langue ou d'un domaine (voir 1.2.3.1). De nombreux auteurs s'accordent en revanche pour reconnaître que l'extraction automatique d'EPL se heurte à un problème de dispersion des occurrences (*data sparseness*). Autrement dit, dans un corpus de taille restreinte (quelques millions de mots), il est peu probable de trouver suffisamment d'occurrences de toutes les EPL d'un domaine, *a fortiori* d'une langue (même en se limitant aux EPL non-compositionnelles sémantiquement), pour les extraire sur une base uniquement statistique. Il faut ici préciser que « l'évidence » statistique qui permet d'identifier une association remarquable entre plusieurs mots nécessite un nombre minimum d'occurrences plus important que celui requis par le lexicographe, qui peut s'appuyer sur son intuition, ou sur des tests complémentaires. Enfin cette dispersion augmente évidemment avec la taille (le nombre de mots) des EPL à extraire.

Plusieurs solutions peuvent être adoptées face à ce manque de données. La première consiste à augmenter la taille des corpus : Charest et al. (2010) utilisent par exemple un corpus d'1,8 milliards de mots, constitué à partir de sources sélectionnées, et analysé syntaxiquement. Le Web est également une source évidente, quoique soulevant certaines difficultés : représentativité, redondance (dépêches, publicités,...), nettoyage du code HTML, textes générés automatiquement, etc. Seretan (2008) présente ainsi un outil destiné à obtenir les collocatifs d'un mot via des requêtes Google automatisées, tandis que Baroni et Bernardini (2004) proposent eux une méthode de constitution de corpus thématiquement homogènes par bootstrapping, réservée cependant à l'extraction terminologique. Une dernière solution consiste à utiliser d'autres propriétés remarquables, comme celles présentées en 1.1.2, en complément d'une mesure d'association plus classique.

1.4.1.2 Combinatoire

La combinatoire obtenue lors de l'extraction des constructions candidates est une autre limite calculatoire, qui a par exemple contraint la méthode d'extraction présentée en 2. La proportion de hapax ou quasi-hapax est en effet très importante en langue (loi de Zipf), même après lemmatisation, le phénomène étant en outre très amplifié lorsqu'il s'agit de combinaisons de plusieurs mots. Il en résulte une multiplication des types de constructions candidates, ce qui alourdit le comptage des occurrences.

Cette combinatoire explique en partie pourquoi la plupart des travaux d'extraction portent sur des cooccurrences binaires. Lorsqu'il s'agit d'EPL *n*-aires, la définition des constructions candidates adoptée (*n*-grammes contigus, *n*-grammes non contigus, ou ensembles de *n* mots syntaxiquement liés) a une incidence importante sur la combinatoire générée.

L'extraction de *n*-grammes contigus reste la méthode la moins coûteuse. Pour une phrase de *t* tokens (après l'éventuelle élimination des ponctuations, mots outils, …), et pour des expressions de 2 à *m* tokens, on obtient $\sum_{n=2}^{m} t-n+1$ occurrences de constructions candidates.

Dans le cas d'une extraction par *n*-grammes non contigus, la taille *k* de la fenêtre d'observation est un paramètre supplémentaire. Pour un *n* donné, le nombre de *n*-grammes candidats pour cette même phrase est de (la démonstration se trouve en Annexe C) :

$$\binom{k}{n}+(t-k)\binom{k-1}{n-1}$$

La taille *k* de la fenêtre d'observation doit évidemment augmenter avec *n*. On peut raisonnablement estimer qu'une fenêtre deux fois supérieure à *n* est un minimum (elle est même traditionnellement plus importante pour les bigrammes, avec $k = 5$). Si $t \geq 2n$, une estimation basse du nombre d'occurrences de constructions candidates est alors de :

$$\sum_{n=2}^{m}\binom{2n}{n}+(t-2n)\binom{2n-1}{n-1}$$

Enfin, en cas d'analyse syntaxique, la combinatoire dépend de la structure syntaxique de la phrase. Si l'on adopte le formalisme des grammaires de dépendances (présenté en 2.2.1), et une contrainte stricte d'arbre enraciné (un seul gouverneur syntaxique par token, et un seul token sans gouverneur syntaxique), on obtient en théorie un nombre de candidats compris entre $\sum_{n=2}^{m} t-n+1$ pour un arbre de dépendances totalement linéaire, et $\sum_{n=2}^{m}\binom{t-1}{n}$ pour un arbre de dépendances à un seul niveau de profondeur, mais ces deux cas de figure sont hautement improbables dans une phrase réelle et de taille moyenne.

A titre d'indice, pour une phrase de 20 tokens (hors ponctuation), et pour des expressions de 2 à 4 tokens, on obtient en tout 54 *n*-grammes contigus, contre 704 *n*-grammes contigus ou non contigus. Pour l'analyse syntaxique, si l'on suppose que l'arbre syntaxique de la phrase a au minimum 2 niveaux de profondeur, et au plus 5 gouvernés par gouverneur (ce qui reste une estimation très pessimiste), on obtient au pire 185 candidats. Du point de vue de la combinatoire, l'utilisation de cooccurrences syntaxiques est donc un compromis, réduisant le nombre de candidats par rapport aux *n*-grammes non contigus, comme le constatent Martens et Vandeghinste (2011), bien qu'il reste largement supérieur au nombre de *n*-grammes contigus.

Pour cette même phrase, et pour cette même configuration syntaxique, le tableau 1 présente le nombre de candidats cumulés lorsque la taille maximale de *n* (soit *m* ci-dessus) augmente. On y voit à quel point l'augmentation de la taille maximale des EPL que l'on souhaite extraire peut multiplier le nombre de candidats.

Taille maximale de *n*	*n*-grams contigus	*n*-grams non contigus	sous-arbres syntaxiques
2	19	54	19
3	37	214	69
4	54	704	185

TABLEAU 1 : Nombre cumulé de constructions candidates pour une phrase de 20 tokens

1.4.2 Taille des constructions extraites

Hors extraction terminologique, les travaux consacrés à l'extraction automatique d'EPL restent très majoritairement centrés sur les constructions binaire. Ceci peut être imputé aux limites calculatoires qui viennent d'être évoquées, ou éventuellement à l'absence d'une mesure d'association authentiquement n-aire reconnue, mais aussi à l'attention portée en linguistique à l'étude des collocations binaires. Cela restreint cependant considérablement la portée de l'extraction, pour de nombreux idiomes par exemple (*se sentir pousser des ailes*, *avoir le vent en poupe,* ...), mais aussi pour de très nombreuses collocations, comme le remarque Tutin (2008). L'extraction d'EPL plus que binaires est donc un des principaux atouts de la méthode que nous allons présenter.

2 MÉTHODE APPLIQUÉE

2.1 APPROCHE ADOPTÉE

Avant une description plus détaillée de la méthode d'extraction implémentée, nous essayons ici de la situer par rapport aux techniques qui ont été présentées en 1.3, tout en justifiant les principaux choix effectués (les choix plus techniques seront quant à eux développés au cours de la présentation de l'outil).

2.1.1 Langue française, corpus de presse généraliste et lexique étudié

Le corpus étudié a été constitué au LIDILEM dans le cadre du projet Emolex (`http://emolex.eu`), qui vise à extraire, étudier et comparer le lexique des émotions dans 5 langues européennes (allemand, français, anglais, espagnol et russe). Nous n'avons ici utilisé que la partie française du corpus, composée exclusivement de textes édités, soit approximativement 85% d'archives récentes de presse généraliste (*Le Monde*, *Le Figaro, Libération* et *Ouest-France,* 2007-2008), et 15% de textes littéraires contemporains (essentiellement des romans, 1950-2009), pour un total d'environ 130 millions de mots. Le choix du français est essentiellement pratique. Travailler dans sa langue maternelle est un avantage très important pour apprécier l'idiomaticité d'une construction, tant cette compétence dans une langue non native est difficile à acquérir (1.2.2). L'évaluation des expressions extraites s'en est trouvée grandement facilitée.

La presse généraliste présente un intérêt similaire : tout locuteur natif adulte est en principe capable d'apprécier l'idiomaticité des expressions extraites, sans connaissances terminologiques particulières, contrairement à un corpus juridique par exemple. Mais il ne s'agit pas de constituer une ressource lexicale représentative d'une hypothétique langue générale. Seule la méthode d'extraction est ici testée, *a priori* applicable à différents types de corpus, sous réserve qu'une analyse syntaxique automatique relativement fiable puisse leur être appliquée en amont.

Enfin, pour les raisons calculatoires exposées en 1.4.1, la méthode d'extraction est testée à partir d'un lexique restreint, élaboré au LIDILEM à des fins didactiques, dont une première version est décrite dans Augustyn et al. (2008). Il est constitué exclusivement de mots simples, soit 538 adjectifs, 270 noms et 393 verbes potentiellement porteurs du trait sémantique « émotion » (éventuellement de façon métaphorique). Seules les expressions contenant au moins un de ces lemmes ont donc été extraites.

2.1.2 Analyse syntaxique préalable du corpus

Le corpus a fait l'objet d'une analyse syntaxique en dépendances dans le cadre du projet Emolex, grâce au parseur *FDG-Connexor* (Tapanainen et Järvinen, 1997), ce qui permet notamment de regrouper diverses réalisations d'une même expression, avec et sans modifieurs (épithètes, adverbes, subordonnées, ...), et ce quelle que soit la distance entre les tokens, ou leur ordre (*une politesse exquise/une exquise politesse, montrer des signes de fatigue/les signes de fatigue qu'il a montrés, ...*), réduisant ainsi l'éparpillement des occurrences (1.4.1.1). Comparée à une approche par n-grammes non contigus, elle réduit aussi fortement le nombre de combinaisons observées (1.4.1.2), tout en étendant la fenêtre d'observation à l'ensemble de la phrase, et permet d'ignorer les cooccurrences « fortuites », comme celle de *furieuse* et *envie* dans *Elle était furieuse et n'avait pas envie de discuter.*

Elle présente aussi quelques inconvénients. La fiabilité de l'analyse en particulier est très variable, les erreurs de rattachement par exemple étant fréquentes. De plus les constructions « agrammaticales », comme *autant que faire se peut* ou *à l'emporte pièce*, à moins d'être intégrées au lexique initial du parseur, sont généralement analysées de façon incorrecte, l'erreur commise pouvant cependant varier selon le contexte, ce qui interdit leur extraction sur une base statistique. Ces constructions se trouvent heureusement être en grande majorité des locutions figées, et donc extractibles grâce à des techniques plus simples, basées sur les n-grammes contigus. Un argumentaire plus complet quant à l'intérêt de l'analyse syntaxique pour cette tâche, bien que limité au cas des collocations binaires, peut être trouvé dans Seretan (2008 et 2011).

2.1.3 Cooccurrence remarquable

Nous avons opté (dans un premier temps du moins) pour la simple cooccurrence remarquable (1.3.2.4) des mots d'une EPL comme critère discriminant. La couverture du critère de figement morphosyntaxique (1.3.2.3) est trop faible ici, tandis que les tests de substitution (1.3.2.1) et d'écarts de traduction (1.3.2.2) restent difficilement applicables, faute de ressources en français. Par ailleurs cette approche offre *a priori* une plus grande robustesse, n'étant dépendante que du corpus utilisé pour l'extraction, et de la fiabilité de l'analyse syntaxique préalable.

2.1.4 EPL n-aires

Étant donné le faible nombre de travaux combinant analyse syntaxique et extraction d'EPL plus que binaires, et l'absence de consensus sur les méthodes, il nous a semblé pertinent d'essayer de développer cette approche. Une attention particulière a été portée aux imbrications entre constructions récurrentes (le traitement est détaillé en 2.3.1).

2.1.5 Approche empirique

Enfin une approche relativement empirique a été privilégiée, basée sur l'observation des sorties aux différents stades de l'extraction. En particulier, aucun filtre catégoriel (1.3.1.3) ni patron morphosyntaxique (1.3.1.2) prédéfini n'a été appliqué, ce qui est assez original pour ce type de travail. Nous avons en effet supposé que l'analyse syntaxique permettrait de réduire suffisamment le bruit pour qu'il ne soit pas nécessaire de contraindre davantage l'extraction.

Des ajustements ont cependant été faits en cours de développement, grâce à l'observation des sorties, par affinages successifs. Ces différents choix sont décrits en détails en 2.2.4, et se veulent relativement indépendants du corpus utilisé (mais pas de la langue). D'autres ajustements peuvent être envisagés, dont certains sont décrits en 3.2.2.

Par ailleurs aucune typologie prédéfinie des EPL n'a été utilisée, conformément à l'hypothèse de travail décrite en 1.1.1.2. Le revers de cette approche est l'utilisation d'une mesure statistique de cooccurrence unique pour tous types d'EPL, en tous cas de même taille (2.3.2.3). Or l'efficacité d'une mesure peut varier en fonction du type d'EPL que l'on souhaite extraire. Krenn et Evert (2001) ont par exemple montré, dans le cas des couples V+PP en allemand, qu'il était pertinent d'utiliser des mesure distinctes pour extraire les constructions à verbe support d'une part et les idiomes d'autre part. Malheureusement, même en anglais, il n'y pas (encore) de consensus quant aux mesures les plus adaptées à l'extraction d'un type particulier d'EPL (idiomes, collocations, noms composés, …).

2.2 EXTRACTION DES CANDIDATS

2.2.1 L'analyse syntaxique en dépendance

Le prétraitement syntaxique du corpus est basé sur le modèle des grammaires de dépendances, dont l'introduction en linguistique moderne est souvent attribuée à Tesnière (Kahane, 2001). Une dépendance syntaxique prend communément la forme d'une relation orientée et typée (sujet, objet, ...) entre un élément gouverneur et un élément gouverné. Chaque élément est représenté par un mot de la phrase analysée, et réfère soit au mot lui-même, soit à un syntagme dont ce mot est la tête.

La représentation des relations syntaxiques d'une phrase sous forme de dépendances constitue une abstraction importante. Elle ne tient pas compte en particulier de l'ordre des mots. Ce formalisme est par conséquent bien adapté à la représentation de motifs syntaxiques dans des langues à ordre des mots moins rigide que l'anglais, ou à ce qui nous intéresse ici, à savoir l'identification de chaînes syntaxiques récurrentes, quel que soit l'ordre sous-jacent des tokens, ou la distance qui les sépare.

La combinaison des dépendances d'une phrase permet de dessiner un graphe simple orienté, dont chaque sommet est un mot, et chaque arc une dépendance syntaxique. Tous les mots d'une phrase doivent y apparaître exactement une fois. Certaines grammaires de dépendances, plus contraignantes, exigent que ce graphe soit un arbre enraciné : un seul mot (la racine) peut être dépourvu de gouverneur, et un même mot ne peut avoir qu'un seul gouverneur (mais plusieurs gouvernés). Voici un exemple d'arbre de dépendances enraciné pour la phrase *La situation se présente sous un jour plus favorable*, où les mots sont représentés par leurs formes de surface. Les types des dépendances y figurent en italique :

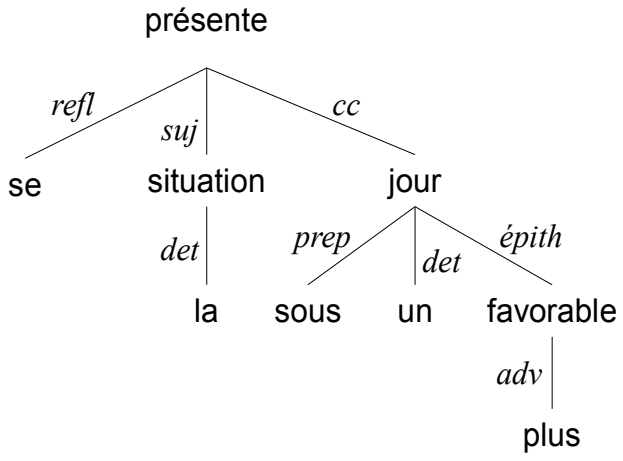


FIGURE 1 : Arbre de dépendances pour la phrase *La situation se présente sous un jour plus favorable*

Cette contrainte d'enracinement est cependant difficile à satisfaire pour plusieurs types de constructions, dont les relatives, comme dans la phrase suivante :

*Les présidents de la République **forment** une étonnante **galerie** de portraits, que cet album invite à **parcourir**.*

Le mot *galerie* peut ici être considéré à la fois comme objet (ou attribut du sujet) de *forment,* et comme objet de *parcourir,* ce qui fait deux gouverneurs pour un même gouverné. Une solution, adoptée entre autres par Hudson (2000) dans le cadre de la *Word Grammar,* consiste à distinguer deux niveaux de dépendances, par exemple des dépendances « de surface » (entre *parcourir* et *que* ici) et des dépendances « profondes » (entre *parcourir* et *galerie*), seules les dépendances de surface pouvant participer à la construction de l'arbre.

Cette solution se révèle cependant peu adaptée à notre objectif d'extraction d'EPL : *il dresse un constat accablant* et *le constat accablant qu'il dresse* doivent par exemple être considérés comme deux occurrences de la même expression (*dresser un constat accablant*), ceci afin de limiter la dispersion des données (1.4.1.1), mais aussi de garantir la cohérence des résultats. Or cette similarité est perdue dans le cas d'un représentation arborescente : l'objet de *dresser* sera *constat* dans le premier cas, et *que* dans le second.

L'analyse produite par de nombreux parseurs, dont *Connexor*, respecte cette contrainte d'enracinement. C'est pourquoi nous avons développé deux modèles d'extraction des candidats, l'un basé sur les sorties arborées directement produite par l'analyseur, et l'autre basé sur une version post-traitée, que nous avons supposée plus adéquate, les dépendances d'une phrase pouvant alors former un graphe simple orienté.

Il faut enfin noter que l'analyse réalisée par *Connexor* peut parfois produire plusieurs arbres de dépendances distincts pour différentes parties d'une même phrase, en particulier une phrase complexe, où certaines propositions pourront être analysées de façon autonome, lorsque le parseur ne parvient pas à fournir un arbre unique cohérent pour l'ensemble de la phrase. Cette robustesse est un atout important pour l'extraction d'EPL, comme le souligne Seretan (2008), ces analyses étant souvent justes localement. Il est ainsi possible d'extraire des candidats à partir de phrases complexes partiellement analysées, fréquentes dans la presse généraliste, phrases qui seraient simplement ignorées par des analyseurs moins robustes.

2.2.2 Extraction à partir d'arbres enracinés

2.2.2.1 Linéarisation

La première méthode employée, qui travaille à partir d'arbres enracinés, considère tout sous-arbre de dépendances comme un candidat au statut d'expression polylexicale. On les appellera simplement arbres et sous-arbres dans ce qui suit.

Afin de compter les occurrences des types de sous-arbres trouvés dans le corpus, il est nécessaire de pouvoir les décrire à l'aide d'une chaîne de caractères. Différentes représentations peuvent être adoptées, certaines privilégiant la concision, d'autres la lisibilité. Dans tous les cas doit exister une bijection entre l'ensemble des chaînes possibles et celui des sous-arbres possibles. Cette condition peut être remplie en simulant un parcours de ce sous-arbre depuis sa racine (*depth-first* ou *breadth-first,* au choix), et en appliquant les deux règles suivantes.

- Les nœuds frères sont parcourus par ordre alphabétique de leurs identifiants. Par exemple, dans l'arbre de la figure 1, le nœud *jour* sera parcouru avant le nœud *se*, qui sera parcouru avant le nœud *situation.*
- Tout nœud d'un sous-arbre doit avoir un identifiant unique dans ce sous-arbre. Si deux nœuds du même sous-arbre ont le même identifiant (par exemple *avoir_V* et *avoir_V* si l'on utilise les lemmes-catégories), ils doivent être distingués par un indice numérique (par exemple *avoir_V(1)* et *avoir_V(2)*), qui indique l'ordre dans lequel ils ont été rencontrés lors du parcours. Si ces deux nœuds sont frères (ce qui est très rare en pratique), leurs descendants sont utilisés pour les ordonner.

Nous avons choisi (dans un premier temps en tous cas) une représentation privilégiant la lisibilité, en particulier la lisibilité des relations de dépendance, plutôt que celle des relations de co-dépendance (rapports entre nœuds frères). Les chaînes de caractères générées correspondent à un parcours en profondeur (*depth-first*) du sous-arbre, avec répétition des nœuds ayant plusieurs gouvernés : une dépendance y est représentée par le signe "→", et l'absence de dépendance par le signe "^^". Ainsi chaque première apparition d'un nœud dans la chaîne est immédiatement précédée de son gouverneur syntaxique, et ses éventuelles apparitions suivantes sont précédées de "^^". A titre d'exemple, la figure 2 représente le sous-arbre ordonné correspondant à *donner des sueurs froides*, les nœuds étant décrits par leurs formes de surface :

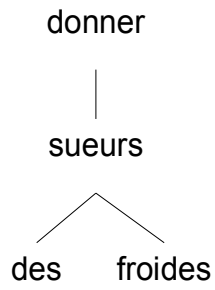


FIGURE 2 : Sous-arbre de dépendances pour l'expression *donner des sueurs froides*

Et voici sa représentation linéaire :

donner → sueurs → des ^^ sueurs → froides

Avec des lemmes-catégories, on obtient :

donner_V → sueur_N → un_DET ^^ sueur_N → froid_ADJ

Et en ajoutant les types des dépendances :

donner_V → obj:sueur_N → det:un_DET ^^ sueur_N → epith:froid_ADJ

2.2.2.2 Extraction de tous les sous-arbres d'un arbre enraciné

Pour chaque arbre enraciné (phrase ou éventuellement proposition), tous les sous-arbres qu'il inclut doivent être extraits en tant que candidats. Nous avons limité cette extraction aux sous-arbres de 2 à 6 nœuds, afin de réduire le temps de calcul nécessaire au comptage des occurrences. Soit la phrase *Elle donne libre cours à ses envies,* représentée par l'arbre ordonné de la figure 3 (où les nœuds son identifiés par leurs formes de surface) :

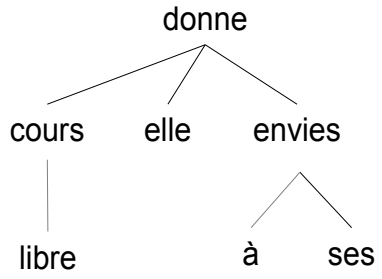


FIGURE 3 : Arbre de dépendances ordonné pour la phrase *Elle donne libre cours à ses envies*

Il faudra entre autres extraire de cette phrase :

donne → cours

donne → cours → libre

cours → libre

donne → cours ^^ donne → elle

donne → cours → libre ^^ donne → elle

donne → elle

donne → cours ^^ donne → envies

donne → cours → libre ^^ donne → envies → à

envies → à ^^ envies → ses

etc.

Nous avons utilisé pour cela un parcours en profondeur préfixé de l'arbre représentant la phrase, avec duplication de certains des sous-arbres déjà extraits lors de l'arrivée sur un nouveau nœud. Chaque sous-arbre $S_{i\ (0<i<n)}$ est une liste de nœuds, et SA est la liste de tous les S_i. A chaque arrivée sur un nouveau nœud N, ayant pour gouverneur le nœud GOUV :

- Une nouvelle liste UNAIRE est créée, qui contient uniquement le nœud courant.
- Chaque liste S_i est parcourue à la recherche de GOUV. S'il est trouvé, S_i est dupliquée, et la nouvelle liste NOUVEAU créée se voit ajouter le nœud courant (ou bien la valeur spéciale '^^', puis GOUV, puis le nœud courant, si GOUV n'était pas en dernière position de S_i).

Soit le pseudo-algorithme suivant :

```
Pour chaque N parcouru
      Créer un tableau TAMPON ;
      Créer un tableau UNAIRE ;
      UNAIRE[0] ← N ;
      TAMPON[0] ← UNAIRE ;
      Pour chaque Si de SA
            Si Si contient GOUV
                 NOUVEAU ← Dupliquer(Si) ;
                 Si GOUV n'est pas le dernier élément de Si
                       Ajouter '^^' à NOUVEAU ;
                       Ajouter GOUV à NOUVEAU ;
                 fsi
                 Ajouter N à NOUVEAU ;
                 Ajouter NOUVEAU à TAMPON ;
           fsi
      fpour
      Déverser TAMPON dans SA ;
fpour
```

Un exemple détaillé est donné en annexe D.

En fin de parcours, chaque S_i représente bien un sous-arbre (mais on ne retiendra pas les sous-arbres unaires). Les chaînes sont écrites en insérant un '→' entre deux mots consécutifs (c'est-à-dire non séparés par un '^^').

La combinatoire peut être importante, et dépend de la structure de l'arbre. A titre indicatif, sur un échantillon de 180 000 phrases du corpus étudié, environ 340 sous-arbres de 2 à 6 nœuds peuvent être extraits en moyenne par phrase (nous avons cependant réduit cette combinatoire en excluant certaines catégories rétablies par suite, exclusions qui seront justifiées en 2.2.4).

2.2.2.3 Rapports d'inclusion

Afin d'identifier, parmi tous les sous-arbres extraits d'un corpus, les sous-arbres significativement récurrents, il peut être nécessaire de disposer de leurs rapports d'inclusion respectifs. Cette question sera abordée plus en détails en 2.3.1, seul l'algorithme utilisé pour extraire ces rapports d'inclusion est ici présenté.

Cette information peut éventuellement être récupérée après l'extraction, en comparant les chaînes obtenues : un sous-arbre inclus est représenté par une chaîne dont toutes les dépendances (c'est-à-dire ici toutes les sous-chaînes du type "*mot1* → *mot2*") sont présentes dans la chaîne représentant le sous-arbre incluant. Cette solution peut cependant s'avérer lourde, compte tenu du nombre très élevé de chaînes obtenues (1.4.1.2). Il est donc plus intéressant d'extraire les rapports d'inclusion pendant le parcours initial des arbres.

Voici l'algorithme utilisé. L'objectif est, pour chaque sous-arbre candidat, d'obtenir la liste de ses sous-arbres immédiatement inclus (*SII* dans ce qui suit). Pour un sous-arbre de taille 5 (c'est-à-dire de 5 nœuds ici), ce sera la liste de ses sous-arbres inclus de taille 4, et pour un sous-arbre de taille 4, ses sous-arbres inclus de 3, etc. Il est alors possible par récursion de retrouver pour un sous-arbre de taille *n* tous ses sous-arbres inclus de tailles 2 à *n*-1.

Si l'on suit la notation de l'algorithme précédent, lors de la création d'un nouveau sous-arbre `NOUVEAU`, celui-ci ne peut avoir que deux types de SII :

- ceux ne contenant pas `N` : il n'y en qu'un, c'est S_i.
- ceux contenant `N` : ce sont tous les SII de S_i, auxquels on a ajouté `N`.

Soit `SII{S}` la liste des SII d'un sous-arbre `S`. Dans l'algorithme précédent, à l'arrivée sur un nouveau nœud `N`, on parcourt `SA` par taille croissante de S_i . Au moment de dupliquer un S_i , on effectue les opérations suivantes :

```
NOUVEAU ← Dupliquer(Si) ;
Créer SII{NOUVEAU} ;
Ajouter Si à SII{NOUVEAU} ;
Pour chaque Sj de TAMPON et de taille immédiatement
inférieure à NOUVEAU
      Si SII{Sj} et SII{Si} ont au moins un élément commun
            Ajouter Sj à SII{NOUVEAU} ;
      fsi
fpour
```

Un exemple détaillé est donné en annexe D.

2.2.3 Extraction à partir de graphes simples orientés

2.2.3.1 Justification linguistique, et conséquences pour l'extraction

Comme il a été montré en 2.2.1, la contrainte d'enracinement génère parfois des analyses peu adaptées à l'extraction d'EPL. Il est en effet souhaitable dans certains cas que des constructions divergentes en surface, mais pouvant être considérées comme des réalisations d'une même expression, soient analysées de la même manière. Par exemple *garder un souvenir* et *le souvenir qu'il en garde* gagnent à être analysés comme deux occurrences de la même construction, ce qui n'est pas le cas si l'on se limite à une analyse syntaxique de surface. Cette analyse syntaxique (légèrement) plus profonde est aussi souvent plus intuitive. Hagège et Roux (2003) décrivent ainsi une liste d'opérations de normalisation de l'analyse syntaxique produite par le *Xerox Incremental Parser* allant en ce sens. Quelques-unes de ces opérations ont été appliquées au corpus en post-traitement dans le cadre du projet Emolex, aboutissant à une seconde version du corpus, qui nous a semblé plus adéquate pour notre tâche d'extraction. Ces traitements ont cependant des conséquences d'un point de vue algorithmique, sur l'extraction des candidats, mais aussi sur leur caractérisation. Voici les trois principaux changements ayant eu une incidence.

- L'antécédent d'une relative devient dépendant de son verbe (sujet, objet, attribut...) : par exemple, pour *le constat qu'il dresse* :
  constat → dresse → qu'
  devient : *dresse → constat ^^ dresse → qu'*
- En cas de coordination, selon le type de la dépendance traitée, les deux termes coordonnés deviennent tous deux soit pères soit fils de cette dépendance.

- père : par exemple, pour *sans foi ni loi* :
  foi → loi ^^ foi → ni ^^ foi → sans
  devient : *foi → loi → sans ^^ foi → ni ^^ foi → sans*
  sans a alors deux gouverneurs (*foi* et *loi*)

- fils : par exemple, pour *gagner du temps et de l'argent* :
  gagner → temps → argent ^^ temps → et
  devient : *gagner → argent ^^ gagner → temps → argent ^^ temps → et*
  gagner a alors deux objets (*temps* et *argent*)

- Le sujet de l'auxiliaire ou semi-auxiliaire pour *Connexor* devient sujet du participe passé ou de l'infinitif. Par exemple, dans *le temps va se gâter :*
  gâter → va → temps
  devient : *gâter → temps ^^ gâter → va*
  Le rapprochement avec *le temps se gâte* devient alors possible.

L'éventuelle combinaison de plusieurs de ces règles (par exemple lorsque le verbe d'une relative a un auxiliaire) est également prise en compte.

La représentation de la phrase dans certains cas n'est plus un arbre, mais un graphe simple orienté, dont un ou plusieurs sommets peuvent être de degré entrant > 2 (autrement dit ils ont plusieurs gouverneurs syntaxiques), et autorisant plusieurs sommets de degré entrant 0 (c'est-à-dire sans gouverneur syntaxique), notamment en cas de relative. Les sous-arbres à extraire deviennent alors des sous-graphes.

La figure 4 représente le graphe obtenu pour la phrase *Je m'étonne des choix qu'ils ont faits.* On y trouve un sommet de degré entrant 2 (*choix*), et deux sommets de degré entrant 0 (*faits* et *étonne*). En cas de relatives multiples, le nombre de sommets de degré entrant 0 augmente d'1. Si elles sont coordonnées, le sommet de degré 2 devient de degré 3, tandis qu'en cas de relatives multiples imbriquées, un autre sommet devient de degré 2.

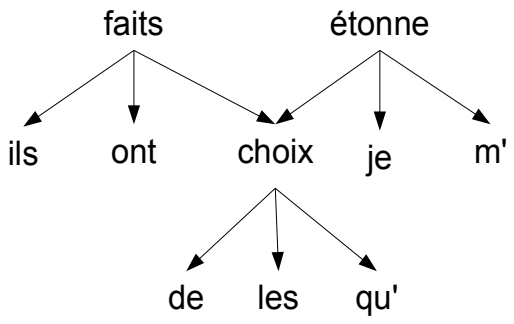


FIGURE 4 : Graphe de dépendances pour *Je métonne des choix qu'ils ont faits* après post-traitement

La figure 5 représente quant à elle le sous-graphe pour *des hommes libres et égaux en droit.* On retrouve un sommet de degré entrant 2 (*égaux*), mais un seul sommet de degré entrant 0 (*hommes*).

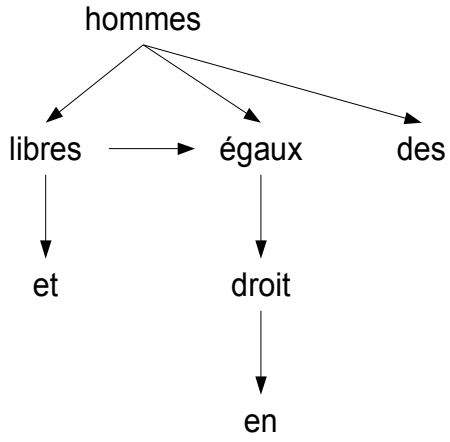


FIGURE 5 : Graphe de dépendances pour *Des hommes libres et égaux en droit* après post-traitement

Une autre méthode d'extraction, plus robuste, a dû être employée pour travailler sur cette seconde version du corpus. Cette méthode autorise aussi l'extraction à partir d'arbres enracinés, et peut donc être appliquée à la première version, mais elle s'avère dans ce cas inutilement coûteuse.

2.2.3.2 Linéarisation

La représentation adoptée pour les sous-arbres peut être conservée ici, moyennant une légère modification : chaque première occurrence d'un token dans la chaîne est immédiatement précédée de son **ou un de ses** père**(s)** syntaxique**(s)**, sauf la racine **et autres nœuds sans père.** Lorsque le graphe ou le sous-graphe a plusieurs sommets de degré 0 (c'est-à-dire plusieurs tokens sans gouverneur syntaxique), ils apparaissent par ordre alphabétique dans la chaîne. Enfin une même dépendance ne doit apparaître qu'une fois.

Pour l'arbre de la figure 4, on obtient :

étonne → choix → de ^^ choix → les ^^ choix → qu' ^^ étonne → je ^^ étonne → m' ^^ faits → choix ^^ faits → ils ^^ faits → ont

Et pour celui de la figure 5 :

hommes → des ^^ hommes → égaux → droit → en^^ hommes → libres → égaux ^^ libres → et

Cependant, pendant l'extraction des candidats d'une même phrase, les sous-graphes sont représentés par de simples listes ordonnées de dépendances. C'est donc cette cette dernière représentation qui sera privilégiée dans ce qui suit. La conversion d'un mode de représentation à l'autre est une opération simple, et ne sera pas détaillée ici.

2.2.3.3 Extraction des sous-graphes

Le cas des sommets ayant deux pères syntaxiques (l'exemple *des hommes libres et égaux en droit*) pourrait être traité assez simplement au moyen d'une structure arborescente, moyennant quelques modifications de l'algorithme précédent. Mais le cas des multiples nœuds sans père nous a semblé nécessiter une autre approche. Entre autres, il est impossible sans parcours préalable de déterminer dans une même phrase si deux mots sans gouverneur appartiennent bien au même graphe. En effet les phrases complexes sont fréquemment analysées par *Connexor* comme plusieurs graphes disjoints (2.2.1).

Nous avons donc préféré utiliser des tables, avec pour lignes des sous-graphes, et pour colonnes des dépendances, afin d'extraire de façon combinatoire tous les sous-graphes d'une phrase. Cette méthode offre l'avantage de ne pas imposer d'ordre de parcours (et donc de nœud racine), et autorise l'extraction simultanée de tous les sous-graphes d'une phrase, même lorsqu'elle est constituée de plusieurs graphes disjoints. La figure 6 représente une structure possible pour une phrase avec relative et coordination. Les indices n'ont pas de signification particulière (l'ordre des tokens dans la phrase peut par exemple être utilisé).

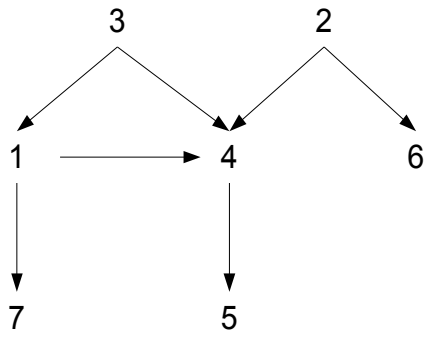


FIGURE 6 : Exemple de graphe de dépendances

Par souci de lisibilité, on représentera ici chaque dépendance par deux chiffres, le gouverneur puis le gouverné. Par exemple la dépendance 1 → 7 est notée 17. On obtient alors les dépendances ordonnées (par indices) : 14, 17, 24, 26, 31, 34 et 45, qui sont stockées en tant que sous-graphes de taille 2 (la taille correspondant au nombre de sommets). Une matrice est

ensuite créée, représentée par le tableau 2, qui signale les arcs partageant exactement un sommet. La valeur 2 indique qu'au moins un des deux arcs contient un sommet de degré entrant > 1 , et la valeur 1 correspond aux autres cas.

	14	17	24	26	31	34	45
14	X	2	2		2	2	2
17	2	X			1		
24	2		X	2		2	2
26			2	X			
31	2	1			X	2	
34	2		2		2	X	2
45	2		2			2	X

TABLEAU 2 : Exemple de matrice de dépendances

Les sous-graphes de taille 3 désignés par les cases ayant la valeur 1 sont dédoublonnés (31+17 et 17+31 désignent par exemple le même sous-graphe 3 → 1 → 7), puis stockés. Les sous-graphes de taille 3 désignés par les cases ayant la valeur 2 peuvent être décrits de façon incomplète par ce couple d'arcs : par exemple le couple 31+34 décrit de façon incomplète le sous-graphe formé des sommets 1, 3 et 4, qui compte un troisième arc (14). On vérifie donc s'il existe un troisième arc entre deux de ses sommets, et, le cas échéant, cet arc lui est ajouté (31+34 devient par exemple 14+31+34). Ils sont ensuite dédoublonnés, puis stockés.

Tous les sous-graphes de taille 3 obtenus sont ensuite réutilisés comme entrées (lignes) du tableau 3, afin d'extraire les sous-graphes de taille 4, en suivant la même procédure. L'opération est ensuite réitérée jusqu'à la taille maximale souhaitée.

	14	17	24	26	31	34	45
14 17	X	X	2		2	2	2
14 24	X	2	X	2	2	2	2
14 45	X	2	2		2	2	X
17 31	2	X			X	2	
24 26	2		X	X		2	2
24 34	2		X	2	2	X	2
24 45	2		X	2		2	X
14 31 34	X	2	2		X	X	2

TABLEAU 3 : Exemple de matrice de dépendances et sous-graphes

2.2.3.4 Rapports d'inclusion

L'extraction des rapports d'inclusion peut ici aussi se faire pendant celle des candidats. Chaque case d'un des tableaux ci-dessus contenant la valeur 1 ou la valeur 2 correspond à la combinaison d'un sous-graphe X_1 de taille n et d'une dépendance. Le sous-graphe Y de taille n + 1 désigné par cette case a donc X_1 pour sous-graphe immédiatement inclus. Or tous les sous-graphes de taille n du graphe en cours sont présents en tant qu'entrées de ce tableau. Par conséquent, l'ensemble de tous les sous-graphes immédiatement inclus dans Y est l'ensemble des X_1, X_2,... X_m pour chacune des m cases désignant Y, avant dédoublonnement.

Par exemple, dans le tableau 3, le sous-graphe 14 17 24 apparaît deux fois : sur la ligne de 14 17, dans la case 14 17+24, et sur la ligne de 14 24, dans la case 14 24 + 17. Il a donc deux sous-graphes immédiatement inclus, 14 17 et 14 24. Ce principe est également valable pour les sous-graphes incomplets. Dans le tableau 2 par exemple, le sous-graphe 14 31 34 apparaît sous 6 formes :

- sur la ligne de 14, dans les cases 14+31 et 14+34,
- sur la ligne de 31, dans les cases 31+14 et 31+34,
- sur la ligne de 34, dans les cases 34+14 et 34+31,

Il a donc pour sous-graphes inclus 14, 31 et 34.

2.2.4 Traits pertinents et ajustements linguistiques

Cette partie présentera quelques choix complémentaires effectués afin d'améliorer l'extraction des candidats, des choix les plus évidents à ceux qui nous ont semblé les plus délicats. Ces ajustements linguistiques sont propres à la langue du corpus (le français), et parfois aussi à l'analyseur syntaxique utilisé (*Connexor*), mais se veulent autant que possible indépendants de la nature du corpus (à dominante journalistique), et des lemmes étudiés.

Contrairement à la très grande majorité des travaux d'extraction d'EPL, nous avons préféré ne pas utiliser de filtres catégoriels prédéfinis (1.3.1.3), du type N+Adj, V+N ou V+N+Prep+N, afin de ne pas restreindre la gamme des constructions candidates, supposant que la motivation syntaxique des cooccurrences permettrait de limiter le bruit. Nous avons à la place utilisé une simple liste de mots et catégories ignorés (*stop words*), constituée et enrichie de façon empirique, par ajustements successifs. Cette solution peut surprendre, mais offre une importante souplesse d'utilisation. L'application de filtres ou patrons lors de futurs essais n'est cependant pas exclue ; pour le français notamment, les patrons les plus productifs en EPL sont particulièrement bien recensés, les travaux de Maurice Gross entre autres y ayant largement contribué.

Enfin un autre enjeu important ici est le degré d'abstraction adopté dans la description des constructions candidates. S'il est inadéquat, il peut conduire soit au regroupement abusif d'occurrences d'EPL distinctes, soit au contraire à une multiplication des types pour une même EPL.

2.2.4.1 Lemmatisation

Afin de limiter la dispersion des occurrences (1.4.1.1), nous avons préféré représenter les mots par leurs lemmes-catégories. Ainsi *un signe encourageant* et *des signes encourageants* sont bien considérés comme deux occurrences de la même EPL. Cela n'empêche pas l'identification de préférences flexionnelles : par exemple *jour* dans *couler des jours heureux* n'admet que le pluriel. Ces préférences sont récupérées dans un second temps, pour chaque EPL validée, à partir de ses occurrences. Cette phase de caractérisation morphosyntaxique sera décrite en (2.4).

La lemmatisation peut cependant très ponctuellement être source de regroupements erronés. Ainsi *faire une avance* et *faire des avances* seront considérés à tort comme deux occurrence d'une même EPL.

Les numéraux cardinaux ont par ailleurs fait l'objet d'une lemmatisation plus abstraite (il prennent tous le même lemme), tout comme la série *dizaine*, *douzaine*, *quinzaine*, *vingtaine*... La forme de surface est alors stockée comme un trait du mot, afin d'être rétablie, pendant la phase de caractérisation morphosyntaxique, pour les quelques expressions n'admettant qu'un lemme précis (*voir 36 chandelles, une douzaine d'huîtres, ...*).

2.2.4.2 Déterminants

Après observation de plusieurs séries de résultats, les déterminants sont apparus comme un facteur important de multiplication des types pour une même EPL, notamment pour de nombreuses EPL compositionnelles sémantiquement. Par exemple *ce digne successeur*, *un digne successeur* et *le digne successeur* pourront être analysés comme des occurrences de trois EPL distinctes (de trois mots chacune). Il est donc globalement plus intéressant de considérer la détermination comme un des traits du nom (ou du pronom), au même titre que ses traits flexionnels. Les éventuelles préférences pour un déterminant particulier (par exemple le déterminant démonstratif dans *digne de ce nom,* ou l'article défini dans *faire froid dans le dos*) sont rétablies après l'extraction, lors de la phase de caractérisation morphosyntaxique, décrite en 2.4. Comme pour la lemmatisation, cette solution peut cependant très ponctuellement donner lieu à des regroupements erronés : *faire une affaire* et *faire l'affaire* seront par exemple considérés comme deux occurrences d'une même EPL.

2.2.4.3 Auxiliaires et semi-auxiliaires

La prise en compte des auxiliaires ou semi-auxiliaires est, plus encore que celle des déterminants, facteur d'éparpillement. *Cela lui a fait froid dans le dos* et *ça fait froid dans le dos* pourront ainsi être considérés comme des occurrences de deux EPL distinctes (*avoir fait froid dans le dos* et *faire froid dans le dos*).

Cependant, contrairement aux déterminants (qui sont des feuilles), les auxiliaires ou semi-auxiliaires ne peuvent être supprimés d'un arbre enraciné sans modifier sa structure : le sujet pour *Connexor* dépend de l'auxiliaire/semi-auxiliaire et non du participe passé/infinitif. C'est l'une des raisons qui nous a poussés à utiliser la seconde version du corpus (2.2.3.1).

2.2.4.4 Types des dépendances

Lin (1999) exploite les types des dépendances (objet, sujet, ...) en tant que critère distinctif entre expressions candidates. Une tentative en ce sens s'est avérée avoir très peu d'incidence sur les sorties, comparée à une extraction sans typage des dépendances. Ce phénomène était largement prévisible : pour un ensemble donné de lemmes-catégories, et pour un ensemble de dépendances non typées donné entre ces lemmes-catégories, si la combinaison obtenue est une EPL, alors les types des dépendances sont généralement stables. Par exemple, dans le sous-graphe *garder_V* → *souvenir_N* → *ému_A*, les dépendances auront presque systématiquement pour types respectifs objet et épithète. Et, une nouvelle fois, cette information peut être assez simplement récupérée lors de la phase de caractérisation morphosyntaxique(2.4).

Il nous a semblé plus intéressant de ne pas utiliser les types des dépendances pendant l'extraction pour une raison spécifique à l'analyseur utilisé : les verbes au passif ne sont pas identifiés comme tels par *Connexor* (seul le complément d'agent est parfois reconnu). Or, toujours pour limiter l'éparpillement, il est utile d'analyser par exemple *ils ont émis un avis défavorable* et *un avis défavorable a été émis* comme deux occurrences de l'EPL *émettre un avis défavorable*. Les types des dépendances étant différents (objet dans le premier cas, sujet dans le second), l'absence de typage est ici un avantage. Une solution alternative aurait été l'utilisation d'une liste exhaustive des verbes prenant l'auxiliaire *être* à la voix active.

2.2.4.5 Adjectifs attributs

Les adjectifs attributs ont été rattachés au sujet, comme les épithètes (mais avec un type de dépendance spécifique), autorisant par exemple le rapprochement entre *un constat accablant*, *le constat est accablant* et *le constat reste accablant,* ce qui n'exclut pas cependant la possibilité

que ces expressions soient autonomes, si leurs occurrences individuelles sont suffisantes. Le revers est ici l'impossibilité de détecter les associations privilégiées entre un verbe d'état et un adjectif attribut sans sujet caractéristique, par exemple *rester calme.*

2.2.4.6 Dépendances entre propositions

Dans une grammaire de dépendances, les propositions sont généralement représentées par leur verbe (ou par le participe pour les participiales), qui est dépendant du verbe de la principale. Ceci peut conduire à l'extraction de motifs difficilement interprétables par un utilisateur humain, quoique correspondant à des régularités avérées. On obtient par exemple des constructions comme *transporter → blesser ^^ transporter → hôpital* , qui correspond à la structure *Blessé* (*grièvement/à la jambe/...*)*, il a été transporté à l'hôpital.* Ces dépendances (ici *transporter → blesser*) ont donc été ignorées pendant l'extraction des candidats, tout comme les relations de coordination entre verbes, pour la même raison, bien que cela puisse encore une fois nuire très ponctuellement au rappel (*à prendre ou à laisser*, *vivre et laisser mourir, ...*).

2.2.4.7 Pronoms personnels

Le cas des pronoms est plus délicat. De nombreuses expressions *a priori* trop longues contenant un pronom personnel à l'indicatif sont ainsi extraites par défaut, comme *je suis curieux de savoir* ou *ils forment le noyau dur.* Le pronom nous a semblé superflu dans ces cas de figure, même avec une préférence marquée du verbe pour une personne et un nombre donné : cette préférence peut en effet être retrouvée à partir des traits flexionnels du verbe, lors de la phase de caractérisation (2.4). Par exemple l'expression *porter ses fruits* n'admet que la troisième personne, mais différents noms ou pronoms peuvent satisfaire cette condition (*l'expérience, il, celui-ci, celles-ci…).* Les pronoms personnels à l'indicatif non toniques, qui sont des feuilles, ont donc été retranchés des arbres et graphes avant extraction des candidats. Ce choix peut cependant lui aussi avoir ponctuellement des conséquences indésirables : les expressions requérant un pronom personnel à l'indicatif spécifique, généralement le pronom impersonnel *il,* sont incomplètes (on obtient par exemple **faire une température glaciale*).

Pour les autres pronoms personnels (accusatif, datif, réfléchis), une lemmatisation plus poussée que celle effectuée par *Connexor* s'est par ailleurs avérée nécessaire, effaçant les distinctions en terme de personne, afin toujours de limiter la multiplication des types pour une même expression : ainsi *me* et *se* (réfléchis) sont bien considérés comme des occurrences d'un même lemme, et par conséquent *je me brosse les dents* et *elle se brosse les dents* comme deux occurrences de la même EPL.

2.2.4.8 Prépositions

Les prépositions sont partie intégrante de très nombreuses expressions (*dormir à poings fermés*, *par le plus grand des hasards, faire froid dans le dos, à guichets fermés...),* et leur omission conduirait à l'extraction d'EPL incomplètes. Dans d'autres cas cependant, une construction avec préposition apparaît plutôt comme une forme possible d'une EPL plus courte. Par exemple *le mauvais temps* et *par mauvais temps*, peuvent être considérés comme deux occurrences d'une même EPL (*mauvais temps*), plutôt que comme deux EPL autonomes présentant un rapport d'inclusion, comme le sont *coup d'œil* et *jeter un coup d'œil* (le traitement des rapports d'inclusion sera détaillé en 2.3.1).

Considérer les prépositions comme des mots à part entière conduit ainsi à l'extraction des couples de constructions suivants, où chacune est considérée comme une EPL autonome :

dénouement heureux | *jusqu'à un dénouement heureux*
faire froid dans le dos | *à faire froid dans le dos*
un congrès extraordinaire | *lors du congrès extraordinaire*
un ton enjoué | *sur un ton enjoué*
la guerre froide | *pendant la guerre froide*
un disque dur | *sur un disque dur*
l'ambiance décontractée | *dans une ambiance décontractée*

Ces couples sont alors analysés de la même manière que :

verre brisé | *un bruit de verre brisé*
un cri déchirant | *pousser un cri déchirant*
sang froid | *assassiner de sang froid*
un air dégoûté | *prendre un air dégoûté*

Les deux cas de figure nous ont semblé suffisamment distincts pour justifier deux types de traitements : considérer *un disque dur* et *sur un disque dur* comme deux occurrences de la même EPL (*disque dur*), mais considérer en revanche *un disque dur externe* comme une expression distincte, qui inclut *disque dur.* Ce qui conduit en pratique à traiter les prépositions de la même manière que les déterminants : elles ne sont pas prises en compte lors de l'extraction des sous-arbres ou sous-graphes candidats, mais rétablies pendant la phase de caractérisation des expressions retenues (2.4). Encore une fois, les prépositions étant des feuilles pour l'analyseur utilisé, cela n'a pas d'incidence sur l'extraction. Un traitement particulier a cependant du être adopté pour les prépositions composées.

2.3 FILTRAGE DES CANDIDATS

La phase d'extraction des candidats est généralement suivie d'un filtrage des types obtenus, qui prend en compte leur nombre d'occurrences, et permet de distinguer les associations de mots jugées fortuites d'associations jugées significatives, les dernières étant alors reconnues comme des EPL. Différentes propriétés des EPL, qui ont été détaillées en 1.3.2, peuvent être utilisées à cet effet : non compositionnalité sémantique, non substituabilité paradigmatique, comportement flexionnel ou syntaxique idiosyncrasique, et enfin cooccurrence significative des mots de l'expression. C'est cette dernière propriété, réputée plus robuste, que nous avons utilisée (2.1.3). Des filtres complémentaires peuvent également être appliqués, afin par exemple de s'assurer de la dispersion de l'expression dans le corpus. Ces différents filtres seront détaillés en 2.3.2.

L'extraction d'EPL de plus de deux termes ajoute une dimension à cette phase de filtrage : un traitement adéquat des phénomènes d'inclusion entre candidats peut être requis. Il permet à la fois d'éliminer des expressions incomplètes (**jeter un coup*, **remuer ciel*, **s'effondrer comme un château, *nager en délire*...), et de calculer un score plus juste pour les expressions restantes, en ne prenant en compte que leurs occurrences autonomes. En pratique, le traitement des inclusions et le filtrage tels que nous les avons implémentés se font de manière simultanée. Par souci de clarté cependant, le premier sera présenté de façon autonome.

2.3.1 Traitement des inclusions

Sans traitement spécifique, de nombreuses constructions incomplètes peuvent être retenues en tant qu'EPL valides. Intuitivement, il s'agit d'éliminer une information redondante. Par exemple, si la construction *faire partie des heureux élus* a été retenue, les constructions incluses **partie des heureux élus* et **faire partie des élus* auront vraisemblablement le même nombre d'occurrences que la construction incluante (ou un nombre très légèrement supérieur), mais seront sans intérêt. En revanche la construction *faire partie de* (et peut-être aussi *les heureux élus*) aura sans doute un nombre d'occurrences très supérieur, auquel cas elle pourra prétendre au statut d'EPL autonome, après soustraction des occurrences correspondant à l'expression complète. A l'inverse, si la construction **homme des neiges* apparaît de façon quasi exclusive dans l'EPL *abominable homme des neiges*, la première ne sera pas retenue.

Chi et al. (2005, in Martens, 2010) proposent un premier traitement allant en ce sens : tout sous-graphe inclus dans un autre sous-graphe au score supérieur, et apparaissant exactement le même nombre de fois que le sous-graphe incluant, est éliminé (ceci correspond à la notion de *sous-segment contraint* pour les segments répétés linéaires, définie par Lebart et Salem, 1994). Ce principe permet d'appliquer un premier filtre, mais pas de traiter les cas (majoritaires) où les

sous-graphes inclus ont un nombre d'occurrences supérieur, en particulier les cas où ce nombre n'est que très légèrement supérieur, et donc où le sous-graphe inclus ne constitue pas une expression autonome.

Nous avons donc préféré employer une méthode similaire à celle décrite par Charest et al. (2010). Le dédoublonnement s'effectue en même temps que le filtrage des résultats (dont les critères sont décrits ci-après), en commençant par les sous-graphes de taille maximale. Si un sous-graphe ne satisfait pas aux critères de filtrage, il est éliminé, sinon ses occurrences sont soustraites une à une des listes d'occurrences de tous ses sous-graphes inclus. Par exemple, si *faire partie des heureux élus* est validé avec 20 occurrences, tous ses sous-abres inclus (*faire partie*, *partie des élus*, *partie des heureux élus*, ...) se voient retrancher ces 20 occurrences (ou une partie, si certaines leur ont déjà été retranchées). La procédure est ensuite réitérée pour les sous-graphes de taille immédiatement inférieure, et ce jusqu'aux sous-graphes de deux tokens. Ainsi seules les occurrences autonomes d'une expression sont prises en compte lors du filtrage. Cet algorithme nous a semblé plus adapté, compte tenu de notre objectif, que le *LocalMax* proposé par da Silva et al. (1999), qui impose de choisir entre deux constructions présentant un rapport d'inclusion immédiate, par exemple entre *coup de fil* et *passer un coup de fil*, alors que toutes deux peuvent être des EPL valides.

Les sous-graphes de taille maximale (6 mots ici) sont par conséquent inutilisables comme tels, car non dédoublonnés. On y trouve ainsi des ensembles de constructions ayant en commun la plupart de leurs occurrences, et partageant deux à deux plusieurs tokens, qui font partie d'une EPL de taille supérieure non identifiée (7, 8, 9 mots...). Ce problème peut également survenir plus ponctuellement pour les sous-graphes de tailles inférieures, lorsque les seuils de filtrage ne sont pas optimaux : par exemple, si *remuer le couteau dans la plaie* a obtenu un score trop faible pour être retenu, il est possible que **remuer le couteau* et **le couteau dans la plaie* soient tous deux validés en tant qu'EPL de taille inférieure.

2.3.2 Filtres appliqués

Rappelons qu'à ce stade un type de sous-graphe est simplement une représentation linéaire de ses lemmes-catégories et dépendances (non typées), sans déterminant, préposition, ni pronom personnel à l'indicatif non tonique, conformément aux choix détaillés en 2.2.4. Ainsi l'EPL *souffrir d'un cruel manque de* est représentée par la chaîne *souffrir_V* → *manque_N* → *cruel_A*.

Les différents filtres appliqués pour valider ou rejeter un candidat prennent simplement en compte son nombre total d'occurrences, la dispersion de ces occurrences dans le corpus, et les occurrences (fréquences marginales) des différents lemmes-catégories inclus dans la chaîne.

Le comptage des occurrences est une opération relativement coûteuse, compte tenu de la combinatoire (1.4.1.2) : même en ignorant les catégories de mots décrites en 2.2.4, 128 sous-graphes candidats sont extraits en moyenne par phrase, dont une grande majorité de hapax (à l'échelle du corpus). Pour une raison pratique, l'extraction des candidats a donc été limitée aux phrases contenant au moins un des lemmes du lexique étudié, la liste des candidats étant réinitialisée à chaque lemme. Cette contrainte a guidé le choix de la mesure d'association utilisée, qui sera décrite en 2.3.2.3.

2.3.2.1 Seuil minimum d'occurrences

Un premier filtre simple permet de se prémunir d'erreurs ponctuelles d'analyse syntaxique, ou de toilettage du corpus. Il s'agit d'une valeur minimale pour le nombre d'occurrences multiplié par la taille du sous-graphe (le nombre de tokens), permettant notamment d'éliminer les hapax. C'est aussi une précaution courante pour garantir un minimum de pertinence aux mesures d'associations. La valeur minimale a été empiriquement fixée à 18 pour les expériences qui seront présentées en 3.1, soit respectivement 9, 6, 5, 4 et 3 occurrences pour les sous-graphes de tailles 2, 3, 4, 5 et 6.

2.3.2.2 Dispersion

Un second filtre s'assure de la dispersion des occurrences de chaque sous-graphe, le corpus étant subdivisé en plusieurs sources. Le critère requis est l'apparition du sous-graphe dans au moins trois des cinq sous-corpus. Cette précaution permet d'écarter certaines entités nommées, événements, ou encore des références entre articles, dans la presse locale notamment, ainsi que les hapax de documents. Cependant, la période couverte étant comparable pour toutes les sources journalistiques, quelques entités comme *le match amical France-Tunisie* subsistent. On peut éventuellement les considérer comme représentatives du corpus utilisé, et donc légitimes ici, bien qu'inutilisables sur un autre corpus ne couvrant pas exactement cette période. D'autres entités, comme *le festival du film fantastique de Gérardmer* ou *la Symphonie fantastique de Berlioz,* ont une portée plus importante. Ce filtrage pourra être affiné, en prenant en compte notamment la dispersion par document. Il pourrait par ailleurs être intégré au calcul du score présenté dans le paragraphe suivant, mais la pondération de ce critère, sans apprentissage supervisé, est assez arbitraire. D'un point de vue pratique, appliquer ces deux filtres de manière autonome garantit une meilleure traçabilité des résultats, et facilite ainsi les ajustements.

2.3.2.3 Mesure d'association

Le filtre suivant mesure la significativité de la cooccurrence des mots de l'expression. De très nombreuses mesures statistiques ont déjà été employées à cet effet, dont un aperçu est donné en 1.3.2.4, et dont on peut trouver une étude approfondie dans Evert (2005). Les scores d'association obtenus sont ensuite utilisés soit pour classer les candidats, soit pour fixer, souvent empiriquement, un seuil limite de validation/rejet. C'est cette seconde solution qui a été adoptée ici, le choix des valeurs seuils étant particulièrement important, du fait du dédoublonnement des constructions incluses, expliqué en 2.3.1 : pour une taille d'expression donnée (nombre de mots), si le seuil est trop permissif, des constructions trop longues seront validées (*mais jeter un coup d'oeil*), tandis que les constructions incluses valides correspondantes (*jeter un coup d'oeil*) pourront être écartées. Au contraire, si le seuil est trop restrictif, des constructions incluses incomplètes (**jeter un coup*) pourront être validées.

Le choix de la mesure retenue a été largement guidé par des contraintes calculatoires, en particulier le problème déjà évoqué du temps nécessaire au comptage des occurrences. Seules les occurrences des sous-graphes contenant au moins un des lemmes étudiés sont comptées, ce qui soulève ici deux obstacles : l'estimation des fréquences des autres lemmes du sous-graphe, et l'impossibilité de calculer des probabilités conditionnelles.

Tout d'abord les fréquences marginales exactes des autres lemmes d'un sous-graphe sont inconnues. Il s'agit en effet, pour chaque lemme, non pas de son nombre d'occurrences f parmi les tokens du corpus (dont nous disposons), mais, pour chaque taille n de sous-graphe, du nombre d'occurrences f_n de ce lemme parmi toutes les occurrences de sous-graphes de taille n. Or f_n ne peut être estimée proportionnellement à f sans un biais important : par exemple, dans une phrase, un même adverbe n'apparaîtra généralement que dans un seul sous-graphe de taille 2 (les adverbes sont généralement des feuilles), tandis qu'un même verbe apparaîtra fréquemment dans 4 ou 5 d'entre eux. Nous avons utilisé cette disparité supposée entre catégories afin d'ajuster l'estimation de f_n à partir de f, en comptant, sur 180 000 phrases réparties dans le corpus, et pour chaque taille de sous-graphe, le nombre d'occurrences de chaque catégorie (cette opération fournissant également une estimation du nombre total T_n d'occurrences de sous-graphes de chaque taille n).

Par ailleurs, pour chaque lemme d'un sous-graphe, le nombre d'occurrences du reste du sous-graphe sans ce lemme est lui aussi inconnu. Il est par conséquent impossible d'obtenir des probabilités conditionnelles exactes, par exemple p(*froid* | *faire XXX dans dos*), comme le préconisent Dias et al. (2000b), et encore moins celles de chaque sous-graphe inclus, par exemple p(*dans dos* | *faire froid XXX YYY*). Par défaut, le nombre d'occurrences attendues d'un

type de sous-graphe est donc évalué sous l'hypothèse de tirages aléatoires indépendants de *n* mots, malheureusement très inexacte dans le cadre de l'analyse de phénomènes langagiers, la probabilité d'apparition d'un lemme étant conditionnée par son contexte.

La mesure d'association utilisée est une adaptation de l'information mutuelle spécifique (Manning et Schütze, 1999:151), afin notamment d'autoriser une comparaison avec les sorties de l'outil d'extraction de cooccurrences linéaires utilisé en parallèle, le *MWE Toolkit*, développé par Ramisch, Villavicencio et Boitet (2010), présenté plus en détail en 3.1.1. Elle a été préférée au *t-score,* également implémenté aux sein du *MWE Toolkit*, pour une raison qui sera donnée ci-après.

Soit $O(w1,w2)$ le nombre d'occurrences observées pour un bigramme, $E(w1,w2)$ le nombre d'occurrences attendues de ce bigrammes sous l'hypothèse de tirages indépendants, et T le nombre total de bigrammes du corpus étudié. Le calcul de l'information mutuelle spécifique (PMI) pour ce bigramme est alors de :

$$PMI(w_1;w_2) = \log\frac{p(w_1,w_2)}{p(w_1)\cdot p(w_2)} = \log\frac{O(w_1,w_2)}{p(w_1)\cdot p(w_2)\cdot T} = \log\frac{O(w_1,w_2)}{E(w_1,w_2)}$$

Ramish et al. (*Ibid*) étendent cette formule à *n* mots contigus. La difficulté pour nous consiste alors à appliquer cette formule non pas à un *n*-gramme, mais à un sous-graphe *SG* de *n* mots. Dans la formule ci-dessus, $O(w_1,w_2)$ peut tout d'abord être remplacé par le nombre $O(SG)$ d'occurrences observées du sous-graphe, et T par le nombre évalué T_n de sous-graphes de taille *n*.

Il reste à trouver l'équivalent *PP* de $p(w_1)\ .\ p(w_2)$, qui dans la formule correspond à la probabilité d'obtenir le bigramme (w_1,w_2) lors d'un tirage ordonné de deux mots. Pour un tirages non ordonné de deux mots, cette probabilité *PP* devient $p(w_1)\ .\ p(w_2)\ .\ 2$, et pour un tirage non ordonné de *n* mots (un « sac » de mots), on a :

$$PP = p\{w_1, w_2, \ldots, w_n\} = p(w_1)\ .\ p(w_2)\ .\ \ldots\ .\ p(w_n)\ .\ n!$$

Mais en théorie (toujours sous l'hypothèse d'indépendance, c'est-à-dire sans tenir compte des co-restrictions linguistiques), plusieurs sous-arbres sont possibles pour un même ensemble de *n* noeuds : selon la formule de Cayley, le nombre d'arbres non ordonnés et non orientés pour cet ensemble de *n* nœuds est de n^{n-2}. Le nombre d'arbres de dépendances (non ordonnés, mais enracinés) est donc de n^{n-1}. Sous l'hypothèse (par défaut) d'une distribution uniforme, on obtient donc finalement la valeur suivante pour *PP* :

$$PP = \frac{p(w_1)\cdot p(w_2)\cdot\ldots\cdot p(w_n)\cdot n!}{n^{n-1}}$$

avec $p(w) = \frac{f_n(w)}{T_n}$

Le score d'un sous-arbre peut alors être calculé grâce à la formule suivante :

$$PMI(SG) = \frac{O(SG)}{PP \cdot T_n} = \frac{O(SG)}{p(w_1) \cdot p(w_2) \cdot \ldots \cdot p(w_n)} \cdot \frac{n^{n-1}}{n! \cdot T_n}$$

Le coefficient $n^{n-1} / (n ! . T_n)$ est artificiellement élevé, du fait notamment du caractère très irréaliste, pour un même ensemble de *n* lemmes, de l'hypothèse par défaut d'équiprobabilité des sous-arbres possibles : il est par exemple peu probable qu'un adjectif ait pour gouverneur un adverbe, ou même que *œil* soit gouverneur de *coup*. Il en résulte une importante surévaluation des scores. Cependant, ce coefficient étant une constante pour un *n* donné, il n'a pas d'incidence sur le simple classement des candidats de même taille *n* lorsque l'on utilise l'information mutuelle spécifique, contrairement au *t-score.* C'est pourquoi nous avons préféré cette première mesure.

Pour chaque taille *n,* le classement des candidats obtenu permet de fixer empiriquement, par observation des sous-graphes, un score minimum d'« idiomaticité ». Ce score minimum est ensuite utilisé en tant que filtre.

2.3.3 Limites

Comme il vient d'être montré, l'hypothèse de tirages indépendants et celle d'équiprobabilité des sous-graphes pour un même tirage non-ordonné de *n* lemmes tendent à fausser les scores. Il est en particulier impossible de comparer entre eux les scores de sous-graphes de tailles différentes sans une normalisation *a priori* délicate. Un comptage exhaustif des occurrences de tous les sous-graphes candidats du corpus permettrait vraisemblablement d'obtenir des scores plus justes, par application d'autres mesures d'association.

Une alternative intéressante et peu coûteuse serait cependant l'utilisation d'une mesure d'association binaire récursive, méthode qui a déjà été implémentée dans le cadre du projet Emolex par Kraif et Diwersy (2012). Il s'agirait alors de partir d'un couple de lemmes-catégories significativement cooccurrents, soit un sous-graphe de deux nœuds, puis de compléter ou étendre l'expression obtenue, en identifiant une cooccurrence significative de cette expression prise comme un tout avec un troisième lemme-catégorie, et ce récursivement jusqu'à l'EPL complète. Par exemple :

- *heureux* + *élus* sont remarquablement cooccurrents (avec *élus* gouverneur de *heureux*),

- lorsqu'il sont cooccurrents, le verbe *faire* leur est souvent associé (en tant que gouverneur de *élus*), autrement dit (*heureux* + *élus*) + *faire* est une association binaire remarquable,
- ((*heureux* + *élus*) + *faire*) + *partie* est également une association binaire remarquable (avec *faire* gouverneur de *partie*).

Cette méthode offre ici un avantage important : le nombre d'occurrences dans le corpus de chaque expression intermédiaire est connu. Dans l'exemple ci-dessus, le nombre exact d'occurrences du sous-graphe correspondant à *heureux* + *élus* est connu, tout comme celui du sous-graphe correspondant à *heureux* + *élus* + *faire*, puisque tous les sous-graphes contenant le lemme *heureux* ont été comptés. Les fréquences individuelles f_n des lemmes peuvent elles être estimées selon la méthode présentée en 2.3.2.3. Il est alors possible, sans alourdir les comptages, de s'affranchir des deux hypothèses erronées vues précédemment, en calculant certaines probabilités conditionnelles, comme *p*(*partie* | *faire heureux élus*), même si d'autres, comme *p*(*heureux* | *faire partie élus*) restent malheureusement inconnues.

On peut attendre de cette méthode un gain en précision, mais peut-être également une perte en rappel. Il est en effet possible que la cooccurrence des *n* mots d'une EPL prise comme un tout soit statistiquement significative, mais qu'aucun des couples gouverneur/gouverné internes à cette EPL ne soit remarquable, notamment lorsque les mots sont fréquents par ailleurs, comme dans *garder la tête froide, prendre les choses en main, voir le bon côté des choses, ne pas en croire ses yeux, en cas de coup dur* ou *arriver bon dernier.* Et même lorsqu'un seul des couples de mots de l'expression est remarquable, il est nécessaire, compte tenu des restrictions de comptage, que le lemme étudié (*heureux* dans l'exemple ci-dessus) fasse partie de ce couple, l'expression étant ignorée dans le cas contraire.

2.4 CARACTÉRISATION MORPHOSYNTAXIQUE

La dernière étape du processus est une phase de caractérisation automatique des expressions validées, à partir de leur occurrences uniquement. Elle remplit plusieurs fonctions :

- rétablir les mots et traits récurrents ignorés pendant la phase d'extraction (les raisons en sont données en 2.2.4) : prépositions, déterminants, auxiliaires, types des dépendances et préférences flexionnelles.
- identifier la ou les configuration(s) linéaire(s) privilégiée(s) pour chaque expression (ordre des mots et distance maximale entre mots), afin de faciliter son identification ultérieure dans des textes ayant fait l'objet d'un simple étiquetage morphosyntaxique, sans qu'une analyse syntaxique complète soit requise.
- générer une forme canonique lisible par un être humain, afin de faciliter l'évaluation, mais aussi la réutilisation des expressions extraites (lexicographie, indexation automatique, ...).

2.4.1 Rétablissement des mots et traits récurrents

La méthode employée ici est presque identique à celle décrite dans Evert et al. (2004) pour identifier les préférences flexionnelles de couples Adj+N en allemand. Pour chaque construction C retenue, pour chacun de ses mots M, et pour chacun des traits T de ce mot, la valeur V_{MT} la plus fréquente de ce trait est extraite (par exemple la valeur 'pluriel' pour le trait 'nombre' du mot *frais* dans *faire les frais*). On utilise alors le nombre d'occurrences observées de V_{MT} et celui de C pour estimer la probabilité réelle $p(V_{MT}|C)$, parmi tous les C du domaine (ou de la langue) dont le corpus se veut un échantillon. Un test binomial exact (à un niveau de confiance de 95 %) fournit un intervalle pour $p(V_{MT}|C)$. Si la limite basse de cet intervalle est supérieure à 0.5, la valeur est considérée comme significative pour le mot M dans la construction C, et donc intégrée à sa description.

La préférence d'un mot M pour une préposition ou un auxiliaire donné (identifiés par leur lemme-catégorie) est calculée de la même manière, tout comme les types de dépendances récurrents (sujet, objet, ...), considérés comme des traits du mot gouverné, mais dans ce dernier cas plusieurs types de dépendances peuvent éventuellement être validés pour un même gouverné, si celui-ci a plusieurs gouverneurs (les quelques cas de figure sont décrits en 2.2.3.1). Le traitement des déterminants est légèrement différent : si un lemme-catégorie pour le trait 'déterminant' du mot M est significatif (par exemple *un_DET*), il est conservé ; sinon, si le

nombre d'occurrences d'un déterminant quelconque pour ce mot *M* est significatif, ce dernier se voit adjoindre un déterminant lexicalement sous-spécifié (*_DET*). Au contraire, si l'absence de déterminant est significative, comme dans *faire bonne figure*, une valeur spécifique (*NO_DET)* est utilisée. Enfin les mots rétablis (auxiliaires et déterminants) peuvent eux-même être caractérisés en termes de flexion et type de dépendance, toujours selon la même méthode.

2.4.2 Ordre(s) et distance entre mots

L'exigence est ici plus lâche : il ne s'agit pas d'identifier un ordre des mots significativement récurrent, mais plutôt les principales configurations linéaires possibles pour une même EPL. Certaines expressions admettent ainsi plusieurs ordres, par exemple avec et sans relative, à l'actif et au passif, ou encore avec et sans déterminant(s). Un seuil minimum de trois occurrences pour une configuration, ou 1/10 du nombre d'occurrences au-delà de 30, est requis afin de se prémunir d'éventuelles erreurs d'analyse.

Outre l'ordre des mots, chaque configuration est caractérisée par la distance maximale observée entre le premier et le dernier mot de l'expression, parmi les occurrences de cette configuration. Par précaution une fois encore la valeur extrême est écartée, ou les 1/10 les plus extrêmes au-delà de 10 occurrences de cette configuration.

En combinant ces informations aux préférences flexionnelles, on obtient des patrons morpho-lexicaux, qui permettent d'identifier ces EPL dans d'autres textes ayant simplement fait l'objet d'un étiquetage morphosyntaxique (lemmes-catégories et traits flexionnels), traitement plus léger et fiable qu'une analyse syntaxique complète, et pour lequel l'harmonisation des étiquettes entre différents systèmes est une tâche relativement simple. Par exemple, si la séquence *donner_V + un_DET{fem,pl} + sueur_N{pl} + froid_A{fem,pl}* apparaît dans une fenêtre maximale de 12 tokens (distance maximale observée après élimination des valeurs extrêmes), il est très vraisemblable qu'il s'agisse d'une occurrence de l'EPL *donner des sueurs froides.*

Il est cependant plus prudent de ne pas utiliser à cette fin les configurations linéaires peu fréquentes d'une EPL, par exemple une tournure passive si elle apparaît seulement 4 fois contre 35 occurrences de la même EPL à l'actif. La vraisemblance est moindre, et il peut également y avoir une divergence entre les traits caractéristiques de cette configuration, et les traits jugés significatifs pour l'expression en général. En l'état, ces configurations secondaires sont donc encore peu adaptées à l'analyse automatique.

Enfin ces patrons lexicaux, s'ils garantissent une bonne précision, peuvent cependant souffrir d'un manque de généralité. Des patrons plus abstraits (3.4.2.1), lexicaux-syntaxiques, voire dans certains cas lexicaux-sémantiques, plus productifs, devraient en particulier permettre d'identifier dans d'autres textes des EPL absentes du corpus ayant servi à l'extraction. On observe ainsi sur

les sorties des parallélismes évidents, comme *toutes classes confondues* / *toutes catégories confondues,* ou *une météo favorable* / *des conditions météorologiques favorables* / *des conditions climatiques favorables,* mais aussi l'absence de *tous genres confondus* ou *un climat favorable*, qui pourraient alors être inférés. La difficulté consiste à utiliser un niveau d'abstraction adéquat, sans surgénérer.

2.4.3 Génération d'une forme canonique

Une forme plus lisible de l'EPL est générée à partir de l'ordre des mots le plus fréquent et des préférences flexionnelles. Ici non plus, un test d'hypothèse n'est pas requis : les formes de surface apparaissant dans plus de la moitié des occurrences de l'EPL sont retenues, ce qui garantit l'accord en genre, nombre et personne. Si aucune forme ne remplit cette condition pour un mot, le lemme-catégorie est utilisé par défaut, mais avec une notation explicite (on utilisera par exemple *heureux_A* au lieu d'*heureux*). Quelques règles ad hoc ont par ailleurs été ajoutées, afin que ces formes se rapprochent de descriptions lexicographiques classiques : par exemple un verbe normalement conjugué sera écrit à l'infinitif, sauf s'il est précédé de son sujet (ce qui permet de générer *faire froid dans le dos* plutôt que *fait froid dans le dos*), et un déterminant récurrent mais sans forme ni lemme privilégié (*_DET*) sera par défaut remplacé par un article indéfini accordé avec le nom dont il dépend.

Ces principes simples permettent généralement de fournir des entrées intuitives, malgré quelques incohérences ponctuelles. Elles facilitent l'analyse des sorties, pour l'évaluation notamment, mais également afin d'ajuster les paramètres de l'extraction (seuils limites, ...). Par exemple, la forme *la bataille fait rage* est plus lisible que le patron correspondant :

> *le_DET{art,fem,sg,def} + bataille_N{fem,sg} + faire_V{ind,p3,sg,pres} + rage_N{sg,fem}*

et *a fortiori* plus lisible que le sous-graphe complet correspondant, reproduit en figure 7 :

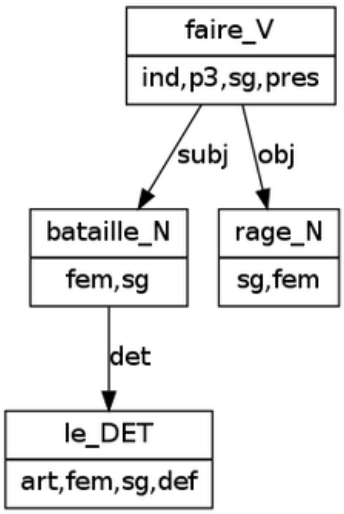


Figure 7 : Sous-graphe complet après la phase de caractérisation pour l'expression *la bataille fait rage*

3 ÉVALUATION, OBSERVATIONS ET PERSPECTIVES

3.1 ÉVALUATION

L'évaluation des EPL extraites est une tâche délicate, la définition d'un *gold standard* restant problématique. Lin (1999) utilise ainsi l'*English Idioms Dictionnary*, mais en reconnaissant que les expressions recensées peuvent varier d'un dictionnaire à l'autre. Villada Moiron (2005) utilise quant à elle deux lexiques constitués et enrichis manuellement à partir de dictionnaires de langues, mais pour des phénomènes très spécifiques (les locutions prépositionnelles d'une part, et les constructions à verbe support et syntagme prépositionnel d'autre part).

En français, le dictionnaire des cooccurrences de Beauchesne a lui aussi une couverture limitée, tandis que le *Dictionnaire des combinaisons de mots* de Le Fur s'intéresse essentiellement aux collocations. Le TLFI quant à lui est peu adapté pour un travail sur un corpus actuel de presse généraliste. Le dictionnaire *Antidote* des collocations de Druide (Charest et al., 2007 et 2010), constitué selon une méthode proche de celle utilisée ici, mais dont les EPL ont aussi fait l'objet d'une validation, voire d'une complétion manuelle, pourrait éventuellement être utilisé.

Cependant la légitimité de ce type de ressource peut également être questionnée. La frontière entre langue générale et langue de spécialité notamment est particulièrement difficile à établir. Même sur un corpus peu technique comme Emolex, de très nombreuses expressions sont caractéristiques de la presse généraliste, ou de la période étudiée (similaire pour les différentes sources journalistiques). Ce phénomène est renforcé par l'acception très large de la notion d'EPL adoptée ici. L'adjectif *amical* par exemple fait partie d'un nombre *a priori* surprenant d'EPL, qui relèvent pour nombre d'entre elles du domaine sportif (*jouer en amical, match amical, disputer une rencontre amicale* …), tandis que d'autres EPL reflètent clairement l'actualité (*les Centres Fermés d'Éducation, les déçus de la gauche/du sarkozysme, ...*). Or notre objectif est bien la mise au point d'une méthode portable d'extraction des EPL d'un domaine donné (ou d'un style, d'une situation de communication, ...), domaine dont le corpus utilisé est idéalement un échantillon représentatif (2.1.1). Un autre mode d'évaluation serait alors l'utilisation d'un corpus comparable, dont les EPL sont annotées, manuellement ou de façon semi-automatique. Malgré sa taille limitée, le *French Treebank* pourrait ainsi être utilisé afin d'évaluer le rappel de la méthode présentée ici.

La dernière option consiste à faire évaluer les expressions extraites par des locuteurs, linguistes ou spécialistes du domaine, sur une base plus ou moins intuitive, comme le font par exemple Evert et Krenn (2001). Là-aussi, les évaluations peuvent être très variables, selon le ou les critère(s) de validation adopté(s) : tournure consacrée, non compositionnalité sémantique, non-substituabilité paradigmatique, actant typique, morphosyntaxe, application visée (traduction, analyse, lexicographie...), etc. Par ailleurs l'accord inter-annotateurs, même linguistes, est en

général faible, faute de test objectif de la « polylexicalité » d'une construction. La plupart des EPL données jusqu'ici à titre d'exemple étaient volontairement peu ambiguës de ce point de vue, mais la réalité est beaucoup plus nuancée, avec de nombreuses décisions délicates, comme *dans des conditions abominables, trouver l'idée amusante, sous l'oeil bienveillant de, un milieu très fermé, froid et calculateur, se sentir frustré, profondément humain, un air dégoûté, etc.*

Plutôt que de proposer une évaluation en terme de rappel et précision forcément artificielle, pour les raisons qui viennent d'être évoquées, nous avons préféré, dans un premier temps du moins, comparer les sorties à celles d'une extraction par n-grammes contigus, menée en parallèle sur le même corpus, en demandant l'avis d'évaluateurs. Les n-grammes non contigus seraient ici une solution beaucoup plus lourde à mettre en œuvre, compte tenu des contraintes calculatoires vues en 1.4.1.2, à moins d'appliquer des filtres catégoriels prédéfinis, et/ou de se limiter à des couples cooccurrents, comme le fait Seretan (2008) dans une évaluation similaire.

Plus précisément, l'objectif est ici d'évaluer l'apport spécifique du traitement des inclusions et de l'analyse syntaxique. Ce dernier point a déjà été mis plusieurs fois en évidence pour des cooccurrences binaires (notamment par Seretan, *Ibid*), mais reste à démontrer pour les expressions plus longues, qui ont une tendance plus prononcée au figement linéaire.

3.1.1 Protocole

Le système utilisé en parallèle est le *MWE Toolkit* (distribué sous licence libre), et décrit dans Ramisch, Villavicencio et Boitet (2010). Il s'agit d'un outil robuste et largement paramétrable destiné à l'extraction d'EPL à partir de corpus étiquetés morphosyntaxiquement. Il permet notamment l'utilisation de patrons linéaires pour l'extraction de n-grammes non contigus, des requêtes Web pour les fréquences individuelles des mots, et l'intégration à un outil d'apprentissage supervisé, si un corpus annoté est disponible. L'expérience menée ici ne tire cependant volontairement pas pleinement parti de ses possibilités. La priorité a été donnée à l'harmonisation des conditions d'extraction entre les deux méthodes comparées, hormis les aspects évalués (l'analyse syntaxique et le traitement des inclusions).

50 des lemmes étudiés ont été sélectionnés aléatoirement. Les lemmes-catégories désambiguïsés par *Connexor* ont servi d'entrée à l'analyse par n-grammes. Tous les n-grammes contigus de 2 à 5 mots contenant un de ces lemmes ont été extraits, sans toutefois tenir compte des déterminants, afin de limiter l'éparpillement des données, et le nombre de doublons (*trou noir, un trou noir* et *le trou noir* pourraient être considérés comme trois expressions distinctes). Ils ont ensuite été rétablis manuellement pour chacune des expressions évaluées. Faute d'analyse syntaxique, les auxiliaires n'ont pas pu faire l'objet du même traitement, ni les

prépositions. Les pronoms personnels à l'indicatif non toniques (*je*, *tu* ,*il*,...), qui sont simplement ignorés (2.2.4.7) lors de l'extraction syntaxique (sans rétablissement lors de la phase de caractérisation), ont été également ignorés.

Pendant l'extraction, des filtres similaires à ceux déjà présentés ont été appliqués pour le nombre minimum d'occurrences (2.3.2.1) et la dispersion dans le corpus (2.3.2.2), et une mesure d'association comparable (2.3.2.3) a été utilisée afin de classer les candidats (l'information mutuelle spécifique, basée sur les fréquences marginales des lemmes, mais non ajustées). Après extraction, les n-grammes interrompus par un signe de ponctuation (généralement une virgule) ont été éliminés, afin toujours de réduire le bruit (les signes de ponctuation étant également absents des sous-graphes syntaxiques). Les variantes de la même expression (avec et sans auxiliaire, actif et passif...) ont été fusionnées. Enfin la forme prototypique de chaque expression a été rétablie manuellement.

L'extraction des candidats sur une base syntaxique a été menée dans les conditions décrites en 2.2.3 pour tous les sous-graphes de 2 à 6 mots, les sous-graphes de 6 mots n'étant pas retenus, mais simplement utilisés pour le dédoublonnement des occurrences incluses (2.3.1). Contrairement à l'extraction par n-grammes, des scores d'association minimums fixés empiriquement ont été utilisés pour chaque taille de sous-graphe, afin de permettre ce dédoublonnement. Enfin les déterminants, prépositions et auxiliaires ont été rétablis automatiquement(2.4.1), tout comme la forme prototypique de chaque construction (2.4.3), moyennant quelques corrections.

Pour chaque taille *n* d'expression (nombre de mots hors mots rétablis), les *x* expressions les mieux classées ont été retenues, *x* étant le plus petit nombre d'expressions de taille *n* trouvées par un des deux systèmes : par exemple, si 40 expressions de taille 3 ont été trouvées via l'extraction syntaxique, et 60 via l'extraction par n-grammes, seuls les 40 n-grammes les mieux classés ont été retenus. Les 108 expressions communes aux deux systèmes, essentiellement binaires, ont été écartées, afin de ne soumettre aux évaluateurs que les expressions trouvées par un seul d'entre eux. Au total, 116 expressions propres à chacun des modes d'extraction ont été évaluées.

5 évaluateurs, tous francophones natifs et possédant une formation linguistique, ont eu pour consigne, volontairement sous-spécifiée, d'identifier celles qu'ils estiment être des tournures consacrées (éventuellement compositionnelles sémantiquement), et/ou représentatives du corpus étudié (dont la composition leur a été communiquée), et/ou pertinentes dans le cadre de l'enseignement du français langue étrangère. Une phrase d'exemple pour chaque expression leur a également été donnée. Les constructions extraites par les deux systèmes ont été mélangées, afin que la comparaison ne soit pas biaisée. Les quatre réponses autorisées étaient *expression*

caractéristique, expression non caractéristique, expression caractéristique mais trop longue et *expression caractéristique mais incomplète*, les deux dernières n'étant pas exclusives. En annexe B se trouvent la consigne et les constructions à évaluer, ainsi que les constructions trouvées en commun par les deux systèmes.

3.1.2 Résultats et interprétation

L'accord inter-annotateur est de 0,46 (kappa de Fleiss), accord modéré dans l'absolu, mais correct pour ce type dévaluation, similaire par exemple à celui observé par Seretan (2008) dans une évaluation comparable pour des collocations binaires. Les catégories *incomplète* et *trop longue* ont été fusionnées pour le calculer, afin que les réponses soit bien exclusives. Le tableau 4 présente les avis cumulés des évaluateurs. Rappelons que seules les expressions non trouvées via une des deux méthodes ont été évaluées : il ne s'agit donc pas d'une évaluation de leurs précisions et rappels respectifs, mais simplement de leurs performances relatives.

	expression caractéristique		expression non caractéristique
	exacte	trop longue et/ou incomplète	
analyse n-gramme sans traitement des inclusions	244(42%)	198(34%)	138(24%)
analyse syntaxique et traitement des inclusions	348(60%)	112(19%)	120(21%)

Globalement, les résultats obtenus pour chacune des deux méthodes d'extraction sont significativement différents, compte tenu de la taille de l'échantillon (un test du X^2 d'indépendance donne une *p-value* < 0,01), mais la proportion de jugements « expression non caractéristique » reste similaire dans les deux cas (138/580 contre 120/580, soit une *p-value* de 0,23). Autrement dit, d'après cette évaluation, la seule différence significative entre les deux méthodes est la répartition des jugements « expression caractéristique », entre « exacte » et « trop longue et/ou incomplète » (avec là encore une *p-value* < 0,01).

Le tableau 5 permet alors d'observer en détail ces jugements « trop longue et/ou incomplète ». On y voit que le nombre de jugements « incomplète » reste proche pour les deux méthodes (64 contre 43 au total), mais que le nombre de jugements « trop longue » est nettement supérieur pour l'extraction par n-grammes (138 contre 71).

	incomplète	trop longue	incomplète et trop longue
analyse n-gram sans traitement des inclusions	60	134	4
analyse syntaxique et traitement des inclusions	41	69	2

TABLEAU 5 : Détail des jugements « expression incomplète » et/ou « expression trop longue »

Une analyse qualitative des réponses permet ensuite de formuler quelques hypothèses afin d'expliquer ces écarts. En voici la liste, classées par fréquence décroissante des phénomènes concernés :

- Les jugements « trop longue » pour l'extraction par n-grammes sont souvent dus à l'absence d'un traitement adéquat des prépositions, du type de celui présenté en 2.2.4.8. On trouve par exemple dans cette catégorie des constructions comme *un vif émoi dans* ou *d'affinités électives,* voire même plusieurs variantes pour une même EPL, comme *à briser le tabou, de briser le tabou, briser le tabou en* et *briser le tabou de.*

- Les jugements « trop courte » pour l'extraction par n-grammes peuvent en partie être imputés à l'absence de dédoublonnement des inclusions (2.3.1). On trouve ainsi **effondrer comme un château* (l'EPL complète étant *s'effondrer comme un château de cartes*), ou encore **jeter la rancune, *jeter la rancune à* et **la rancune à la rivière* (l'EPL complète étant *jeter la rancune à la rivière*).

- Les jugements « trop courte » pour l'extraction syntaxique cette fois-ci semblent en partie dus à l'absence de la préposition introduisant un complément sous-catégorisé obligatoire, comme dans *garder un souvenir ému* (où il manque la préposition de *de*). Ce point sera abordé plus en détail en 3.2.1. Il arrive aussi qu'un modifieur interne à l'expression (linéairement) et requis ne soit pas extrait, car son lemme varie . C'est le cas de l'adjectif dans :

  **jouir d'une immunité/impunité/popularité/réputation.*

- Pour l'extraction syntaxique toujours, certains jugements « trop longue » sont dus à la présence de conjonctions de coordination (*mais avoir douché les espoirs*). Un traitement particulier sera proposé en 3.2.2.

- Un autre cas intéressant de jugements « trop longue » pour l'extraction syntaxique est celui des actants typiques d'une EPL, jugés superflus, comme *affaire* dans *l'affaire avait suscité l'émoi.* Ces phénomènes nous semblent cependant pertinents pour des tâches

comme la traduction, la génération ou la désambiguïsation syntaxique. Pour l'extraction d'informations ou l'indexation en revanche, ils constituent plutôt du bruit : l'association de *affaire* et *susciter l'émoi* par exemple est compositionnelle sémantiquement. Enfin, dans une approche lexicographique classique, ils sont généralement ignorés. Une classification automatique adéquate des constructions extraites doit alors être envisagée, en suivant par exemple la méthode évoquée en 3.4.2.2.

Nous avons par ailleurs effectué une évaluation subjective des 108 expressions trouvées en commun par les deux systèmes. 95 d'entre elles nous ont semblé exactes, 2 trop longues, 7 trop courtes, et 4 non caractéristiques. Ces chiffres doivent cependant être pris avec précaution, en l'absence d'évaluateurs multiples.

3.2 PISTES POUR L'OPTIMISATION

3.2.1 Sous-catégorisation syntaxique

Comme l'a montré l'évaluation, un des traits manquants les plus évidents pour une caractérisation morphosyntaxique pertinente des EPL extraites est la sous-catégorisation, notamment lorsqu'elles contiennent un verbe (ou un déverbal). Par exemple, dans *se croire revenu, revenu* prend nécessairement un complément introduit par *à.* Mais le complément sous-catégorisé peut aussi dépendre d'un nom (voire d'un adjectif), comme dans *faire l'amère expérience **de**,* ou *s'apitoyer sur le sort **de**/s'apitoyer sur **son** sort.*

Or, pour de nombreux parseurs en dépendances, dont *Connexor*, la préposition dans un syntagme prépositionnel est rattachée à la tête (nominale, verbale) du syntagme. Par conséquent elle est ici rattachée au complément sous-catégorisé, et ce dernier n'est pas extrait lorsque son lemme peut varier (*se croire revenu au temps*/*au moment*/*à l'époque*/*au début...*). La préposition n'est donc pas identifiée comme faisant partie de l'EPL.

Une solution consisterait à faire de la préposition un des traits du mot régisseur, adoptant un traitement proche de celui des déterminants (2.2.4.2), afin d'identifier après extraction la ou les prépositions récurrentes pour un même régisseur. Afin d'éviter du bruit cependant, il peut être utile de s'appuyer sur les types des dépendances syntaxiques concernées, bien que les erreurs d'analyse syntaxique soient fréquentes sur ce point (la distinction entre complément circonstanciel et complément d'attribution par exemple est difficile à établir).

3.2.2 Extension de la liste des mots ignorés

Plusieurs classes de mots outils ont été conservées par défaut pendant la phase d'extraction. Il s'agit notamment des négations, de nombreux pronoms (dont les réfléchis), ou encore de certains adverbes (*plus, moins,* voire *très*...). Ces mots font donc partie des sous-graphes initialement extraits, au même titre que les mots appartenant à des classes ouvertes, et contrairement aux déterminants (2.2.4.2) et prépositions (2.2.4.8), qui sont rétablis après la phase d'extraction, à partir des occurrences des sous-graphes retenus (2.4.1).

Par exemple, au moment de l'extraction, l'EPL *ne pas y croire un seconde* correspond au sous-graphe représenté par la figure 8. Seul le déterminant est rétabli après la phase d'extraction. Dans quelques cas cependant, cette approche conduit à l'extraction de constructions très proches, lorsque plusieurs configurations autonomes sont avérées, comme *réunir en session extraordinaire* et *se réunir en session extraordinaire.* Deux points de vue peuvent alors être

adoptés : le premier considérera qu'il s'agit d'expressions distinctes, mais présentant un rapport d'inclusion, comme *disque dur* et *disque dur externe.* Le second, qui nous semble plus intuitif, considérera qu'il s'agit de différentes formes de la même expression. Il faut alors prévoir un traitement adéquat, du type de celui déjà adopté pour les déterminants et prépositions.

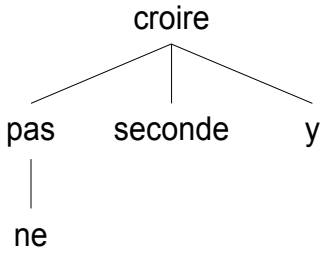


FIGURE 8 : Sous-graphe pour *ne pas y croire une seconde* au moment de l'extraction des candidats

Le cas des conjonctions de coordination est *a priori* différent, mais un traitement similaire peut être appliqué : la conjonction peut être ignorée pendant l'extraction, et considérée comme un simple « trait » du premier élément coordonné, par exemple un trait de *faits* dans *faits et gestes*. Si tous les éléments coordonnés font bien partie de l'expression extraite (ici le mot *faits* et le mot *gestes*), la conjonction peut alors être rétablie pendant la phase de caractérisation, éventuellement sous-spécifiée en l'absence de préférence marquée pour un lemme donné (comme dans l'expression *cruel, inhumain et/ou dégradant*).

3.2.3 Analyse syntaxique

La fiabilité de l'analyse syntaxique conditionne nécessairement les résultats. Quelques erreurs récurrentes propres à l'analyseur utilisé pourraient cependant être corrigées, grâce à un traitement ad hoc. On observe par exemple plusieurs cas de doublons pour des expressions de type V N Adj, lorsque le verbe admet un attribut de l'objet. *Laisser un goût amer* donne ainsi lieu à l'extraction de deux constructions jugées distinctes : dans le premier cas, l'adjectif est rattaché au nom (épithète), tandis que dans le second, il est rattaché au verbe (attribut de l'objet). Pour d'autres EPL, la structure syntaxique sous-jacente est fausse, mais, cette erreur étant systématique, l'EPL a été correctement extraite (du moins sa forme prototypique). C'est le cas par exemple de *rendre un avis favorable* (où *favorable* est systématiquement analysé comme attribut de l'objet et non comme épithète), ou de constructions plus complexes syntaxiquement, comme *aussi bizarre que cela puisse paraître.*

Par ailleurs quelques post-traitements supplémentaires par rapport à ceux décrits en 2.2.3.1 pourraient être appliqués aux sorties de l'analyseur, non pas pour corriger des erreurs, mais pour obtenir une analyse plus adaptée à l'extraction d'EPL. Ainsi certains pronoms (*beaucoup de, ...*), voire noms collectifs (*une tonne de,...*) gagneraient à être analysés comme des déterminants complexes, afin d'autoriser par exemple le rapprochement entre *susciter l'émoi* et *susciter beaucoup d'émoi* (qui est analysé par défaut comme *susciter_V* → *beaucoup_PRON* → *émoi_N* → *de_PREP*).

Une autre type d'ajustement pertinent serait la prise en compte d'informations dérivationnelles lors de l'extraction des sous-graphes candidats, afin notamment de rapprocher un verbe et son déverbal. Par exemple *la remise au goût du jour* et *remettre au goût du jour* sont analysés par défaut comme deux EPL distinctes. Le fait de ne pas prendre en compte les types des dépendances (sujet, objet...) lors de la phase d'extraction des candidats (2.2.4.4) serait ici un avantage (on a par exemple ici la même chaîne de dépendances non typées *remettre/remise* → *goût* → *jour*).

3.2.4 Optimisation algorithmique

La priorité a été donnée pendant l'implémentation à la souplesse du paramétrage et à la lisibilité des sorties intermédiaires, ce qui a facilité les ajustements linguistiques sur une base empirique, décrits en 2.2.4. L'efficacité des algorithmes et structures de données choisis s'en est parfois trouvée limitée. La réduction du temps de traitement est donc une piste d'amélioration : si elle est suffisante pour autoriser un comptage exhaustif des sous-graphes du corpus, des mesures d'associations plus exactes que celle décrite en 2.3.2.3 pourraient être utilisées (en abandonnant notamment l'hypothèse des tirages indépendants).

Ce coût pourrait notamment être réduit en tenant compte des fréquences individuelle des lemmes-catégories des mots de la phrase cours au moment de l'extraction des candidats. Par exemple, si le mot *chrysanthème* n'apparaît que 4 fois dans le corpus, il est inutile (compte tenu des seuils minimums d'occurrences donnés en 2.3.2.1) d'extraire comme candidats les sous-graphes de tailles 2, 3 et 4 qui le contiennent. Une part importante des sous-graphes contenant des noms propres pourrait ainsi vraisemblablement être écartée d'emblée. Une autre possibilité pour réduire le coût du comptage serait l'utilisation d'une représentation linéaire des candidats plus concise que celle présentée en 2.2.2.1.

3.3 OBSERVATION DES EPL EXTRAITES

8648 constructions de tailles 2 à 5 ont été extraites à partir des 1201 lemmes étudiés. Le score d'association minimum pour les constructions de taille 2 est relativement élevé, privilégiant la précision au rappel. Pour les constructions de tailles 3 à 5 en revanche, ce type de paramétrage n'est pas autorisé, car un score minimum trop élevé ou trop faible pour une taille donnée a une incidence directe sur les constructions obtenues pour les tailles inférieures, du fait du traitement des constructions incluses (2.3.1). Les sorties permettent d'observer quelques phénomènes intéressants.

3.3.1 Pertinence des rapports d'inclusion entre EPL

C'est un des phénomènes les plus évidents. Après dédoublonnement des occurrences incluses (2.3.1), les rapports d'inclusion restants (c'est-à-dire entre EPL validées par le système) sont un critère de regroupement particulièrement intuitif. Il est fréquent d'observer une construction qui, bien qu'apparaissant par ailleurs de façon autonome, est incluse dans une ou plusieurs construction(s) plus longue(s), comme *une session extraordinaire* et *convoquer en session extraordinaire.* Afin d'étudier plus en détail ce phénomène, nous avons observé 800 cas d'inclusions choisis aléatoirement (après élimination manuelle des erreurs d'extraction), pour lesquels nous avions une évidence, c'est-à-dire pour lesquels l'EPL incluse (*session extraordinaire*) contient un des mots du lexique étudié (*extraordinaire*).

Trois cas de figure peuvent être distingués. Le plus fréquent (476 cas sur 800) est celui des EPL « récursives », c'est-à-dire composées d'un (voire plusieurs) mot(s) et d'une EPL autonome par ailleurs, comme *une logique de guerre froide*, qui se décompose en *une logique* + [*guerre froide*] (et non pas **une logique de guerre* + *guerre froide*). La relation syntaxique ou sémantique entre l'EPL incluse et le (ou les) élément(s) complémentaire(s) est très variable :

verbe support d'un groupe nominal : *sueurs froides* / *donner des sueurs froides,*

verbe support d'un adverbe complexe : *cul sec* / *boire cul sec,*

nom régisseur : *ressources humaines* / *directeur des ressources humaines,*

nom support de l'adjectif : *peu enviable* / *un sort peu enviable,*

épithète : *un disque dur* / *un disque dur externe*

adverbe : *un trafic perturbé* / *un trafic fortement perturbé,*

circonstant *: semer la confusion* / *semer la confusion dans les esprits*

modal : *garder la tête froide* / *savoir garder la tête froide,*

actant typique : *faire froid dans le dos* / *les chiffres font froid dans le dos,*

« argument typique » : *la politesse du désespoir* / *l'humour est la politesse du désespoir,*

etc.

Le second cas de figure le plus fréquemment observé (280 cas sur 800) est la combinaison de deux (voire trois) EPL partageant un mot plein, comme *poursuivre son irrésistible ascension,* qui est vraisemblablement une combinaison de *poursuivre son ascension* + *l'irrésistible ascension* (bien que nous n'ayons ici aucune évidence de l'autonomie de *poursuivre son ascension*, puisque seules les expressions contenant *irrésistible* ont été extraites). De la même manière, *un trafic de drogues dures* peut vraisemblablement être décomposé en *trafic de drogues* + *drogues dures.* On peut encore mentionner *faire une dépression nerveuse* (*faire une dépression* + *dépression nerveuse*)*, un curieux mélange des genres, une situation extrêmement préoccupante, un établissement d'enseignement supérieur*, *aborder les sujets qui fâchent* ou *durement frappé par la crise* .

Enfin le troisième cas de figure, plus rare (44 cas sur 800), est celui des combinaisons de deux EPL sans chevauchement. Par exemple *faire l'effet d'une douche froide* se décompose en *faire l'effet de* + *une douche froide.* C'est aussi le cas de *une société de conseil en ressources humaines,* qui se décompose en *une société de conseil* + *ressources humaines,* ou encore *plaider non coupable des chefs d'accusation* (*plaider non coupable* + *chefs d'accusation*). On pourrait aussi voir dans ce dernier exemple une combinaison avec chevauchement (*plaider non coupable* + *coupable des chefs d'accusation*), mais ce n'est pas la combinaison qui a été extraite.

Il est également fréquent d'observer plusieurs niveaux d'inclusion entre EPL, comme dans l'exemple suivant :

à guichets fermés
donner à guichets fermés
jouer à guichets fermés
se jouer à guichets fermés
la rencontre se joue(ra) à guichets fermés

Cet exemple illustre par ailleurs la question soulevée en 3.2.2. Il serait probablement plus pertinent ici de ne pas distinguer *jouer à guichets fermés* et *se jouer à guichets fermés*, ce qui donne l'arborescence suivante :

à guichets fermés

donner à guichets fermés

jouer à guichets fermés

la rencontre se joue(ra) à guichets fermés

Il faut enfin préciser que toutes les EPL de plus de deux mots pleins ne peuvent évidement pas être décomposées, certaines d'entre elles n'incluant aucune expression de taille inférieure. C'est le cas de *l'abominable homme des neiges* ou *faire amende honorable.*

3.3.2 Compositionnalité sémantique et inclusion

Dans la plupart des exemples d'inclusion qui viennent d'être mentionnés (comme *donner* + *des sueurs froides*), il semble que l'EPL incluante (*donner des sueurs froides*) soit régulière sémantiquement, c'est-à-dire que son sens peut être déduit de la combinaison du sens de l'EPL incluse (*sueurs froides*) et de celui du mot (ou de l'EPL) complémentaire (*donner*). Cette règle reste apparemment valable lorsque le complément est polysémique, et employé dans un sens second. Par exemple *donner* dans *donner à guichets fermés* n'a pas le même sens que dans *donner des sueurs froides* (ni le même sens que dans *donner un euro*), mais il s'agit tout de même d'un des sens avérés du mot *donner,* que l'on retrouve par exemple dans *donner une représentation/un spectacle/un concert...* Par conséquent, on peut aussi considérer *donner* + *à guichets fermés* comme une composition sémantiquement régulière, tout comme *climat* + *guerre froide,* car ce sens de *climat* est avéré par ailleurs (*le climat est tendu, ...*).

Nous avons testé cette hypothèse sur les 800 cas d'inclusions présentés en 3.3.1. Malgré l'inévitable subjectivité du jugement de compositionnalité, l'hypothèse semble globalement valide, avec en tout 796 cas réguliers sur 800.

Pour les EPL récursives, la règle est vérifiée dans 473 cas sur 476. Les seules exceptions que nous ayons trouvées sont les suivantes :

- *dormir à poings fermés* (dont le sens n'est pas équivalent à celui de *dormir* + *les poings fermés*),
- *un bruit de verre brisé* (dont le sens n'est pas équivalent à celui de *bruit* + *verre brisé,* car *verre brisé* désigne ici le procès, par métonymie),
- *avancer en terrain miné* (différent de *avancer* + *en terrain miné*, car ce sens d'*avancer* n'est pas avéré par ailleurs),

En ce qui concerne les combinaisons de deux EPL partageant un mot plein (comme *poursuivre son irrésistible ascension*), dans 279 cas sur 280, on peut déduire le sens de l'EPL complète par récursion. Les deux compositions sont parfois possibles (par exemple, on peut obtenir le sens de

susciter un vif émoi à partir de *susciter l'émoi+ vif*, ou bien à partir de *susciter + un vif émoi*), ou alors une seule d'entre elles (par exemple on peut obtenir le sens de *trafic de drogues dures* à partir de *trafic + drogues dures*, mais pas à partir de *trafic de drogues + dures* , car ce sens de *dur* n'est pas avéré par ailleurs). La seule exception que nous ayons trouvée est *tomber en panne sèche*.

Enfin les 44 combinaisons de deux EPL sans chevauchement (comme *faire l'effet d'une douche froide*) sont toutes régulières. Autrement dit on peut obtenir le sens de l'EPL complète à partir du sens des deux EPL incluses (*faire l'effet de + une douche froide*).

3.4 PERSPECTIVES

3.4.1 Exploitation immédiate

3.4.1.1 Lexicographie

En l'état, les données extraites peuvent déjà être utilisées en tant qu'aide au travail lexicographique. Un exemple de description obtenue pour l'expression *se réduire comme peau de chagrin* est donné en annexe A. Ces informations sont un bon complément de celles fournies par un outil comme le *Sketch Engine* (Kilgariff, 2004), qui permet d'étudier le profil combinatoire d'un mot et ses quasi-synonymes. Elles offrent quelques facilités supplémentaires pour l'étude des EPL contenant ce mot.

- Les EPL n-aires, comme *l'abominable homme des neiges,* sont extraites automatiquement, alors qu'elles doivent être reconstituées lorsque les cooccurrents du mot étudié sont présentés individuellement (extraction d'EPL binaires). Il faudrait ici identifier une hypothétique cooccurrence binaire remarquable entre *abominable* et *homme* ou *homme* et *neige*, puis compléter l'expression en observant les phrases contenant ce couple.
- L'organisation par inclusion des EPL valides est relativement intuitive. Les EPL de plus de deux mots pleins sont rattachées aux EPL qu'elles incluent syntaxiquement. Par exemple *laisser un goût amer* et *le goût amer de la défaite* sont toutes deux rattachées à *un goût amer.*
- Une forme lisible est générée (2.4.3), avec détermination et flexions prototypiques notamment, ce qui facilite grandement la reconnaissance de l'EPL, et ce malgré quelques erreurs ponctuelles de génération.
- Au delà de cette forme prototypique, une description détaillée et statistiquement significative de l'expression est également produite : présence ou absence de déterminant, prépositions, types des dépendances, flexions, etc. Il manque encore une caractérisation complète des différentes alternances attestées, mais certaines d'entre elles peuvent déjà être déduites des patrons morpho-lexicaux générés (2.4.2) : adjectif antéposé et postposé, passivation, relativisation, non contiguïté (grâce à la distance maximale observée), etc. On trouve par exemple deux configurations privilégiées pour *une politesse exquise* (adjectif antéposé ou postposé), ou pour *entretenir une relation amoureuse* (avec et sans relativisation).

3.4.1.2 Désambiguïsation syntaxique

Les patrons générés lors de la phase de caractérisation morphosyntaxique (2.4.2) constituent un indice fort pour la reconnaisance de l'EPL concernée dans d'autres textes. Ils peuvent donc être utilisés afin de contraindre une analyse syntaxique en dépendance, à condition que l'analyse initiale de l'expression soit juste, et que les deux analyseurs aient des jeux d'étiquettes comparables. Voici par exemple le (seul) patron obtenu pour *faire l'effet d'une douche froide,* avec une distance maximale observée de 7 tokens entre le premier et le dernier mot (c'est à dire autorisant ici l'insertion d'un token) :

faire_V + le_DET{msc,art,def,sg} + effet_N{msc,sg} + de_PREP +
un_DET{indef,art,sg,fem} + douche_N{sg,fem} + froid_A{sg,fem}

Si toutes ces conditions sont remplies (lemmes-catégories, flexions, ordre des mots et distance maximale), il est très vraisemblable que nous soyons en présence de l'EPL *faire l'effet d'une douche froide*, et donc également du réseau de relations syntaxiques (typées) extrait pour cette expression. Ceci est également valable pour des expressions admettant un nombre supérieur de mots insérés. La distance maximale observée pour *se heurter à des portes fermées* est ainsi de 11 tokens, même après élimination par précaution de la valeur extrême (2.4.2). Les patrons alternatifs (variation d'ordre des mots en cas de relativisation par exemple, ou de passivation) peuvent eux aussi être exploités, au-delà d'un certain seuil d'occurrences. Intuitivement, il s'agit d'une analyse syntaxique guidée par la reconnaissance d'unités complexes préfabriquées et lexicalisées. Laurent et al. (2009) décrivent un traitement de ce type pour améliorer les performances de l'analyseur syntaxique *Cordial*, à partir de cooccurrences binaires remarquables.

3.4.2 Approfondissements

3.4.2.1 Abstraction supplémentaire des patrons

La dispersion des occurrences (1.4.1.1) limite en pratique la couverture de l'extraction. Une solution consiste alors à extraire des patrons plus abstraits, et donc plus productifs, en combinant critères lexicaux, morphosyntaxiques, voire sémantiques, afin de détecter dans d'autres textes des EPL absentes du corpus utilisé (ou présentes mais en quantité insuffisante pour avoir été extraites). Détecter ces EPL non répertoriées peut par exemple constituer une aide à l'analyse syntaxique, en particulier pour les constructions exocentriques (voir 1.2.3.2), ou encore à l'indexation, en présumant la non compositionnalité sémantique de la construction identifiée.

La principale difficulté ici consiste à ne pas surgénérer. Certains patrons syntagmatiques traditionnellement utilisés lors de l'extraction sont beaucoup trop productifs (N Prep N, Adj N, Adv Adj), tandis que d'autres, comme N N, ou V N (sans déterminant) semblent plus spécifiques aux EPL, quoique encore trop abstraits. Un niveau intermédiaire d'abstraction entre patrons fortement lexicalisés, comme ceux déjà extraits, et patrons morphosyntaxiques semble donc plus approprié. Green et al. (2011) font ainsi émerger par apprentissage supervisé, à partir du *French Treebank*, des patrons comme *coup de N, V de N* ou *au N de.* L'apprentissage est une solution vraisemblablement plus réaliste ici qu'une élaboration manuelle, compte-tenu de la variété des combinaisons possibles. Les patrons obtenus semblent en revanche peu adaptés à l'identification d'EPL disjointes.

3.4.2.2 Compositionnalité sémantique

Il s'agit ici de répondre au second point soulevé en introduction, à savoir le rattachement (ou non-rattachement) des EPL aux mots ou EPL qu'elles incluent, sur un critère sémantique. Ceci correspond à la distinction faite par Mel'čuk entre d'une part les phrasèmes non compositionnels sémantiquement, comme *casser les pieds,* qui doivent selon lui faire l'objet d'une entrée lexicographique distincte, et d'autre part les phrasèmes compositionnels (les collocations), comme *proprement scandaleux,* qui doivent être rattachées à leur base. La base (*scandaleux*) désigne dans une collocation binaire le mot choisi librement (en situation d'énonciation), qui contraint lexicalement le choix du collocatif (*proprement*).

D'un point de vue plus pratique, ce rattachement est nécessaire pour des tâches comme l'indexation de documents, ou l'extraction d'information. Considérer l'EPL *témoignages accablants* comme une entrée lexicale distincte de celle de *témoignage* peut par exemple nuire à l'analyse : aucun rapprochement ne sera fait entre *témoignage* dans *les témoignages de ses proches* et *témoignage* dans *des témoignages accablants*. Il est donc plus intéressant d'analyser *témoignages accablants* comme une combinaison sémantiquement régulière, quoique consacrée. A l'inverse, pour les expressions non compositionnelles sémantiquement, comme *jeter l'éponge* ou *cousu de fil blanc*, cela générerait évidemment des contresens.

L'objectif est donc d'obtenir une organisation du lexique sur le modèle suivant (où les entrées lexicales sont en gras) :

témoignage

des témoignages accablants

éponge

jeter l'éponge

guichet

tenir le guichet

à guichets fermés

donner à guichets fermés

jouer à guichets fermés

la rencontre se joue(ra) à guichets fermés

Les mots et EPL en gras peuvent être considérés comme des unités de sens autonomes, tandis que les EPL en italique sont des associations privilégiées, plutôt destinées à la traduction automatique, la génération ou la désambiguïsation.

Différentes méthodes peuvent être envisagées pour obtenir automatiquement ce type d'organisation des EPL extraites. Le figement morphosyntaxique est une première piste, de par sa corrélation souvent admise avec la non-compositionnalié sémantique. Villada Moiron (2005) justifie ainsi son étude de la variabilité morphosyntaxique de constructions à verbe support. Par exemple *l'éponge qu'il a jetée, l'éponge a été jetée* ou *jeter les éponges* n'ont pas le même sens que *jeter l'éponge*, et on ne peut pas dire **les devants qu'elle a pris,* ni **prendre le devant.* Il est cependant facile de trouver des contre-exemples (*les plâtres qu'elle a essuyés*, *le sucre qu'il lui a cassé sur le dos...*). Surtout, compte-tenu de la variété des phénomènes concernés (ordre des mot, insertions,…), ce critère semble difficile à modéliser, comme cela a déjà été expliqué en 1.3.2.3.

Une alternative pour décider de la compositionnalité d'une EPL est alors l'étude sémantique de son contexte d'apparition. L'analyse sémantique vectorielle (Manning et Schütze, 1999:296) est une des pistes envisageables. Par exemple, en observant quelques-unes des phrases du corpus Emolex contenant *à guichets fermés* (en tant qu'EPL autonome), on voit clairement apparaître un paradigme sémantique (*concert, prolongation, pièce, musique, match, scène, représentation, tournée,...*) qui devrait permettre de distinguer le sens de l'EPL de celui de *guichet* (et *a fortiori* de celui de *fermé*). Une analyse vectorielle robuste peut alors être envisagée, à l'échelle du paragraphe et non de la phrase, basée par exemple sur des requêtes Web. Les règles de composition vues en 3.3.2 peuvent ensuite être appliquée pour traiter les expressions incluantes (*jouer à guichet fermé*, *donner à guichet fermé,* …). Une des difficultés de ce type d'approche reste cependant la prise en compte de l'éventuelle polysémie des mots simples inclus. La

polysémie des EPL quant à elle semble être un phénomène plus marginal. Et même lorsqu'il y a polysémie (*passer l'éponge, sortir de ses gonds, laisser un goût amer*), le sens propre est souvent cantonné à un usage bien spécifique.

CONCLUSION

Nous avons montré au cours de ce travail qu'il est possible d'extraire automatiquement des mots et expressions composés avec une précision à priori satisfaisante à partir d'un corpus analysé syntaxiquement, en utilisant pour critère la simple cooccurrence significative des mots de l'expression, et surtout sans limiter la taille des expressions extraites à deux mots, ni utiliser de patrons prédéfinis, comme *N de N* ou *V N prep N*.

L'avantage d'une telle méthode est double. Tout d'abord la qualité de l'extraction est sensiblement supérieure à celle obtenue via une extraction par cooccurrence linéaire, comme le montre l'évaluation comparative que nous avons menée avec un système d'extraction de n-grammes contigus, utilisé en parallèle (3.1). L'extraction sur un base syntaxique telle que nous l'avons implémentée permet en particulier d'écarter des constructions trop courtes, comme **remuer le couteau* et **toute allure,* ou trop longues comme **à briser le tabou* et **un vif émoi dans*. Les premières sont écartées grâce à un traitement spécifique des rapports d'inclusion entre constructions récurrentes (2.3.1), et les secondes grâce à des ajustements linguistiques propres à la langue étudiée, mais basés sur l'analyse syntaxique des constructions candidates (2.2.4).

L'autre avantage de cette méthode est la possibilité d'obtenir une description morphosyntaxique précise des EPL extraites : relations syntaxiques sous-jacentes, préférences flexionnelles, variations en terme d'ordre et distance entre mots, présence ou absence de déterminants, etc. Ceci facilite la constitution de lexiques riches, et donc relativement portables, utilisables dans divers cadres applicatifs (analyse de surface ou plus profonde, traduction, voire génération), c'est-à-dire non orientés par la tâche, mais dépendant simplement du domaine étudié. C'est aussi une étape supplémentaire vers l'automatisation de l'encodage de ces EPL, tâche qui est encore souvent effectuée manuellement.

Enfin l'étude des phénomènes d'inclusion entre EPL (comme *coup d'œil* et *jeter un coup d'œil*) est particulièrement intéressante. Elle permet tout d'abord d'identifier automatiquement les préférences d'une expression en termes de verbe support, compléments typiques ou actants typiques, étendant parfois la gamme des phénomènes traditionnellement pris en compte en phraséologie. Dans le corpus étudié, l'EPL autonome *faire froid dans le dos* par exemple a parmi ses emplois consacrés la tournure *les/des chiffres (qui) font froid dans le dos,* tandis que l'EPL autonome *une fusée de détresse* peut prendre pour verbe support *tirer*. Par ailleurs l'exploitation de ces phénomènes d'inclusion entre EPL (et bien sûr entre EPL et mots simples) facilite leur organisation sur une base sémantique (3.4.2.2).

ANNEXES

Annexe A. Exemple de description obtenue pour une EPL

- **chagrin_N**
  - **peau de chagrin** + ++ +++
    - **réduire_V à peau de chagrin** + ++ +++
      - **se réduire comme peau de chagrin** - -- ---

Dépendances

réduire_V -> le/se_PRON ^^ réduire_V -> peau_N -> chagrin_N

Exemples

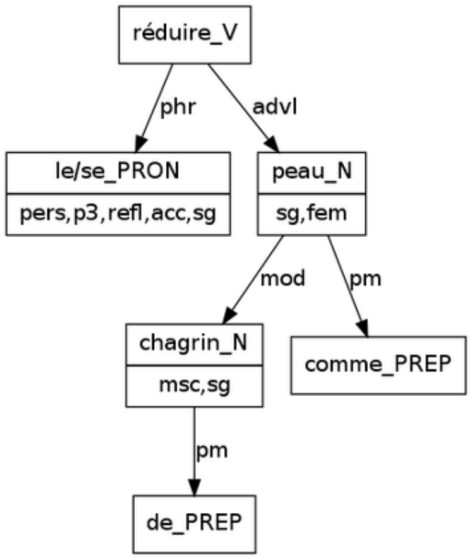
Il a dû se reconstruire», confie un autre membre de ce cénacle qui **s'**est **réduit** comme **peau** de **chagrin**.
« Sur le marché du zinc, par exemple, le déficit entre l'offre et la demande devrait **se réduire** comme **peau** de **chagrin**, passant de 420 000 à 100 000 tonnes cette année», estime Greig Gailey, directeur exécutif du numéro deux mondial, l'australien Zinifex.
Leurs programmes ne cessent de **se réduire**, au fil des jours, comme **peau** de **chagrin**.
» Développements :« Quand on a des ministres qui humilient la France en baragouinant l'anglais à Bruxelles, il ne faut pas s'étonner que l'ex-première langue du monde **se réduise** comme **peau** de **chagrin**.
Si bien que la Rue de Valois, qui, au milieu des années 1970, finançait seulement un Palais Garnier et une BNF, doit faire fonctionner aujourd'hui tant d'établissements qui ponctionnent mécaniquement son budget, que la marge de manoeuvre d'un nouveau ministre **s'**est **réduite** comme **peau** de **chagrin**.

Dépendances et traits privilégiés

```
réduire_V -> le/se_PRON{pers,p3,refl,acc,sg} ^^ réduire_V ->
peau_N{sg,fem,prep:comme_PREP,det:NO_DET} ->
chagrin_N{prep:de_PREP,msc,sg,det:NO_DET}
```

Occurrences : 69

Mots pleins : 4



Ordre privilégié

```
le/se_PRON{pers,p3,refl,acc,sg} + réduire_V + comme_PREP
+ peau_N{sg,fem} + de_PREP + chagrin_N{msc,sg}
```

Occurrences : 69 / 69

Maxspan : 6

le nombre se réduit comme peau de chagrin + ++ +++

rétrécir comme peau de chagrin + ++ +++

ANNEXE B. ÉVALUATION

Consigne donnée aux évaluateurs

Il s'agit de comparer deux méthodes d'extraction automatique d'expressions composées, sur un corpus constitué à 85 % d'articles de journaux, et à 15 % de textes littéraires. Les expressions trouvées en commun par les deux outils d'extraction ont été écartées, vous n'avez dans le document joint que les expressions trouvées par un seul d'entre eux ; le nombre d'erreurs peut donc être important.
Pour chaque expression, il faut indiquer si elle vous semble être une tournure consacrée (en général, ou bien compte tenu de la nature du corpus). Par exemple *jeter un coup d'œil* me semble être une expression consacrée. Une phrase du corpus est donnée pour chaque expression à titre d'exemple.
L'acception est ici très large, il peut s'agir de simples associations de mots qui intuitivement vous semblent caractéristiques, ou bien présentent un intérêt dans le cadre de l'enseignement du français langue étrangère. Le sens peut être totalement transparent, ou plus métaphorique, cela n'a pas d'importance. Voici quelques expressions qui me paraissent correspondre à cette définition : *une canne à pêche, prendre des proportions alarmantes, un conseil avisé, donner son avis, c'est la moindre des choses, en avance sur son temps...*
Les réponses possibles sont :

- expression caractéristique/consacrée/pertinente : une expression valide peut éventuellement contenir une autre expression valide, cela n'a pas d'importance : par exemple *jeter un coup d'œil* contient *coup d'œil*, et toutes les deux me semblent valides.
- expression caractéristique mais incomplète : par exemple l'expression **jeter un coup* est a priori incomplète.
- expression caractéristique mais trop longue : par exemple **faire ses valises et*
- association de mots quelconque.

Une expression peut éventuellement être à la fois trop longue et trop courte, par exemple **de jeter un coup.* C'est normalement le seul cas de figure où plusieurs réponses sont possibles pour une même expression.
Quelques précisions :

- Les flexions (y compris verbales), les auxiliaires et les déterminants n'ont aucune importance : ils sont simplement là pour faciliter la lecture ; vous pouvez donc les supprimer, ou en imaginer d'autres.
- Tous les autres mots de l'expression doivent être présents, notamment les négations, les pronoms réfléchis ou les prépositions : par exemple **pas y aller de main morte* est incomplet (d'après moi en tous cas), car il manque le *ne*, tout comme **brosser les dent,* car il manque le pronom réfléchi *se.*

Enfin, ne vous posez pas trop de questions (inutile de faire des tests savants de substitution, déplacement, etc...), il s'agit essentiellement d'intuition ici, le tout ne devrait pas prendre plus de 30 mn.

Merci beaucoup

Constructions soumises aux évaluateurs

Les deux listes qui suivent contiennent les constructions évaluées, c'est-à-dire trouvées par un seul des deux systèmes. Lors de l'évaluation, elles ont été fusionnées et « anonymisées », afin de ne pas influencer le jugement. Nous les présentons ici de façon distinctes pour faciliter l'appréciation des erreurs commises par chacun des deux systèmes. Faute de place, les phrases d'exemple données aux évaluateurs n'ont pas été reproduites.

Voici tout d'abord les constructions trouvées uniquement via l'extraction par n-grammes :

aucune affinité avec
d'affinités électives
les vertus apaisantes de
calmer les ardeurs de
doucher les ardeurs de
freiner les ardeurs de
l'ardeur réformatrice de
modérer les ardeurs de
refroidir les ardeurs de
tempérer les ardeurs de
balancer d'avant en arrière
balancer doucement
balancer les jambes
balancer mes cendres
balancer mes cendres sur
balancer Nicolas Sarkozy
balancer par la fenêtre
balancer pas mal
et balancer mes cendres
se balancer au-dessus de
se balancer sur une chaise
son cœur balance entre
à briser le tabou
briser comme du verre
briser l'isolement
briser la colonne
briser la glace
briser la vitre et
briser le monopole
briser le silence
briser le silence sur
briser le tabou de
briser le tabou en
briser le tabou et
briser les barrières
briser les os
de briser le carcan
de briser le tabou
en brisant la vitre
lui briser le cœur
culpabilité civile
de la culpabilité d'Yvan Colonna
la culpabilité civile d'un malade mental
la culpabilité d'Yvan
la culpabilité du condamné
la culpabilité ou l'innocence de
prouver sa culpabilité
déstabiliser le Liban
déstabiliser le Pakistan
déstabiliser le régime
déstabiliser Nicolas Sarkozy
déstabiliser ses adversaires
déstabiliser un peu plus
pour déstabiliser le régime
pour déstabiliser ses adversaires
doucher l'enthousiasme de
doucher l'espoir de
un gel douche
effondré en larmes
effondré sur la banquette
effondré sur le sol
effondrer comme un château

effondrer comme un château de
effondrer comme un château de cartes
effondrer dans les sondages
effondrer en Bourse
effondrer en larmes dans
après l'émoi suscité
après l'émoi suscité par
l'émoi provoqué par
l'émoi suscité par
un vif émoi à
un vif émoi dans
un souvenir ému de
exprimer sa frustration
la frustration est d'autant plus
honorer l'échéance
honorer la mémoire de
honorer les promesses faites
pour honorer la commande
pour honorer la mémoire
pour honorer ses promesses
de l'irrésistible ascension
l'irrésistible ascension de
l'irrésistible montée en puissance de
une irrésistible montée en puissance
jouir d'un statut particulier
dans un joyeux désordre
de ces joyeux luron
joyeux Noël
donner la nausée
avec un plaisir évident
avec un plaisir non dissimulé
éprouver du plaisir
faire plaisir à tout le monde
procurer du plaisir
savourer ce plaisir
un malin plaisir à
un plaisir sensuel
varier les plaisirs
jeter la rancune
jeter la rancune à
la rancune à la rivière
sentir rassuré
révolté contre
se révolter contre
cale sèche
des feuilles sèches
du pain sec
du pain sec et
sec comme un coup de trique
sec et humide
sec et nerveux
un bruit sec
un geste sec
un licenciement sec et

Voici maintenant les constructions trouvées uniquement via l'extraction syntaxique :

donner un spectacle affligeant
une affligeante banalité
anxiété et dépression
avoir des vertus apaisantes
un geste apaisant
tant d'ardeur
balancer de droite à gauche
balancer un coup de pied
briser la vitre arrière
briser le carreau d'une fenêtre
briser un mouvement de grève de fonctionnaires nationaux
dont les vitres ont été brisées
l'homme a brisé la vitre du véhicule
la vitrine d'un magasin a été brisée
ne pas hésiter à briser le tabou
une volonté de briser des tabous
apporter la preuve de sa culpabilité
éprouver un sentiment de culpabilité
être convoqué sur reconnaissance préalable de culpabilité
la culpabilité de l'accusé
la culpabilité ou l'innocence
la procédure de reconnaissance de culpabilité
reconnaissance préalable de culpabilité
rongé par la culpabilité
un aveu de culpabilité
un fort sentiment de culpabilité
un sentiment de culpabilité
un verdict de culpabilité
une présomption de culpabilité
accuser de chercher à déstabiliser le pays
avoir participé à une machination visant à déstabiliser Sarkozy
déstabiliser le régime iranien
visant à déstabiliser Nicolas Sarkozy
mais avoir douché des espoirs
la toiture s'est effondrée
le communisme s'est effondré
le marché immobilier s'effondre
le mur de Berlin s'effondrait
les prix de l'immobilier s'effondrent
s'effondrer en pleurs
s'effondrer en sanglots
s'être effondré avec fracas
s'être effondré comme un château de cartes
s'être effondré sous le poids de la neige
une école s'effondre à Haïti
Wall Street s'effondre
avoir provoqué un certain émoi
avoir provoqué un vif émoi
avoir suscité un certain émoi
avoir suscité un vif émoi
l'affaire avait suscité l'émoi
semer l'émoi
susciter beaucoup d'émoi
en garder un souvenir ému
garder un souvenir ému
déception et frustration
engendrer une frustration
un sentiment de frustration
être honorés parmi les morts pour la patrie
honorer les criminels de guerre
Jacques Chirac a honoré
une distinction qui honore
lever les inhibitions
l'irrésistible montée
jouir d'un prestige
jouir d'une bonne image
jouir d'une certaine liberté
jouir d'une cote de popularité
jouir d'une forte popularité
jouir d'une immunité
jouir d'une impunité
jouir d'une popularité
jouir d'une réputation
jouir d'une totale liberté
jouir un privilège
l'ambiance joyeuse

l'humeur joyeuse
ne pas être un joyeux drille
des appels malveillants
des appels téléphoniques malveillants
une intention malveillante
bouder son plaisir hier
ça fait tellement plaisir
le plus grand plaisir des spectateurs
le public a boudé son plaisir
mais ne pas bouder son plaisir
ne pas bouder son plaisir
on ne boude pas son plaisir
prendre énormément de plaisir
prendre toujours autant de plaisir
prendre un malin plaisir
prendre visiblement plaisir
se faire un malin plaisir
sembler prendre un malin plaisir
se sentir rassuré
se révolter contre l'injustice

supprimer [NUM] emplois sans licenciements secs
assurer qu'il y aura des licenciements secs
boire cul sec
en cale sèche
être mis au pain sec
éviter les licenciements secs
l'encre n'est pas sèche
n'avoir plus un poil de sec
n'y avoir aucun licenciement sec
plus chaud et sec
prévoir aucun licenciement sec
procéder à des licenciements secs
tomber en panne sèche
un claquement sec
un mur de pierres sèches
un nettoyage à sec
un temps chaud et sec
un vin blanc sec
des similitudes troublantes
la ressemblance est troublante

Constructions trouvées en commun par les deux systèmes

Voici les 108 constructions qui ont été trouvées en commun via l'extraction syntaxique et l'extraction par n-grammes, et n'ont donc pas été soumises aux évaluateurs.

une affinité particulière
des affinités électives
un spectacle affligeant
les vertus apaisantes
des déclarations apaisantes
des paroles apaisantes
tempérer les ardeurs
refroidir les ardeurs
modérer les ardeurs
l'ardeur réformatrice
freiner les ardeurs
doucher les ardeurs
calmer les ardeurs
moins attirant
se balancer doucement
balancer mes cendres sur Mickey
balancer à bout de bras
Ça balance pas mal
briser une vitre
briser l'hégémonie
briser net
briser les reins
briser le tabou
briser le carcan
briser le blocus
briser la nuque
briser l'omerta
briser l'élan
briser un carreau
briser une vitrine
oser briser le tabou
briser la vitre d'une voiture
briser le plafond de verre
briser le cercle vicieux
briser la loi du silence
briser la colonne vertébrale
briser la vitre du véhicule
la culpabilité d'Yvan Colonna
un peu déconcerté
doucher les espoirs
doucher l'enthousiasme
s'être partiellement effondré
s'effondrer en larmes
s'effondrer brutalement
un émoi vif
susciter l'émoi
provoquer l'émoi
honorer la mémoire
visiblement ému
un souvenir ému
très ému
la frustration sexuelle
la frustration grandissante
colère et frustration
honorer une promesse
honorer ses engagements
honorer ses dettes
honorer les commandes
honorer un rendez- vous
une irrésistible envie
un charme irrésistible
l'irrésistible ascension
poursuivre son irrésistible ascension
jouir sans entraves
jouir d'une notoriété
jouir d'une réelle popularité
jouir d'une grande popularité
jouir d'une excellente réputation
jouir d'une bonne réputation
un joyeux luron
un joyeux drille_N
un joyeux désordre

un joyeux bordel
la Veuve Joyeuse
de ces joyeux drilles
les rumeurs malveillantes
les nausées et les vomissements
un plaisir solitaire
gâcher le plaisir
de menus plaisirs
bouder son plaisir
un plaisir non dissimulé
prendre un plaisir évident
une rancune tenace
garder rancune
avoir la rancune tenace
avoir jeté la rancune à la rivière
un peu rassuré
une mine réjouie
révolté par l'injustice

une toux sèche
une guitare sèche
des raisins secs
une panne sèche
des licenciements secs
des haricots secs
des herbes sèches
des fruits secs
des toilettes sèches
des gâteaux secs
des biscuits secs
cul sec
raisins secs et
chaud et sec
le pain sec et à l'eau
une coïncidence troublante
un détail troublant
des éléments troublants

ANNEXE C. COMBINATOIRE

Voici le mode de calcul des valeurs données en 1.4.1.2.

N-grammes non contigus

Si l'on extrait tous les *n*-grammes éventuellement non contigus dans une fenêtre de *k* tokens, on obtient $\binom{k}{n}$ possibilités. En effet chaque ensemble (non ordonné) de *n* tokens choisis parmi les *k* tokens de la fenêtre correspond à un seul *n*-gramme non contigu. Par exemple, si $n = 3$ et $k = 6$, dans la fenêtre $(t_1, t_2, t_3, t_4, t_5, t_6)$, l'ensemble $\{t_2, t_5, t_6\}$ ne correspond qu'à un trigramme, qui est (t_2, t_5, t_6).

A chaque décalage de cette fenêtre d'un token vers la droite, on ajoute tous les *n*-grammes qui contiennent le dernier token sur la droite, et *n*-1 tokens parmi les *k*-1 autres tokens de la fenêtre, soit $\binom{k-1}{n-1}$ nouveaux *n*-grammes.

Dans une phrase de *t* tokens, on effectuera $(t-k)$ décalages de la fenêtre, ce qui donne en tout $\binom{k}{n} + (t-k)\binom{k-1}{n-1}$ possibilités.

Sous-arbres syntaxiques

La figure 9 représente l'arbre de 20 nœuds, avec au maximum 5 gouvernés par gouverneur, générant le plus de sous-arbres. Il s'agit d'une estimation très pessimiste, un arbre syntaxique réel comptant généralement plus d'étages (et donc moins de sous-arbres).

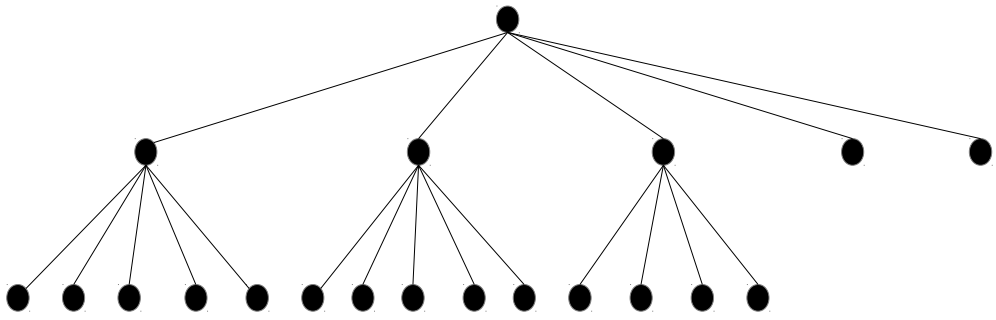


Figure 9 : Arbre de dépendances de 20 nœuds produisant le plus de sous-arbres pour un maximum de 5 gouvernés par nœud

ANNEXE D. ILLUSTRATIONS DES ALGORITHMES

Extraction de tous les sous-arbres d'un arbre enraciné

Est ici illustré l'algorithme donné en 2.2.2.2.

L'arbre de la figure 10 représente la phrase *Elle donne libre cours à ses envies.*

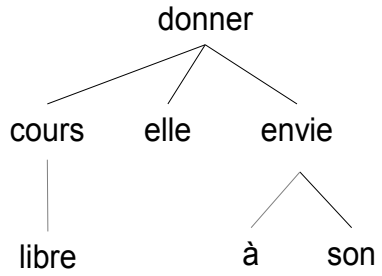


FIGURE 10 : Arbre de dépendances ordonné et lemmatisé
pour la phrase *Elle donne libre cours à ses envies*

En lisant la racine, une liste est créée, qui ne contient que le nœud *donner* :

donner

Puis en arrivant sur *cours* :

- une nouvelle liste est créée, qui contient uniquement *cours*
- la liste initiale est dupliquée (car elle contient le gouverneur de *cours*), et la pile jumelle se voit ajouter *cours*. L'ensemble des listes est alors :

donner	
cours	
donner	cours

Puis, en arrivant sur *libre* :

donner		
cours		
donner	cours	
libre		
cours	libre	
donner	cours	libre

Et en arrivant sur *elle* :

donner					
cours					
donner	cours				
libre					
cours	libre				
donner	cours	libre			
elle					
donner	elle				
donner	cours	^^	donner	elle	
donner	cours	libre	^^	donner	elle

Rapports d'inclusion

Est ici illustré l'algorithme donné en 2.2.2.3. L'arbre de la figure 11 est utilisé comme exemple.

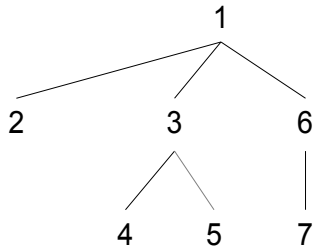


FIGURE 11 : Exemple de parcours d'arbre

Pour plus d'efficacité, les sous-arbres déjà extraits dans la phrase en cours sont regroupés par taille, l'extraction ayant lieu comme décrit précédemment, par parcours en profondeur et duplication des sous-arbres existants. Cependant cette duplication se fait de manière ordonnée : les sous-arbres comptant un seul nœud sont dupliqués (le cas échéant) en priorité, puis ceux de 2 nœuds, etc. Les règles suivantes sont alors appliquées.

- Un sous-arbre de taille 2 a pour SII les deux sous-arbres unaires composés chacun d'un de ses nœuds.
  Par exemple ici le sous-arbre '3→4' a pour SII les sous-arbres unaires '3' et '4'.
- Un sous-arbre de taille > 2 a pour SII le sous-arbre S_i dont il est issu par duplication (qu'on appellera ici son *original*)

- Un sous-arbre de taille > 2 a également pour SII tout sous-arbre dont le dernier token est le nœud courant (c'est-à-dire un sous-arbre créé depuis l'arrivée sur le nœud courant), et qui partage au moins un SII avec son original.

Par exemple, après avoir lu le nœud 3, et créé les sous-arbres correspondants, on doit avoir la configuration représentée par la figure 12, où les traits signalent les rapports d'inclusion immédiate :

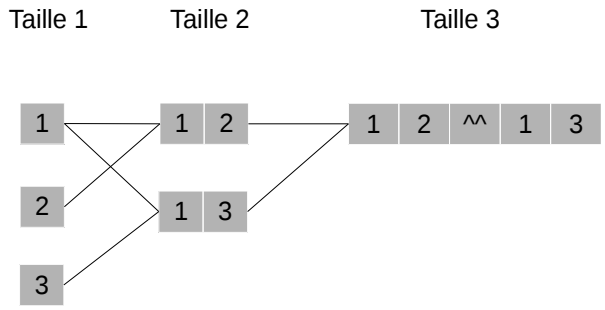


FIGURE 12 : Exemple d'identification des sous-arbres immédiatement inclus : situation initiale

En arrivant sur le nœud 4, on crée :

- le sous-arbre unaire "4"
- puis le sous-arbre de taille 2 "3→4", qui a pour SII les sous-arbres unaires "3" et "4", puis le sous-arbre de taille 3 "1→3→4", qui a pour original, et donc aussi pour SII, le sous-arbre "1→3".

Les sous-arbres et rapports d'inclusion ajoutés sont représentés en rouge sur la figure 13.

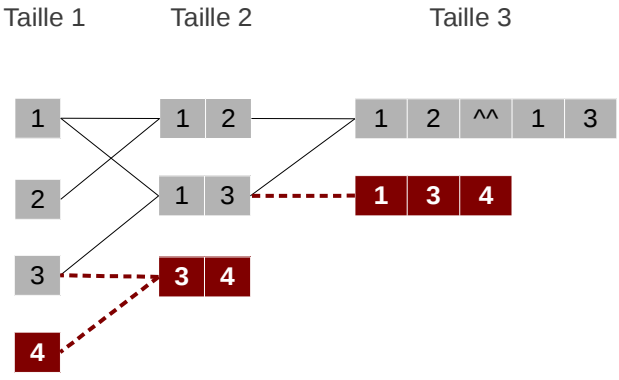


FIGURE 13 : Exemple d'identification des sous-arbres immédiatement inclus : première étape

- On cherche ensuite à savoir si "1→3→4" a d'autres SII. Si c'est le cas, ils doivent faire partie de l'ensemble des sous-arbre de taille 2 créés depuis l'arrivée sur le nœud 4. Il n'y en n'a qu'un ici (c'est "3→4").

- On vérifie donc si "3→4" et l'original de "1→3→4" (c'est-à-dire "1→3") ont un SII commun. C'est effectivement le cas (il s'agit ici de "3"). Par conséquent "1→3→4" a aussi pour SII "3→4".

  Sur la figure 14, les traits bleus pointent vers le SII partagé par "3→4" et l'original de "1→3→4", et le trait rouge indique le rapport d'inclusion immédiate ajouté.

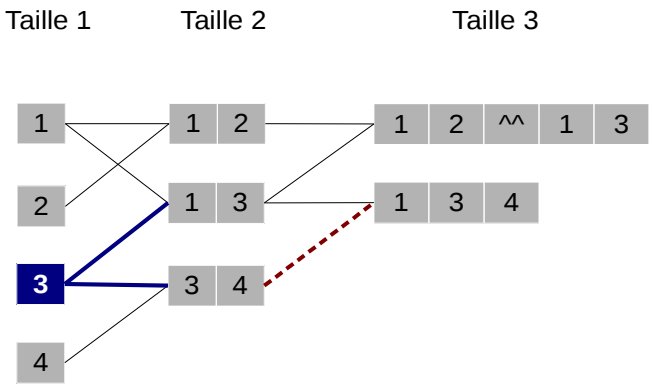


FIGURE 14 : Exemple d'identification des sous-arbres immédiatement inclus : seconde étape

- On ajoute ensuite un sous-arbre de taille 4 ("1→2^^1→3→4"), qui a pour original "1→2^^1→3".

  Et on vérifie si les autres sous-arbres de taille 3 ajoutés depuis l'arrivée sur le nœud 4 (ici il n'y en n'a qu'un, c'est "1→3→4"), partagent un SII avec l'original. C'est effectivement le cas (l'original et "1→3→4" partagent "1→3"). Par conséquent "1→2^^1→3→4" a également pour SII "1→3→4".

  Sur la figure 15, la couleur rouge indique toujours les ajouts, et la couleur bleue permet d'identifier le SII partagé.

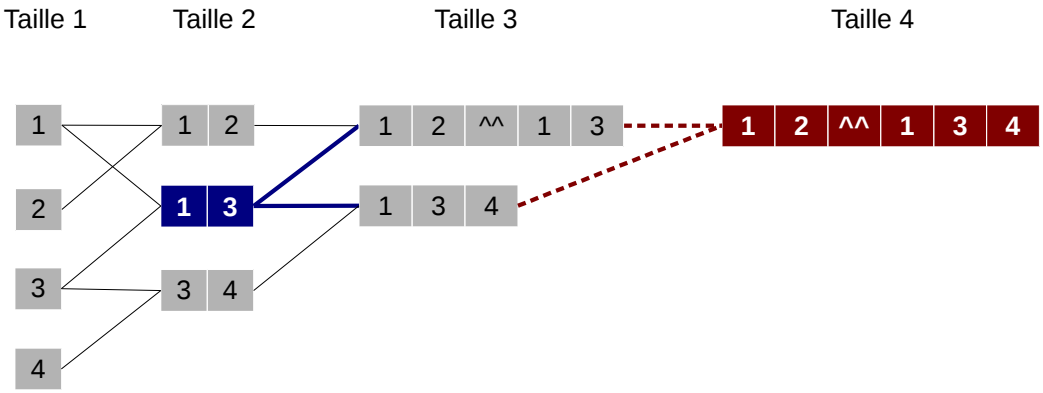


FIGURE 15 : Exemple d'identification des sous-arbres immédiatement inclus : troisième étape

RÉFERÉNCES

Aubin, S., & Hamon, T. (2006). Improving Term Extraction with Terminological Resources. *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL)*.

Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., & Rinck, F. (2008). Lexique des affects: constitution de ressources pédagogiques numériques. *Autour des langues et du langage: perspective pluridisciplinaire* (p. 407-414). Presses Universitaires de Grenoble.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. *4th International Conference on Language Resources and Evaluation (LREC)* (Vol. 4).

Charest, S., Brunelle, E., Fontaine, J., & Pelletier, B. (2007). Élaboration automatique d'un dictionnaire de cooccurrences grand public. *14e conférence TALN*.

Charest, S., Brunelle, E., & Fontaine, J. (2010). Au-delà de la paire de mots: extraction de cooccurrences syntaxiques multilexémiques. *17e conférence TALN*.

Church, K. W., & Hanks, P. (1991). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.

Cowie, A. P. (1998). Introduction. *Phraseology: Theory, Analysis, and Applications*. Oxford University Press.

da Silva, J. F., & Lopes, G. P. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. *Sixth Meeting on Mathematics of Language* (p. 369–381).

Dias, G., Guilloré, S., & Lopes, J. G. P. (2000a). Normalization of Association Measures for Multiword Lexical Unit Extraction. *International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Engineering and Industrial Applications (ACIDCA)* (p. 207–216).

Dias, G., Guilloré, S., & Lopes, J. G. P. (2000b). Extraction automatique d'associations textuelles à partir de corpora non traités. *5th International Conference on the Statistical Analysis of Textual Data* (p. 213–221).

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, *19*(1), 61–74.

Evert, S. (2005). *The statistics of word cooccurrences* (PhD Thesis). University of Stuttgart.

Evert, S., Heid, U., & Spranger, K. (2004). Identifying morphosyntactic preferences in collocations. *4th International Conference on Language Resources and Evaluation (LREC)* (Vol. 4).

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. *39th Annual Meeting on Association for Computational Linguistics* (p. 188–195).

Fazly, A. (2007). *Automatic acquisition of lexical knowledge about multiword predicates* (PhD Thesis). University of Toronto.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. *Phraseology : An interdisciplinary perspective* (p. 27–50). Amsterdam: John Benjamins Publishing Co.

Grossmann, F., & Tutin, A. (Éd.). (2003). *Les collocations : analyse et traitement*. Travaux et recherches en linguistique appliquée. Amsterdam: de Werelt.

Green, S., de Marneffe, M. C., Bauer, J., & Manning, C. D. (2011). Multiword expression identification with tree substitution grammars: a parsing tour de force with French. *Conference on Empirical Methods in Natural Language Processing* (p. 725–735).

Hagège, C., & Roux, C. (2003). Entre syntaxe et sémantique: Normalisation de la sortie de l'analyse syntaxique en vue de l'amélioration de l'extraction d'information à partir de textes. *10e conférence TALN*.

Hudson, R. (2000). Discontinuity. *TAL. Traitement automatique des langues*, *41*(1), 15–56.

Hunston, S., & Francis, G. (2000). *Pattern grammar : a corpus-driven approach to the lexical grammar of English*. Studies in corpus linguistics ; 4. Amsterdam ; Philadelphia: J. Benjamins.

Kahane, S. (2001). Grammaires de dépendance formelles et théorie Sens-Texte. *8e conférence TALN*.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. *Information Technology*, *105*, p. 116.

Kraif, O., & Tutin, A. (2009). Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. *7th Conference of Teaching and Language Corpora (TaLC)*.

Kraif, O., Diwersy, S. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. *19e conférence TALN*.

Krenn, B., & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. *ACL Workshop on Collocations* (p. 39–46).

Laurent, D., Nègre, S., & Séguéla, P. (2009). Apport des cooccurrences à la correction et à l'analyse syntaxique. *16e conférence TALN*.

Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Lin, D. (1999). Automatic identification of non-compositional phrases. *37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (p. 317–324).

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.

Martens, S. (2010). Varro: an algorithm and toolkit for regular structure discovery in treebanks. *23rd International Conference on Computational Linguistics: Posters* (p. 810–818).

Martens, S., & Vandeghinste, V. (2010). An efficient, generic approach to extracting multi-word expressions from dependency trees. *CoLing Workshop: Multiword Expressions: From Theory to Applications*.

Mel'čuk, I. (2011). Tout ce que nous voulions savoir sur les phrasèmes, mais…. consulté le 12/05/2012, de `http://olst.ling.umontreal.ca/pdf/MelcukPhrasemes2011.pdf`

Pearce, D. (2001). Synonymy in collocation extraction. *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics* (p. 41–46).

Pecina, P., & Schlesinger, P. (2006). Combining association measures for collocation extraction. *COLING/ACL conference poster sessions* (p. 651-658).

Ramisch, C., Villavicencio, A., & Boitet, C. (2010). Mwetoolkit: a Framework for Multiword Expression Identification. *7th International Conference on Language Resources and Evaluation (LREC)*.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. *CICLing* (p. 189–206).

Seretan, V. (2008). *Collocation extraction based on syntactic parsing* (PhD Thesis). University of Geneva.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Dordrecht: Springer.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, *19*(1), 143–177.

Tapanainen, P., & Järvinen, T. (1997). A non-projective dependency parser. *5th conference on Applied natural language processing* (p. 64–71).

Tutin, A., & Grossmann, F. (2002). Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, *7*(1), 7–25.

Tutin, A. (2008). For an extended definition of lexical collocations. *13th EURALEX conference*.

Tutin, A. (2010). Les collocations dans les dictionnaires monolingues spécialisés de collocations. *2e Congrès Mondial de Linguistique Française (CMLF)*.

Van de Cruys, T., & Villada Moirón, B. (2007). Semantics-based multiword expression extraction. *ACL : Workshop on a Broader Perspective on Multiword Expressions* (p. 25–32).

Villada Moirón, B. (2005). *Data-driven identification of fixed expressions and their modifiability* (PhD Thesis). University of Groningen.

Résumé

Ce document présente une méthode d'extraction d'expressions polylexicales à partir de corpus analysés syntaxiquement. Ces expressions restent un problème central pour le traitement automatique des langues naturelles, et leur extraction et encodage automatiques sont des tâches encore non résolues.

L'approche implémentée permet en particulier d'extraire des expressions de plus de deux mots, et une attention particulière a été portée aux constructions récurrentes imbriquées. Une description morphosyntaxique fine des unités extraites est également générée, en termes de relations syntaxiques, ordres des mots possibles, distance entre mots, flexion ou détermination, informations qui nous semblent nécessaires à leur bonne intégration aux lexiques, pour des applications comme l'extraction d'information ou la traduction automatique.

Mots-clés : expression polylexicale, corpus arboré, grammaire de dépendances

Abstract

This document describes a method for extracting multiword expressions from syntactically analyzed corpora. These expressions are still a main issue for NLP applications. Their automatic extraction, and appropriate description, remain largely unsolved problems.

This approach most notably allows the extraction of expressions composed of more than two words, and focuses on the issue of nested recurrent structures. It also yields a fine-grained morphosyntactic description of the extracted units, including syntactic relations, possible word orders, contiguity, inflection or determination, which we think is necessary to a proper encoding of these expressions in lexicons, for applications such as information extraction or machine translation.

Keywords : multiword expression, MWE, treebank, dependency parsing