

1 RECHERCHE D'EXPRESSIONS FIGÉES DU FRANÇAIS EN BELGIQUE. MÉTHODE ET OBSERVATIONS.

Anne DISTER, Cédrick FAIRON, Patrick WATRIN
Université catholique de Louvain

1. Introduction

Malgré ce que suggère leur nom, nombre d'expressions « figées » peuvent être soumises à des variations de formes et de structure (dans ce cas, on les qualifie parfois d'expressions semi-figées). Il s'agit de phénomènes bien connus tels que l'insertion d'adverbes ou d'adjectifs, le remplacement ou l'omission d'un membre de l'expression, etc. (cf., entre autres, Moon 1998 ; Mejri 2002, 2003) Il n'en reste pas moins que ces expressions polylexicales se caractérisent par une liberté syntaxique restreinte et une certaine opacité sémantique (Gross 1996). Les quelques exemples suivant en sont l'illustration¹ :

<attraper> (<quelqu'un>+<E>) par le collet / se <faire> attraper par le collet
<attraper> un (<A>+<E>) coup de froid
(<attendre>+<chercher>+<trouver>) le prince charmant
<aller> chercher (<ADV>+<E>) loin

Depuis longtemps, on insiste sur la nécessité d'identifier ces structures figées dans les tâches de traitement automatique des langues (par exemple, Laporte 1988), mais leur nombre et leur caractère mouvant rendent relativement complexe une identification automatisée. Des méthodes statistiques ont été proposées pour identifier ces groupes dont la structure syntaxique peut être atypique et dont les formes cooccurrent parfois de manière remarquable par rapport à leur usage moyen en corpus. Les méthodes « linguistiques » de repérage sont souvent écartées, car elles nécessitent des ressources descriptives importantes qui ne sont pas toujours disponibles pour toutes les langues. Pour le français, les ressources du lexique-grammaire composées par le LADL, sous la direction de Maurice Gross, sont très détaillées

¹ Dans la notation que nous utilisons, <attraper> représente le lemme du verbe *attraper* et le symbole <E> marque l'absence d'un élément. Par exemple, (<quelqu'un>+<E>) signifie que le lemme <quelqu'un> est optionnel. <A> et <ADV> signalent la présence d'un mot représenté par sa catégorie grammaticale (respectivement *adjectif* et *adverbe*).

(plus de 30 000 expressions figées cataloguées par type de structure syntaxique). Des expériences pratiques ont déjà été tentées pour utiliser ces ressources à des fins de repérage automatique en corpus. La méthode la plus aboutie est celle de Senellart (1998) basée sur l'exploitation des tables d'expressions figées à l'aide de « graphes paramétrés ».

L'objectif de notre recherche est d'appliquer cette méthodologie sur un vaste corpus de presse écrite afin d'apprécier les variations que l'usage impose au figement.

Nous avons choisi de travailler sur les données du projet BFQS qui envisage l'étude comparée des expressions figées dans quatre variantes du français : le français en Belgique (B), en France (F), au Québec (Q) et en Suisse (S) (Leclère 2000, Lamiroy *et al.* 2003)². Les linguistes impliqués dans cette étude, qui se fait dans le cadre théorique du lexique-grammaire (Gross 1975 ; Gross 1982), proposent une classification des expressions figées non seulement en fonction de leur structure syntaxique mais également selon leur utilisation, leur vitalité dans ces quatre régions de la francophonie.

2 Présentation des données

2.1 Les expressions figées BFQS

Les données du projet BFQS sont recensées dans des tables qui se présentent de la manière suivante : la 1^{re} colonne note la structure de l'expression qui figure dans la 2^e colonne, les 4 colonnes suivantes représentent les 4 espaces géographiques envisagés (le signe “+” indiquant que l'expression est utilisée, le signe “-” qu'elle ne l'est pas), la dernière colonne est une traduction « passe-partout » (dans le sens de non ambiguë pour les 4 variétés envisagées) de l'expression figée.

²Nous remercions les auteurs qui ont mis leurs bases de données à notre disposition pour cette étude.

Struc.		B	F	Q	S	TRADUCTION
N <u>V Adv</u>	Aller <tout> schuss	-	+	-	+	Foncer sans précaution
N <u>Nég V Adv</u>	Aller <très> fort (ne pas)	+	+	+	+	Être dans une phase difficile
N <u>V Prép N</u>	Aller à (ça)	+	+	+	+	Ça convient à
N <u>V Prép N</u>	Aller à Canossa	+	+	-	+	Reconnaître humblement son erreur
N <u>V Prép N</u>	Aller à l'abattoir / au carnage	+	+	+	+	Être engagé dans une action vouée à un échec grave
N <u>V Prép N</u>	Aller à la / aux bécosse<s>	-	-	+	-	Aller aux toilettes
N <u>V Prép N</u>	Aller à la boucherie	+	+	-	+	Aller se faire massacrer
N <u>V Prép N</u>	Aller à la broue	-	-	+	-	Aller boire un coup une bière
N <u>V Prép [N Prép N]</u>	Aller à la chasse aux lapins	+	-	-	-	Se fourrer les doigts dans le nez
N <u>V Prép N Nég N</u>	Aller à la chasse pas de fusil	-	-	+	-	Ne pas être équipé adéquatement
N <u>V à N</u>	Aller à la cour / à la culotte	+	-	-	-	Aller aux toilettes

Illustration 1. Extrait de la table BFQS

2.2 Le corpus

Notre corpus compte huit années du quotidien national *Le Soir* (1994-2001)³, ce qui représente près de 194 millions de mots. Il s'agit d'un corpus de presse écrite, par conséquent peu varié stylistiquement (pour le dire vite, de l'écrit très standardisé). Nos résultats sont donc inévitablement partiels, beaucoup d'expressions BFQS étant utilisées, notamment, dans un contexte plus familier. On sera donc attentif au fait que nos remarques et observations doivent être remises dans la perspective de cette même presse écrite et ne peuvent servir de base à des généralisations dépassant ce cadre particulier.

3 Méthodologie

Nous suivons la méthodologie proposée par Senellart (1998). Elle repose sur une description linguistique exacte des phénomènes traités et n'utilise aucune information statistique. La partie principale de cette description consiste en un ensemble de tables qui associent aux prédicats étudiés les structures et propriétés syntaxiques qu'ils vérifient. Sur cette base, nous générons automatiquement, pour chaque entrée de la table, une grammaire que nous appliquons sur nos corpus au moyen du logiciel de traitement de corpus Unitex⁴ (Paumier 2003).

³ Il s'agit de l'un des principaux journaux de la presse belge francophone.

⁴<http://www-igm.univ-mlv.fr/~unitex/>

Les tables du lexique-grammaire (cf. illustration 2) présentent, sous la forme de matrices, une description formelle des prédicats d’une langue. Chaque ligne correspond à un prédicat et chaque colonne représente une propriété syntaxique ou distributionnelle. Aux intersections ligne-colonne, un signe ‘+’ ou ‘-’ signifie, respectivement, que l’entrée actualise ou n’actualise pas la structure correspondante. Par exemple, la propriété *Prep N2* := *Ppv* (pronominalisation du complément prépositionnel N2) est vérifiée à la fois pour *assurer une descendance* à *qqun* et pour *arracher le coeur de qqun*. Dès lors, nous pouvons recontrer des phrases du type :

- [illegible]

Par contre, seul *assurer une descendance à qqun* vérifie la transformation de pronominalisation. Dès lors, si, nous pouvons observer des structures telle que :

- Notons que les tables BFQS ne possèdent pas à l'origine (cf. illustration 1) toutes les caractéristiques indispensables à la méthodologie que nous nous proposons d'utiliser. Nous avons donc dû les adapter de la manière suivante :

1. Répartition des entrées lexicales selon la structure élémentaire qu'elles actualisent. Le lexique-grammaire repose son étude des langues sur la notion de phrase élémentaire, unité sémantique minimale. Une même forme verbale peut, en effet, avoir plusieurs significations. Dès lors, afin d'évacuer l'ambiguïté, il est nécessaire d'appréhender leur contexte constitué des compléments essentiels et formant une structure élémentaire.
2. Identification, pour chaque expression, des zones libres et fixes et des variations possibles.
3. Recensement des transformations. Nous prenons en compte les transformations unaires agissant au niveau de la phrase élémentaire sans considérer les transformations complexes (c'est-à-dire celles combinant plusieurs phrases élémentaires).

3.2 Graphes paramétrés

Généralement, le contenu des tables est exploité au moyen de « graphes paramétrés » (Roche 1993, Senellart 1999, Paumier 2003). Ces graphes formalisent les constructions lexico-syntaxiques pertinentes pour une table donnée. Chaque construction, c'est-à-dire la phrase élémentaire définitoire de la table ainsi que l'ensemble de ses propriétés syntaxiques et distributionnelles, correspond à un parcours du graphe (cf. illustration 3).

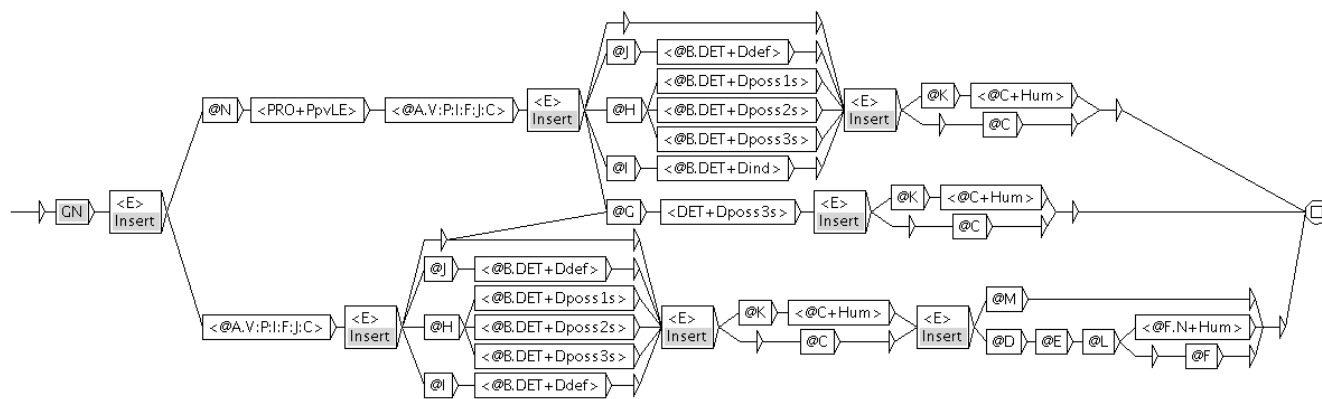


Illustration 3. Graphe patron de la table N0VN1PrepN2

Le graphe contient un ensemble de paramètres (ou variables). Le nom des paramètres renvoie aux différentes colonnes de la table. Ainsi, la première colonne est notée @A, la seconde, @B, etc. La substitution des variables s'exécute de trois manières différentes selon que la colonne contient une entrée lexicale ou une des deux marques de propriétés ('+' ou '-')

1. si la variable @X désigne un signe '-', le chemin est supprimé ;
2. si elle désigne un signe '+', le chemin est validé ;
3. si elle désigne une entrée lexicale, elle est remplacée par cette entrée.

Un processus automatique génère, pour chaque entrée de la table, un graphe qui contient uniquement les constructions grammaticales et acceptables de cette entrée (cf. illustration 3).

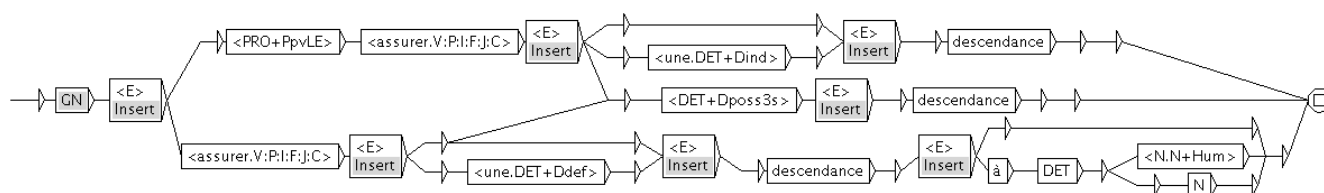


Illustration 4. Graphe généré pour "assurer une descendance à qqun"

Parce qu'elle repose sur une description exacte du fonctionnement des prédicats de la langue, cette méthode est particulièrement adaptée à l'analyse syntaxique. Toutefois, dans le cadre qui est le nôtre, nous ne pouvons nous limiter aux seules propriétés décrites au sein des tables. Nous visons, en effet, l'étude des variations possibles au sein des figements. Les variations n'étant pas prises en compte par les tables BFQS, nous devons nécessairement « relâcher » les structures décrites dans les graphes paramétrés. Ce « relâchement » opère à trois niveaux :

1. suppression des déterminants (<DET>), afin de permettre de retrouver des expressions avec d'autres déterminants que ceux indiqués dans l'expression telle qu'elle est répertoriées dans BFQS ;
2. ajout de la possibilité d'insertions entre les différents composants de l'expression, qui peut alors être retrouvée même lorsque sa linéarité est brisée (typiquement, on peut avoir un adverbe entre le verbe et le complément de l'expression) ;
3. modification du complément figé : a) lemmatisation des noms et adjectifs, afin de pouvoir retrouver des occurrences qui seraient d'un genre et / ou d'un nombre différents de ceux donnés dans l'expression, b) simplification (suppression de certains éléments) et c) possibilité de variation graphique (*clé* et *clef* ; *renom* vs *renon* par exemple).

Suite à nos modifications, l’expression de la table BFQS *jouer un pied de cochon à quelqu’un* “jouer un tour de cochon à quelqu’un” devient chez nous : *<jouer> <E> <pied> de <cochon>*.

Les angles signifient que l’on recherche toutes les formes fléchies du lemme (donc non seulement *jouer* mais aussi *jouera, jouaient, joues*, etc., que l’on trouvera non seulement *pied* et *cochon* mais aussi *pieds* et *cochons*) ; on supprime le déterminant (*<E>* signifie ensemble vide) afin de permettre de retrouver n’importe quelle forme à cet endroit grâce à l’insertion dans le graphe ; on simplifie également l’expression en supprimant le /N2/ à *quelqu’un*, qui a peu de chances d’apparaître tel quel en corpus. On trouvera ainsi, grâce à cette modification de l’expression et à la possibilité d’insérer des incises, l’expression sous la forme suivante : *Ils ont toujours joué des pieds de cochon à leur institutrice*, ce qui aurait été impossible en se basant strictement sur la structure de l’expression telle qu’elle est décrite dans la table.

En « relâchant » les contraintes, nous pouvons donc prétendre à une moisson plus large mais, dans le même temps, nous obtenons un bruit inévitablement plus important (cf. 4.3). Par conséquent, il est nécessaire de garder un ensemble suffisant de contraintes structurelles afin de ne pas obtenir en sortie un ensemble trop important de solutions concurrentes (cf. illustration 5).

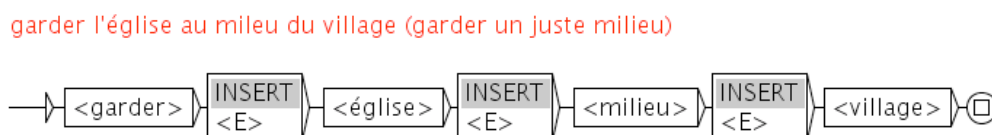


Illustration 5. Graphe relâché

4 Observations linguistiques

4.1 Expressions non attestées

Dans la liste des expressions répertoriées par le projet BFQS, nous nous sommes penchés sur les 557 expressions notées B, c’est-à-dire attestées en Belgique francophone. Parmi celles-ci, 163 (soit 29,26 %) n’ont pas été trouvées : *avoir la tête dans le cul* “avoir mal à la tête, suite à un excès d’alcool”, *dépenser un argent bête* “dépenser beaucoup d’argent”, *faire bouf* “sortir d’une affaire sans perte ni bénéfice”, *pousser bobonne dans les bégonias* “exagérer”, etc. De manière générale, on ne s’étonnera pas devant la liste des expressions non récupérées : elles appartiennent la plupart du temps à un registre de langue peu soutenu, familier voire parfois vulgaire (cf. la 1^{re} expression citée), que l’on a peu de chances de voir figurer dans un corpus de textes journalistiques.

4.2 Limites de la méthode

4.2.1 Le silence

4.2.2

Outre les expressions que l'on a peu de chances de trouver dans le type de textes sur lequel nous travaillons⁵, nous avons un certain nombre d'occurrences d'expressions qui ne sont pas récupérées, parce qu'elles échappent au repérage tel que nous l'avons formalisé. En effet, notre méthode d'extraction ne nous permet pas de repérer des tournures, pourtant susceptibles d'apparaître en corpus, mais dont la structure syntaxique n'est pas décrite dans notre graphe. Ainsi, notre méthode produit du silence notamment dans les cas suivants :

- l'extraction : l'expression *dépenser un argent bête* ne sera pas reconnue dans une phrase où le /N1/ est mis en évidence par l'extraction en *c'est... que*, comme dans : *c'est un argent bête que j'ai dépensé* ;
- lorsque le complément figé /N1/ ou /N2/ est antéposé : *brosser un cours* "sécher un cours" n'est pas reconnu dans *Il ne réussira jamais les cours qu'il a brossés durant toute l'année* ;
- lorsque l'expression est enchâssée dans une autre structure syntaxique, ou dans une autre expression figée. Dans l'exemple suivant, on a les 2 expressions *faire le jardin* "nettoyer le jardin" et *faire les poussières* "prendre les poussières", sans reprise du verbe support *faire* dans le 2^e cas :

Mon dernier espoir, ce sont les « petits boulots » : faire le jardin, les poussières chez les gens.

Selon notre méthode d'extraction, seule cette dernière expression est reconnue par notre graphe, la première étant considérée comme une insertion entre le verbe support *faire* et le complément *les poussières* ;

- lorsque l'insertion permise entre le verbe et le complément figé dépasse 10 mots graphiques, l'expression n'est pas reconnue. Étant donné nos conclusions en ce qui concerne l'insertion (cf. ci-dessous), on peut supposer que les insertions de plus de 10 mots sont rarissimes.

4.2.3 Le bruit

⁵ Notons que nous avons certaines expressions uniquement attestées dans des exemples métalinguistiques. C'est le cas de *faire son trottoir* "nettoyer son trottoir", pour laquelle les 4 attestations correctes relèvent d'un usage métalinguistique. *Faire son trottoir/son seuil/son, jardin* sont des expressions fréquentes dans la (...)

394 expressions sont attestées dans le corpus, qui correspondent à 100 224 occurrences. Parmi celles-ci, nous pouvons estimer à plus ou moins 40 % le nombre d'occurrences erronément reconnues et ne correspondant pas à l'expression figée recherchée. C'est dire que le bruit engendré par une telle méthode de recherche, volontairement approximative, est énorme. En effet, la possibilité d'insertion augmente considérablement le bruit, comme dans l'exemple suivant où nous cherchions *remettre le couvert* "recommencer une activité"⁶ :

Il remet à neuf une ménagère complète de couverts

Ici, le bruit est non seulement dû à la possibilité de l'insertion entre le verbe et son complément, mais également au fait que l'on autorise dans le graphe de recherche la variation en nombre (*couverts* vs *couvert* dans l'expression de départ).

On a la même chose dans la phrase suivante, où l'expression *faire son trottoir* a été erronément reconnue :

Enfin, les maîtres qui laissent leur chien faire des besoins sur les trottoirs sont obligés de les ramasser.

Ici aussi, les licences permises par les graphes (l'insertion – *des besoins sur* ; la variation du nombre – *trottoirs* vs *trottoir* et le changement de déterminant – *les* vs *son*) ont entraîné à tort la reconnaissance de l'expression.

Dans les cas qui génèrent du bruit, on est évidemment aussi confronté au problème classique de l'ambiguïté entre une suite libre et une suite figée. C'est le cas de l'exemple suivant où l'on n'est pas en présence d'une séquence figée *acheter un enfant* "être enceinte" mais bien d'une suite libre⁷ :

(...) enfant à la disposition de leurs clients. On peut même acheter des enfants pour quelques centaines de dollars.

Le phénomène des faux-amis, entre les variantes B et F, entre également en jeu parmi les facteurs de bruit. Ainsi, par exemple, *faire le poirier* est répertorié parmi les expressions B dans le sens particulier de "se tenir en équilibre sur les mains sans que la tête touche au sol", alors que dans le sens F de l'expression, la tête touche le sol (ce qui correspond en fait à l'expression B *faire le trépied*). Dans les occurrences que nous recueillons, il est la plupart du temps impossible de faire le départ entre les sens B et F. De la même manière, *défendre une thèse* est répertorié parmi les expressions B pour "soutenir une thèse de doctorat". La grande

⁶ Il s'agit toutefois de graphes fortement relâchés. Il est évident que la description systématique et précise des variations observables pour chaque expression devrait réduire considérablement le phénomène du bruit.

⁷ Nous ne recensons aucune attestation dans le corpus de l'expression figée mais uniquement des suites libres.

majorité des occurrences récupérées (147 exemples sur 160, soit 92 %) concernent le sens classique de “défendre une opinion”, sens connu en Belgique francophone et en France.

4.3 La variation au sein des parties figées

Nous avons observé en corpus de nombreuses variations par rapport à la forme de l’expression telle que décrite dans les tables. Celles-ci concernent le nom, le déterminant ou le compliment figé. La variation paradigmatique du verbe n’a pas été permise dans les graphes et ne sera donc pas abordée ici⁸.

4.3.1 La variation du nom

Cette variation concerne le nombre du nom, comme dans les exemples suivants, où le singulier alterne avec le pluriel, alors que seule l’un des deux est décrit dans la table :

*La crème ne doit pas **prendre couleur** vs l’arc-en-ciel **prend couleurs** (“se colorer”)*

*Les membres de nANcy ne s’arrêtent en effet pas une seconde de **faire les gugusses** et de tout tourner en dérision vs J’ai toujours adoré **faire le gugusse**, faire rire les copains (“faire le clown”)*

Nous n’avons pas d’exemple de modification du genre, les expressions concernées ne se prêtant pas à ce type de variation, sauf dans *faire le bob* (ou *la bobette*) “ne pas boire pour conduire”, qui n’est pas attestée au féminin dans le corpus.

4.3.2 La variation du déterminant

Il arrive fréquemment que l’on rencontre une variation sur le déterminant : cette variation concerne le nombre (cf. dessus), mais également la sous-catégorisation ou encore simplement la présence de celui-ci. Ainsi, les expressions (*additionner + mélanger*) *des pommes et des poires* “mélanger des torchons et des serviettes” se rencontrent une fois sur deux avec un déterminant défini au lieu de l’indéfini présenté dans la table.

L’expression *remettre le bonjour* “faire le bonjour” est attestée à 4 reprises, toujours avec un déterminant possessif (*mon* et *son*), et non avec le déterminant défini comme l’indique l’entrée de la table.

⁸ Remarquons que dans certains cas cette variation a été explicitement prévue dans les tables, comme pour les entrées de (*additionner+mélanger+comparer*) *des pommes et des poires* “mélanger des torchons et des serviettes”. Dans ce cas, les différents verbes sont recherchés dans le corpus.

À l'inverse, *donner son C4* (à quelqu'un) "renvoyer", listée avec le possessif de la 3^e personne du singulier, se rencontre aussi en corpus avec le possessif de la première personne (*mon employeur ne m'a pas donné mon C4*) ou avec un déterminant autre que possessif (*l'employeur donne le C4 à l'employé désireux de quitter sa fonction*).

Dans notre corpus, les exemples de variation de ce type se rencontrent par centaines.

4.3.3 La modification de la structure du nom figé

De nombreuses expressions permettent une variation de la structure du complément figé. C'est par exemple le cas de *faire les pauses* ("faire les trois huit") qu'on rencontre sous la forme *faire la première pause* et *faire les trois pauses* (ajout d'un déterminant), *faire la pause de nuit* (ajout d'un complément du nom), etc. De même, l'expression *couper au court* ("prendre au plus court") apparaît sous la forme *couper au plus court*, avec l'ajout d'un adverbe, ou *couper court*, avec suppression des morphèmes contractés à *les* (*au*).

En général, on constate que les modifications de la structure du nom figé sont assez minimales, et qu'elles consistent principalement en l'ajout d'un élément de type déterminatif.

4.4 Les insertions

Contrairement à ce que nous attendions, nous trouvons peu d'insertions entre les différentes composantes de l'expression figée. Alors que ce problème d'insertion est typiquement mis en avant dans la reconnaissance automatique (et l'analyse) des expressions figées, constituant d'une certaine manière la pierre d'achoppement du traitement, ces cas sont relativement peu fréquents dans notre corpus. Ainsi, si l'on prend l'exemple de l'expression *remettre le couvert* ("recommencer une activité quelconque"), les 959 occurrences repérées se répartissent comme suit :

- 5 erreurs (segments reconnus à tort comme étant l'expression figée, cf. 4.2.2.) ;
- 9 variations sur le genre : *remettre les couverts* ;
- 3 variations de la structure : *remettre le même couvert* ;
- 60 insertions, principalement des adverbes (*sans doute, annuellement, souvent, donc, en tout cas, une dernière fois, aussitôt, vraisemblablement, mardi, une troisième fois, en quelque sorte, chaque semaine, pour sa part, seul et définitivement, eux aussi, à présent, toutefois, généreusement, magistralement, d'ailleurs, ensuite, rapidement, etc.*).

On a donc, pour cette expression, 92,5 % des occurrences qui correspondent en fait à la structure de base telle que décrite dans la table, sans modification d’aucune sorte ni insertion.

De la même manière, un certain nombre d’expressions, très fréquentes, semblent très figées, au point de se présenter de manière très « stable » en corpus. On peut citer *poser un acte* “accomplir un acte” (425 occurrences), *manger un bout* “manger un morceau” (47 bonnes occurrences), *manger sa parole* “manquer à sa parole” (55 exemples corrects). De plus, on constate, de manière générale, que les insertions observables sont de taille très réduite, dépassant rarement 3 ou 4 mots graphiques (cf. les exemples ci-dessus). Notons au passage qu’en permettant des insertions plus longues on augmente considérablement le bruit lors de l’opération de recherche.

5 Vitalité des expressions

Le sondage que nous avons réalisé dans le journal *Le Soir* nous apporte également une information sur la vitalité des expressions figées B dans la presse. Comme on le voit dans la partie gauche du tableau ci-dessous, les expressions « belges » sont loin d’être des phénomènes linguistiques excentriques et occasionnels⁹. Si la fréquence absolue¹⁰ de la plupart de ces expressions reste modeste au vu de la taille importante de notre corpus, elles n’en sont pas moins comparables aux fréquences des expressions non marquées régionalement (colonne de droite). Dans plusieurs des exemples de ce tableau, la variante B apparaît avec une fréquence pouvant même être supérieure à celle de l’expression équivalente F/B. Dans tous ces cas, les deux variantes (B et B/F) ont le même niveau de langue. Dans le cas de l’expression *arranger les bidons*, on pourrait faire l’hypothèse que c’est la différence de niveau de langue qui donne l’avantage à son équivalent *arranger les affaires*. Mais pour être validée, celle-ci devrait bien sûr être vérifiée sur un plus large échantillon.

Expression B	Occ.	Expression F/B	Occ.
arranger les bidons	12	arranger les affaires	128
battre le beurre	46	pédaler dans la choucroute	49
faire les poussières	8	enlever les poussières	9

⁹ La collecte serait sans doute plus variée en travaillant sur d’autres types de textes (cf. 4.1), notamment sur des corpus oraux.

¹⁰ Le nombre d’occurrences affiché dans ce tableau ne peut être qu’indicatif et ne représente pas précisément le nombre d’occurrences dans le corpus. En effet (cf. 4.2.1.), notre méthode ne permet pas de retrouver des expressions ayant subi des transformations telle que l’extraction, etc.

faire la file	410	faire la queue	248
manger un bout	40	manger un morceau	12
pendre le linge	16	étendre le linge	10
manger sa parole	55	manquer à sa parole	17
faire/tourner les pauses	9	faire les trois huit	1

6 Conclusions

La méthode de recherche des expressions figées que nous avons utilisée dans ce travail a pour but la reconnaissance des expressions et non leur analyse. Elle se veut donc volontairement « approximative » et produit naturellement un bruit important nécessitant une relecture attentive des phrases extraites. Mais par la même occasion, elle permet de retrouver à partir d'une liste d'expressions présentées sous une forme canonique, des actualisations en corpus, qui peuvent apparaître sous des formes très variées, difficilement prédictibles. Grâce à cette souplesse, on obtient une meilleure vision de la variation que connaissent dans les expressions figées (puisque l'on ne présuppose pas de variation paradigmatique ou syntagmatique particulière). Bien entendu, il y a également un aspect contraignant dans la mesure où le bruit généré implique une relecture des candidats extraits. Ce travail nous a permis en particulier de remarquer que :

- les insertions dans les expressions figées sont généralement très courtes. En effet, même si l'on peut aisément forger des exemples comprenant de longues séquences insérées dans la structure d'une l'expression figée, force est de constater que dans notre corpus, les insertions se limitent principalement à un ou deux mots ;
- la structure de la partie nominale figée de l'expression subit peu de variation (et principalement de type déterminatif) ;
- le nom peut parfois varier en nombre ;
- le déterminant peut également être l'objet de variations.

Si l'on se conforme au principe descriptif des tables du lexique-grammaire, la variation observée devrait être codée dans les tables. On se rend naturellement compte de la difficulté d'une telle entreprise, les jugements d'acceptabilité étant sans doute encore plus difficiles à poser qu'ils ne le sont dans le cas des énoncés libres. Pour cette raison, il ne serait pas aisé de choisir de manière très systématique les variantes qui devraient faire ou non l'objet d'une description dans les tables.

Notre méthode nous permet également d'obtenir une information plus précise de la vitalité des expressions figées dans les corpus. Les observations réalisées en ce sens montrent que la présence d'expressions figées B est importante et doit être prise en compte dans les études.

6.1 BIBLIOGRAPHIE

- BOURIGAULT, Didier. 1996. « Lexter, a Natural Language Processing Tool for Terminology Extraction », *Proceedings of the 7th EURALEX International Congress*.
- DANLOS, Laurence (éd.). 1990, *Les expressions figées, Langages* 90.
- DEKANG, Lin. 1998. « Extracting collocations from text corpora », *First Workshop on Computational Terminology*, Montréal.
- DAILLE, Béatrice. (1995) « Study and Implementation of Combined Techniques for Automatic Extraction of Terminology », dans P. Resnik et J. Klavans (éds), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, 49-66.
- DIAS, Gaël. GUILLORÉ, Sylvie. LOPES, Gabriel. 1999. « Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora », *Actes de TALN 1999*, Cargèse, Institut d'Études Scientifiques.
- GOLDMAN, Jean-Philippe. NERIMA, Luka, WEHRLI, Éric. 2001. « Collocation extraction using a syntactic parser », *Proceedings of the ACL 01 Workshop on Collocation*, 61-66.
- GROSS, Gaston. 1996. *Les expressions figées en français : noms composés et autres locutions*, Paris, Ophrys.
- GROSS, Maurice. 1975. *Méthodes en syntaxe*, Paris, Herman.
- GROSS, Maurice. 1982. « Une classification des phrases « figées » du français », *Revue québécoise de linguistique* 11/2, 151-185.
- KLEIN, Jean-René. LAMIROY, Béatrice. 1995. « Les expressions figées du français de Belgique », dans J. Labelle et Chr. Leclère (éds.), *Lexiques-grammaires comparés en français : Actes du colloque international de Montréal*, Amsterdam, Benjamins, 37-52.
- LAMIROY, Béatrice. KLEIN, Jean-René. LECLERE, Christian. LABELLE, Jacques. 2003. « Expressions verbales figées dans quatre variétés du français : le projet BFQS », *Cahiers de lexicologie* 83, 153-172.

- LECLERE, Christian. 2000. « Expressions figées dans la francophonie : le projet BFQS », *BULAG* (numéro hors série), 321-331.
- LAPORTE, Éric. 1988. « La reconnaissance des expressions figées lors de l'analyse automatique », *Langages* 90, 117-126.
- MEJRI, Salah. 2000. « Syntaxe et figement ». *Bulag* (Mélanges offerts à Gaston Gross numéro hors série), 333-342.
- MEJRI, Salah. 2002. « Le figement lexical : nouvelles tendances ». *Cahiers de lexicologie* 80, 213-225.
- MEJRI, Salah. 2003. « Le figement lexical », *Cahiers de Lexicologie* 82, 23-39.
- MOON, Rosamund. 1998. *Fixed expressions and idioms in English : a corpus-based approach*. Oxford, Clarendon press
- PAUMIER, Sébastien. 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse non publiée, Université de Marne-la-Vallée.
- SENEILLART, Jean. 1998. « Reconnaissance automatique des entrées du lexique-grammaire des expressions figées », *Travaux de linguistique* 37, 109-125.
- SILBERZTEIN, Max. 1998/1999. « Traitement des expressions figées avec INTEX », *Analyse lexicale et syntaxique. Le système INTEX, Lingvisticae Investigationes* 22, 425-449.
- SMADJA, Frank. 1993. « Retrieving collocations from text: X-tract », *Computational Linguistics*, 19/1, 143-177.
- ZINGLE, Henri. 2003. « Constitution et exploitation d'une base de données phraséologiques en français », *Actes du 17^e Congrès international des linguistes*. Prague.

Adresse des auteurs :

Anne Dister – Cédric Fairon – Patrick Watrin

Université catholique de Louvain, Cental

Collège Érasme, Place Blaise Pascal

1348 Louvain-la-Neuve

Belgique

{dister,fairon,watrin}@tedm.ucl.ac.be