Les syntagmes définis dans le TLFi : Normalisation et classification pour le traitement automatique des langues

Evelyne Jacquey, Jean-Marc Humbert, Line Heckler

Les syntagmes définis dans le TLFi : Normalisation et classification pour le traitement automatique des langues

Evelyne Jacquey[1], Line Heckler[1], Jean-Marc Humbert[1]
[1]*UMR ATILF - CNRS - Université de Lorraine*
Prenom.Nom@atilf.fr

1. Introduction

Les recherches linguistiques ou de type TAL autour de la question de la phraséologie connaissent un renouveau accru depuis une dizaine d'années. Parmi les indicateurs de cette dynamique, renouvelée avec l'ouvrage de G. Gross (1996) sur le figement syntaxique, on peut citer de nombreux travaux donnant une vision synthétique (Bolly, 2008), (Mejri, 2010), (Bolly, 2012), ou des propositions plus ciblées comme (Grossman et Tutin, 2003), (Blumenthal, 2006), (Polguère, 2008), (Mel'čuk, 2010) et (Mel'čuk et Polguère, à paraître).

Dans un contexte TAL, une ressource sur la phraséologie du français aurait de nombreux avantages utilisée en prétraitement pour améliorer le traitement des textes : étiquetage morphosyntaxique (TreeTagger), analyse syntaxique en constituants ou plus élaborée, annotation sémantique et profilage thématique, détection de candidats termes en texte intégral, etc. Du point de vue de la segmentation par exemple, une ressource en phraséologie du français enrichirait le lexique des expressions complexes (notamment concernant ce que (Bolly, 2012) appelle à la suite de (Wray, 2002, 2007) les unités phraséologiques référentielles -*avoir peur*, *point de vue*, *à la louche*-, organisationnelles -*au vu de*, *dans la mesure où*- et interactionnelles -*à l'aide!*-).

Dans cette proposition, nous décrirons comment nous avons construit la ressource TLFPHRASEO[1] à partir de la ressource des syntagmes définis du TLFi, son état actuel d'avancement ainsi que sa mise en oeuvre pour la reconnaissance de ces syntagmes en texte intégral.

2. Présentation de la ressource des syntagmes définis dans le TLFi

Les syntagmes définis du TLFi se présentent sous la forme de 56649 objets lexicographiques particuliers encadrés par des balises xml <syntita n="d"> et bénéficiant d'une définition lexicographique spécifique.

Avoir peur *de* (suivi d'un verbe à l'inf.). *Hésiter, répugner à (faire quelque chose).*

Outre les nombreux intérêts de cette ressource, son utilisation dans un contexte TAL comme dans des études linguistiques futures nécessite deux étapes incontournables : la normalisation formelle des intitulés de ces syntagmes et une classification phraséologique minimale. En effet, sous cette étiquette unique de 'syntagme défini', les lexicographes du TLF ont regroupé l'ensemble de ces catégories. A l'heure actuelle, seule l'étape de normalisation est achevée.

3. Normalisation formelle

En l'état les syntagmes définis du TLFi présentent plusieurs difficultés pour une reconnaissance automatique :

- Présence d'informations métalexicographiques

  Avoir peur de [$_{meta}$(suivi d'un verbe à l'inf.)$_{meta}$] => *Avoir peur de*

  Abaisseur de la langue[$_{meta}$(Ac. Compl. 1842 ; BESCH. 1845 ; LITTRÉ)$_{meta}$] => *Abaisseur de la langue*

- Séquences mono-lexicales supprimées (*Arrivée* dans *Arg. milit.* ***Arrivée***. Arrivée d'obus) sauf si un doute subsistait (report à l'étape de classification, *L'acier* dans *TECHNOL.* ***L'acier***. Industrie, le grand commerce de l'acier).

1 Cette ressource sera mise à disposition de la communauté dans son premier état intermédiaire, version normalisée non classifiée, sur le site du CNRTL : http://www.cnrtl.fr/lexiques/tlfphraseo.

- Traitement des alternatives manuellement : non automatisables

Abattre du minerai, de la houille	*Abattre du minerai* *Abattre de la houille*
Fer, acier cendreux	*Fer cendreux* *Acier cendreux*
Champ de feu, de tir d'une arme	*Champ de feu d'une arme* *Champ de tir d'une arme*

A l'issue de cette étape, la ressource atteint le nombre de 72 100 formes normalisées de syntagmes définis. La section suivante présente l'utilisation qui peut être faite de la ressource en l'état TLFPHRASEO, c'est-à-dire indépendamment sa classification phraséologique à venir.

4. Reconnaissance automatique des syntagmes définis normalisés dans les textes

La première exploitation que nous avons mise en oeuvre se veut avant tout robuste. Elle consiste à rechercher les syntagmes définis et normalisés de la ressource TLFPHRASEO dans un texte, préalablement étiqueté (en morpho-syntaxe) et lemmatisé par TreeTagger[2]. Chaque phrase est analysée selon les trois étapes :

- Délimitation de l'ensemble de tous les syntagmes contenant au moins un item de la phrase analysée : dans la phrase *On le sait, le quartier Oudinot subira la plus importante* [...], le pronom *on* apparaît dans 187 syntagmes de la ressource.

Mot	on	le	savoir	,	le	quartier	Oudinot	subir	le	plus	important
Nb Syntita	187	15551	174	21	15551	41	0	11	15551	380	0

- Restriction au sous-ensemble des syntagmes dont tous les éléments sont présents dans la phrase analysée : dans *Avec cette déclinaison en ligne , « faite maison » par la petite équipe qui s'occupe du site au quotidien , nous ne célébrons pas les vertus de l'interactivité , ni ne proposons de mirifiques contenus exclusifs : il s'agit d'abord de vous permettre de lire le journal , partout et le mieux possible*[3], le système retourne 7 séquences candidates.

1- Syntita possible : en ligne => en ligne
2- Syntita possible : les petites maisons => maison le petit NON RETENU !
3- Syntita possible : les petites - maisons => maison le petit NON RETENU !
4- Syntita possible : au quotidien => au quotidien
5- Syntita possible : le pas => pas le NON RETENU !
6- Syntita possible : lire le journal => lire le journal
7- Syntita possible : les possibles => le possible

- Prise en compte de l'ordre linéaire.

2 Pour cet étiquetage, nous avons utilisé le nouveau fichier de paramètre constitué récemment à l'Atilf pour l'écrit et l'oral (cf projet PERCEO : http://www.cnrtl.fr/corpus/perceo/).

3 Extrait de l'éditorial du Monde Diplomatique du 29 avril http://www.monde-diplomatique.fr/2013/05/A/49059

0- Syntita trouvé : en ligne => en ligne (Syntita n° 53944)
--- > tronçon de phrase : en ligne

0- Syntita trouvé : au quotidien => au quotidien (Syntita n° 32541)
--- > tronçon de phrase : au quotidien

0- Syntita trouvé : lire le journal => lire le journal (Syntita n° 20818)
--- > tronçon de phrase : lire le journal

0- Syntita trouvé : le possible => les possibles (Syntita n° 79680)
--- > tronçon de phrase : le mieux possible

Ces trois étapes permettent d'avoir un rappel important (rappel de 1 sur l'extrait ci-dessus).

En revanche, comme le montre la quatrième séquence, *le mieux possible*, la précision doit être améliorée. À cette fin, il est envisagé d'appliquer des règles fondées sur les patrons morpho-syntaxiques des syntagmes définis de la ressource (par exemple, un syntagme ne contenant aucune catégorie majeure comme *de plus* devra être retrouvé dans sa totalité) ou sur les différentes séquences candidates (par exemple, une séquence comme *croire a priori nécessaire* sera mise en correspondance avec le syntagme *croire nécessaire* parce que les éléments du syntagme appartenant à une catégorie majeure sont présents).

5. Classification

La classification est encore à l'état prospectif (rédaction d'un guide non encore appliqué à l'échelle). A l'heure actuelle, nous envisageons de classer les syntagmes de la ressource normalisée selon la typologie adoptée par deux projets connexes Definiens et RLF au sein du laboratoire, inspirée de la LEC (Mel'čuk, 2010) et (Mel'čuk et Polguère, à paraître). Trois catégories de haut niveau sont différenciées :

1. **les locutions** : sont syntaxiquement et paradigmatiquement contraintes et sémantiquement opaques (c'est-à-dire non compositionnelles), par exemple *enculer les mouches, couper les cheveux en quatre*.
2. **les collocations** : sont paradigmatiquement contraintes mais syntaxiquement libres et sémantiquement compositionnelles.
3. **les clichés linguistiques** : sont une sorte de phrasèmes qui correspondent à des situations particulières toujours associées à une même expression linguistique, par exemple *Au secours !*

Dans cette typologie, deux axes de classification apparaissent structurants : la liberté syntaxique et la liberté paradigmatique.

La liberté syntaxique est testée via plusieurs transformations phrastiques ou à l'intérieur du groupe nominal (Gross, 1996). A la suite de Gross, Polguère et Mel' čuk, le degré de liberté syntaxique constaté est utilisé de manière scalaire (variations en intensité et en portée). Les séquences totalement figées correspondent le plus souvent aux locutions et aux clichés linguistiques.

- *avoir les yeux plus gros que le ventre*
- *à fond la caisse*
- *Au secours !*

Les séquences dont seul un sous-ensemble est figé correspondent plutôt à des collocations, simples ou mixtes.

- *toucher **en plein dans le mille***
- *chevaucher **à tombeau ouvert***

Afin de déterminer si l'on est bien face à une collocation ou non, la liberté paradigmatique peut être testée en utilisant deux indices applicables au collocatif et à son pivot.

Les éléments d'une collocation relèvent d'une fonction lexicale définie dans la LEC.

Le nombre relativement de quasi-synonymes est réduit avec les collocations.

6. Conclusion

En conclusion, cette proposition décrit une ressource phraséologie du français en cours de développement. Elle est issue des syntagmes définis du TLFi. Dans sa version normalisée, elle comporte 72100 formes normalisées et lemmatisées. Leur classification selon la typologie de Polguère et Mel' čuk (2010 ; à paraître) est en cours.

7. Bibliographie

Blumenthal, P. et Haussmann, F-J. (2006), Collocations, corpus, dictionnaires, *Langue française*, 150, Paris : Larousse.

Bolly, C. (2008), *Les unités phraséologiques : un phénomène linguistique complexe ?* (Thèse de doctorat non publiée), Louvain-la-Neuve, Université catholique de Louvain.

Bolly, C. (2012), Flou phraséologique, quasi-grammaticalisation et pseudo marqueurs de discours : un no man's land entre syntaxe et discours, *LINX 62 (première épreuve).*

Gross, G. (1996), *Les expressions figées en français. Noms composés et autres locutions*, Editions OPHRYS.

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

Mejri, S. (2010), Structuration sémantique des séquences figées, dans *Les configurations du sens*, (Blumenthal, P. et Mejri, S. éds), Zeitschrift für französische Sprache und Literatur - Beihefte. Neue Folge (ZFSL-B), n°37.

Mel'čuk, I. (2010), Phrasology: It's place in the Language, in the Dictionary, and in Natural Language Processing, présenté à la *10th International Conference on Greek Linguistics*, 1-4 Septembre 2011, Komotini.

Polguère, A. (2008), *Lexicologie et sémantique lexicale, Notions fondamentales*, Nouvelle édition revue et augmentée, Presses Universitaires de Montréal.

Polguère, A. et Mel'čuk, I. (à paraître), *Lexique [Chap 2 : Entités lexicale du type "expressions" : phrasèmes.*

Wray, A. (2002), *Formulaic language and the lexicon*, Cambridge, Cambridge University Press.

Wray, A. (2007), 'Needs only' analysis in linguistic ontogeny and philogeny, dans Lyon, C., Nehaniv, L. et Cangelosi, A. (éds.), *Emergence of communication and language (Part 1)*, Londres, Springer Verlag, p. 53-70.