Coluccia, Rosario/Brincat, Joseph M./Möhren, Frankwalt (éd.) (2016). *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 5 : Lexicologie, phraséologie, lexicographie*. Nancy, ATILF : http://www.atilf.fr/cilpr2013/actes/section-5.html.

Phraséologie historique et outils d'extraction pour un corpus non-standardisé. Le journal de ma vie de Jacques Louis Ménétra.

Résumé

L'identification des collocations et des phraséologismes dans un corpus est "a pain in the neck" (cf. Filatkina 2009) dans la domaine de la linguistique informatique et de la linguistique du corpus. Un corpus non-standardisé renforce les problèmes méthodologiques. Une approche traditionelle (étiqueter le corpus (' part-of-speech-tagging '), définir des patterns et tester si les résultats d'extraction sont significatifs) est très vite limitée. Cette communication a pour but deux objectifs : Premièrement, nous proposons une méthode pour améliorer l'identification des collocations et des phraséologismes dans un corpus non-standardisé en utilisant un outil avec une implementation des expressions régulières (chapitre 3). Deuxièment, on discutera la hypothèse que la combination d'une approche quantitative et qualitative est utile à l'analyser la contribution des phraséologismes pour créer textualité (chapitre 4).

1. Introduction

Le but de la communication se situe autour de deux axes : Le premier se penche sur la méthodologie permettant d'identifier des collocations et des phraséologismes dans un corpus non-standardisé et le second est dédié à la contribution des phraséologismes pour la formation d'un texte. Le terme de phraséologisme est ici défini dans la lignée de Burger / Buhofer / Sialm 1982. Dans ce cadre, il y a deux critères obligatoires : la polylexicalité et le figement. Le critère facultatif – (l'idiomaticité) – concerne la sémantique et implique que le sens d'un phraséologisme n'est pas déductible de la somme de ses composants.

Dans la première partie de la contribution, le corpus sera présenté et son choix sera justifié (chapitre 2). Dans la seconde partie, la méthode employée pour l'extraction des phraséologismes sera légitimée (chapitre 3). La dernière partie (chapitre 4) éclaire la relation entre la phraséologie et la textualité à l'aide d'exemples issus du corpus.

2. Le corpus

Afin de tester une méthode adaptée à l'analyse de graphies non-standardisées et dans le but d'examiner l'emploi authentique des phraséologismes, l'analyse part d'un corpus du dix-huitième siècle qui comprend environ quatre-vingt mille tokens, le journal de ma vie de l'artisan Jacques-Louis Ménétra. Dans ce texte privé, l'auteur décrit les différentes étapes de sa vie : Ménétra est originaire de Paris, il exerce la profession de vitrier et il est enclin aux liaisons amoureuses extra-conjugales. En somme, Ménétra est peu lettré. Cela est évident au niveau d'écriture. En général, au XVIIIème siècle, il n'y avait pas encore de norme orthographique obligatoire. Dans le cas particulier de Ménétra, son écriture est marquée par l'idiosyncrasie. A titre d'exemple, on pourrait évoquer que l'auteur n'utilise pas la ponctuation de façon cohérente, et ne respecte pas l'écriture en majuscule et minuscule. Autre particularité, il écrit plusieurs mots en un seul mot. Ces caractéristiques compliquent le processus d'identification automatique des phraséologismes dans le corpus.

3. Quelques méthodes quantitatives et leur efficacité

En ce qui concerne le choix d'une méthodique permettant d'extraire les phraséologismes, il faut tout d'abord soulever les deux questions suivantes : 1. Comment identifier de manière systématique les phraséologismes typiques et spécifiques d'un corpus? 2. Comment réduire le problème de l'écriture non-standardisée qui complique le processus d'identification des phraséologismes?

3.1. Extraction basé sur des 'patterns' avec 'log likelihood ratio'

Il y a en somme aujourd'hui plus que 50 'association measures' qui permettent d'extraire de manière systématique les phraséologismes des corpus (cf. Evert 2008, 1243). Même après plus que 50 ans de recherche en linguistique informatique, trouver le test adéquat n'est pas un but de recherche trivial. Afin d'illustrer ce point, nous aborderons de façon exemplaire, une association measure bien connue, 'log likelihood ratio' (Dunning 1993). Pour l'application, une solution qui s'établie : étiqueter le corpus ('part-of-speech-tagging'), définir des 'patterns' et tester si les résultats d'extraction sont significatifs. Toutefois ce procès est limité comme on peut voir au tableau 1 :

1	dit	100	que	135	33	188.09614
2	mon	148	pere	42	26	184.95002
3	que	135	cest	87	28	56.75925
4	ce	35	quil	119	19	135.66373
5	un	251	jour	31	20	121.81455
6	chez	59	lamere	11	10	104.13436
7	la	150	conduite	14	12	97.983083
8	que	135	cettoit	12	11	97.777167

9	mes	49	adieux	13	9	87.209868
10	un	251	coups	17	13	86.364489
11	avec	62	moy	79	14	81.325985
12	douze	9	franc	10	6	79.933036
13	jeluy	44	dit	100	13	76.766786
14	quy	96	etoit	48	13	75.276444
15	chez	59	luy	58	12	74.636016

Tableau 1 : Extrait des résultats d'extraction avec log likelihood ratio; patterns : verbe-substantif

Le tableau 1 présente les résultats du pattern verbe-substantif avec la méthode ' log likelihood ratio '. On peut cependant constater que les résultats comme dit que ou mon pere ne correspondent pas au pattern verbe-substantif. En effet, la liste contient d'autres classes de mots : des prépositions, des adjectifs numéraux etc.

Les résultats de la première étape (tableau 2) – le corpus étiquetté avec le ' treetagger '– présentent également un haut pourcentage d'erreurs en ce qui concerne le classement de la catégorie grammaticale et de la forme de base pour chaque ' token ' :

texte	NOM	texte
situvas	VER :simp	‹unknown›
cequejay	NOM	‹unknown›
delarepublique	ABR	‹unknown›
Maprofessiont	NOM	‹unknown›
jenefut	NOM	‹unknown›

Table 2 : Extrait du corpus étiqueté

Ce résultat est dû à l'écriture idiosyncratique de Ménétra. L'outil ' treetagger ' se base sur un dictionnaire. La norme orthographique joue donc un rôle important. En raison des divergences formelles par rapport au français standardisé, il est souvent impossible d'identifier automatiquement la forme de base et la catégorie grammaticale pour beaucoup de tokens. En conséquence, les résultats de l'extraction basés sur des patterns contiennent de nombreuses erreurs. C'est pourquoi nous proposons une autre méthode.

3.2. Réduire les difficultés de l'écriture idiosyncratique avec des expressions régulières

Pour répondre au problème de l'écriture idiosyncratique et non-standardisée, nous avons recours à une technique appelée expressions régulières. Il s'agit d'une technique efficace qui présente de multiples possibilités de décrire des chaînes de caractères permettant de trouver des structures dans un texte. Les expressions régulières suivent une syntaxe stricte. Au fond, chaque caractère d'alphabet garde

sa signification. On peut ajouter quelques caractères spéciaux, les métacaractères.[1]1 Les quantifieurs se réfèrent au caractère précédent. Un astérisque (*) signifie que le caractère précédent doit se produire 0 fois ou plus. Un point d'interrogation (?) permet que le caractère précédent doit se produire 0 ou 1 fois. Ce qui est utile aussi, c'est un autre métacaractère, le point (.) qui signifie n'importe quel caractère. Par exemple, l'ensemble des mots latette, atette, satette et tette peut être comprimée en la forme (.*tette) pour trouver les variantes graphiques de tête. Il existe de nombreuses implémentations d'expressions régulières. Une entre elles est le logiciel de concordance AntConc. A l'aide de la fonction ' regex ', on peut d'un côté chercher des variantes dans la graphie et de l'autre on peut dépister les modifications d'un phraséologisme.

3.3. Travailler corpus-driven et corpus-based pour identifier des phraséologismes

AntConc permet d'identifier des phraséologismes. Le point de départ est la comparaison du corpus d'analyse avec un corpus de référence[2] pour lequel il est possible d'identifier les mots-clés ayant recours à la méthode ' log likelihood ratio '. Cette approche est dite ' corpus-driven ' : c'est-à-dire qu'on ne teste pas un hypothèse préalable, mais au contraire, de cette manière, on peut plutôt découvrir les caractéristiques typiques d'un corpus. Ainsi, il est possible de trouver des structures jusqu'à présent inconnues et de formuler des hypothèses originales (cf. Bubenhofer 2009, 99sqq.). Le graphique 1 présente un extrait des résultats de la comparaison, visualisé avec wordle. Tandis que l'ordre des mots est fortuit, la taille des mots correspond à sa valeur significative :

Figure 1 : Un extrait des mots-clés

[1] Pour une description détaillée des expressions régulières, voir Friedl 2008.

[2] Le corpus de référence employé pour la présente étude comprend des cahiers de doléances de 1789 avec 85.675 ' tokens '.

Nous complétons cette procédure inductive par une perspective déductive. Autrement dit, nous avons recours à l'approche ' corpus-based '– qui est la seconde étape obligatoire selon Bubenhofer (2009, 102) : Nous vérifions les mots-clés – s'ils sont un composant d'un phraséologisme – avec les méthodes ' mutual information 'et ' T-Score 'qui sont implémentées en AntConc. On peut réduire les désavantages[3] des association measures avec leur combinaison. Cela permet de trouver d'une part des phraséologismes en général moins fréquents et d'autre part ceux qui sont plus nombreux.

4. L'analyse

Les résultats de l'extraction sont à la base de l'analyse qualitative des phraséologismes qui se focalise sur la textualité. Les questions centrales sont :

1. Comment fonctionne le lien entre les phraséologismes et son contexte?
2. Quels sont les mécanismes qui aboutissent à la textualité?
3. Quelles sont les fonctions des phraséologismes dans le corpus?

4.1. Le cadre théorique : Le concept du potentiel de formation textuelle

Avec Gardt (2002), un texte peu être défini comme un système émergent : on ne peut pas comprendre son sens à partir du sens de chacun de ses composants. C'est-à-dire que le sens d'un texte et d'un phraséologisme dépasse la somme de ses parties. Le point commun entre la textualité et la phraséologie est que leur sens n'est pas compositionnel : le sens de l'expression dépasse la somme de ses parties.

Dans un texte, les phraséologismes jouent un rôle important. Ils marquent un texte sur les niveaux d'organisation formelle, communicative et de contenu (Gréciano 1994, 94). Pour examiner les phraséologismes dans le corpus, l'analyse se rapporte au niveau théorique à la théorie du potentiel de formation textuelle (en allemand ' textbildende Potenzen '). Ce concept presque inconnu hors de la frontière germaniste, est issu de la recherche sur la phraséologie soviétique des années 1970 et a été développé en premier lieu par Dobrovol'skij (1980), Gréciano (1989 et 1994), Fleischer (1997) et Sabban (2004 et 2007).

Le concept est centré autour du potentiel phraséologique de structurer et d'organiser un texte. Les caractéristiques phraséologiques – la polylexicalité, la stabilité relative au niveau morpho-syntaxique et sa sémantique complexe – sont responsables de ce potentiel. En outre, leur distribution joue un rôle important. En effet, il y aurait une corrélation entre certains phraséologismes et les genres textuels. Les observations centrales gagnées avec le concept de potentiel de formation textuelle sont les suivantes :

[3] «Mutual information estimates based directly on counts are subject to overestimation when the counts involved are small» (Dunning 1993, 62).

1. Le potentiel et sa réalisation dépendent du type de phraséologisme, de la situation communicative etc. (cf. Fleischer 1997, 214-215). C'est-à-dire que plusieurs types phraséologiques disposent d'un potentiel, mais sa réalisation n'est pas obligatoire.
2. Les phraséologismes sont un mécanisme de la récurrence. Ils sont un moyen pour créer cohérence et cohésion dans un texte (cf. Gréciano 1994, 95). Donc, ils mènent à la textualité.

4.2. L'analyse de quelques exemples du corpus

4.2.1. Jeux de mots

Pour l'analyse, nous avons choisi un extrait des phraséologismes qui réalisent le potentiel de formation textuelle. Une des particularités centrales de quelques phraséologismes réside en leur potentiel ludique. Pour enchaîner des phrases de manière linéare, Ménétra emploie souvent la technique de jeux de mots. Voici le premier exemple :

(1) jevois unhomme decoree que ledomestique traite de monseigneur quy me demande si jene suis point coureur jeluy reponds que non il me propose daitre lesien et demontee dans son carosse jeluy dit jevous entends jesuis coureur defille et non autrement adieu (217)

Ménétra reçoit l'offre de travailler comme coureur, en d'autres termes comme domestique. Il refuse l'offre en reprenant dans sa réponse un mot simple déjà évoqué – coureur. Ce mot simple fait partie du phraséologisme suivant être coureur de filles qui signifie 'celui qui court de femme en femme, a de nombreuses aventures galantes'. De cette manière, Ménétra explique qu'il est uniquement à la recherche de liaisons amoureuses. C'est-à-dire Ménétra prend l'idiome au premier degré pour le nouer avec le mot simple coureur. En outre, coureur est une cataphore de coureur de filles. Avec cette technique, l'auteur emploie le potentiel ludique des phraséologismes pour démontrer qu'il peut s'exprimer avec aisance et pour déclencher une fonction affective d'un phraséologisme. Inversement, il y a aussi des usages anaphoriques :

(2) nous trouvont une espese dangart ou son des tonneaux pres des privee nous nous metons ajouer a cache cache nicolas apres les petite fason dusage en pareil ocasion nous etions bien cachee (244)

Le vitrier a fait la connaissance d'une femme et il couche avec elle dans un jardin. Il exprime cela avec le phraséologisme jouer a cache-cache Nicolas. A l'origine, cache-cache mitoulas désigne un jeu d'enfants qui consiste à se passer secrètement un objet que le joueur désigné doit localiser. Par erosion au niveau phonétique, on reçoit cache-cache Nicolas. Par ailleurs, la désignation du membrum virile avec des prénoms masculins est un procédé usuel (Ernst / Wolf 1997, 217). Ménétra commente son plaisir avec un composant du phraséologisme – cachee – qu'il prend au premier degré. En outre, cachee est sémantiquement lié avec le phraséologisme ajouer a cache cache nicolas par une anaphore fidèle (cf. Adam 2008, 85).

Enfin, dans l'exemple (3), le mécanisme fonctionne de la même manière que dans l'exemple (2) :

(3) jelorgnois depres une jolie repasseuse quy demeuroit dans notre maison et que je desiroit bien repassee (112)

Ménétra s'intéresse à une jeune femme qui travaille comme repasseuse. Il exprime son intérêt avec le phraséologisme métonymique jelorgnois depres une jolie repasseuse. Son intention est claire, il veut passer une nuit avec elle. Pour exprimer ce tabou, il emploie une stratégie qui peut être caractérisée d'euphémisme : il reprend un composant du phraséologisme – repasseuse – et dit qu'il veut la repasser. Cette expression est une métaphore euphémique pour l'acte sexuel.

En somme, cette stratégie déclenche d'un côté une fonction affective et de l'autre, elle garantit la progression textuelle.

4.2.2. Isotopie

Il y a aussi d'autres mécanismes. Gréciano (1989, 418) a montrée que certains auteurs combinent des phraséologismes avec des termes similaires dans le contexte antérieur ou postérieur. La fonction principale de cette technique est de produire de la cohérence dans un texte.

Dans l'exemple (4), Ménétra combine un mot simple avec un phraséologisme :

(4) lafemme cris que jene peut restee ala maison [...] apres un baiser donnee quy est rue tous deflame quantre mr et madame letorchons brule tres fort (280)

Ménétra fait la connaissance d'une femme mariée. Cette dernière commet un adultère avec Ménétra. La liaison entre la femme et le vitrier est marquée par la passion, comme en témoigne la métaphore un baiser de flame. Au niveau littéral, flame entretient une relation d'isotopie avec le phraséologisme suivant, le torchons brûle qui signifie 'il y a une querelle'. Ce phraséologisme – qui date du fin du XVIIIème siècle – se réfère à l'homme trompé et à sa femme. L'interprétation qui s'impose spontanément est celle d'une querelle domestique, symbolisée par l'image du torchon ménagée. En fait, le mot est à prendre dans son acceptation ancienne de petite torche (Rey / Chantreau 2007, 874). L'effet de ce montage d'isotopies est un contraste qui contribue à la progression textuelle et qui a une fonction affective et ironique : La passion de l'amour est opposée à la passion d'une dispute entre un couple.

Dans (5), il y a une accumulation de deux idiomes :

(5) javois un competiteur oupour mieux dire un concurent quy secroioit audesus detout etant brigadier duguet chacque fois quil me voyoit parle acette aimable fille il vouloit men imposer moy jeluy montroit les dent il me disoit quil avoit les bras longs (112)

Ménétra et son rival désirent la même femme. A l'égard de son concurrent, Ménétra fait une démonstration de sa force avec le phraséologisme montrer les dents, qui signifie 'être menaçant'. Il s'agit d'un somatisme, c'est-à-dire d'un phraséologisme qui

contient un membre du corps entre ses composants. Ce somatisme métonymique est motivé par l'image d'un chien qui gronde en montrant ses crocs (Rey / Chantreau 2007, 299). Les dents symbolisent l'offensive.

La réplique du concurrent – avoir les bras longs 'avoir un grand pouvoir' – participe à la cohérence du texte. Il y a en effet une relation d'isotopie entre les deux phraséologismes : il s'agit de montrer sa force pour gagner la bienveillance d'une femme. Au moyen des somatismes accumulés en (5), la force se rapporte vraiment au domaine physique, donc, les phraséologismes sont pris au sens littéral.

Voici un autre exemple :

(6) malgre ma mauvaise nuit jefut travailier entre dix et onze heur jevois arivee le domestique quy me demande si jebien ousassee lanuit en me ricanant japersois quil asa main mon bas je change de couleur [...] il me dit quil mavoit vue montee que cettoit le cocher quy mavoit vue et que lemaitre luy avoit bien lavee latette (175)

Ménétra a manqué l'occasion de passer la nuit avec une femme à cause d'un autre homme qui lui a joué un tour. Un autre domestique lui raconte ce qu'il s'est passé. Il s'avère que le cocher est responsable de son occassion manquée, mais ce dernier a reçu une punition, parce que son maître a pensé que le cocher ment : lemaitre luy avoit bien lavee latette. Ce somatisme signifie 'faire de violents reproches à qn'. Les autres parties du corps mentionées – comme main, bas etc. – déclenchent un réseau sémantico-pragmatique qui contribue à la textualité.

5. Conclusion et perspectives

Résumons. Pour répondre au problème de l'écriture idiosyncratique et non-standardisée, nous avons proposé d'utiliser des expressions régulières. Les méthodes quantitatives – appliquées à une perspective inductive et déductive – sont efficaces, parce qu'elles donnent des indices qui permettent l'identification des phraséologismes. Mais comme nous avions vu, une approche qualitative est indispensable.

A l'aide de la théorie du potentiel de formation textuelle – qui a servi de cadre explicatif quant à la contribution des phraséologismes à la textualité – nous avons vu que les phraséologismes interagissent avec le texte à différents niveaux. Ils contribuent à l'enchaînement linéaire du texte et développent la textualité tout en participant à la progression textuelle. En effet, les phraséologismes sont plus qu'une « pierre de construction » d'un texte. Avec des techniques comme les jeux de mots et les isotopies, ils développent une fonction affective par laquelle l'auteur fait preuve de son aisance et de sa virtuosité en ce qui concerne l'emploi de la langue.

Université de Kassel Sandra ISSEL-DOMBERT

Références

Adam, Michel, 2008. *La linguistique textuelle. Introduction à l'analyse textuelle des discours*, Paris, Armand Colin.

Bubenhofer, Noah, 2009. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*, Berlin/New York, de Gruyter.

Burger, Harald / Buhofer, Annelies / Sialm, Ambros (1982) : *Handbuch der Phraseologie*. Berlin / New York, de Gruyter.

Dunning, Ted, 2003. «Accurate methods for the Statistics of Surprise and Coincidence», *Computational Linguistics* 19/1, 61-74.

Ernst, Gerhard / Wolf, Barbara, 1997. «Jacques Louis-Ménétra (1738-1803 (?)) und sein „gibier féminin". Ein Mann über Frauen, Liebe und Sexualität», in : Dahmen, Wolfgang / Holtus, Günther / Kramer, Johannes (ed.) : *Sprache und Geschlecht in der Romania. Romanistisches Kolloquium X*, Tübingen, Narr [*Tübinger Beiträge zur Linguistik* 417], 205-221.

Evert, Stefan, 2008. «Corpora and collocations», in : Lüdeling, Anke / Kytö, Merja (ed.) : *Korpuslinguistik. Ein internationales Handbuch*, Berlin / New York, de Gruyter [*Handbücher zur Sprach- und Kommunikationswissenschaft* 29], 1212-1248.

Filatkina, Natalia, 2009. «Historische formelhafte Sprache als «harte Nuss» der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt», *Linguistik online* 39, 3/2009, <www.linguistik-online.de/39_09/filatkina.org>.

Fleischer, Wolfgang, 1997. *Phraseologie der deutschen Gegenwartssprache*, Tübingen, Niemeyer.

Friedel, Jeffrey E., 2008. *Reguläre Ausdrücke*, Beijing, O'Reilly.

Gardt, Andreas, 2002. «Wort, Text und Bedeutung. Aspekte der semantischen Erschließung von Texten», Ágel, Vilmos / Gardt, Andreas / Haß-Zumkehr, Ulrike / Roelcke, Thorsten (ed.) : *Das Wort. Seine strukturelle und kulturelle Dimension*, Tübingen, Niemeyer, 111-132.

Gréciano, Gertrud, 1989. «Le signe idiomatique et la production textuelle», in : Tobin, Yishai (ed.) : *From sign to text. A semiotic view of communication*, Amsterdam / Philadelphia, John Benjamins [*Foundations of semiotics* 20], 415-425.

Gréciano, Gertrud, 1994. «Phraseologische Text(in)kohärenz», *Cahiers d'études germa*niques 27, 93-103.

Rey, Alain / Chantreau, Sophie, 2007. Dictionnaire d'expressions et locutions, Paris, Le Robert.

Sabban, Annette 2004. «Zur Rolle der Phraseme für die Konstitution und Funktion des Textes. Ein Beitrag zum Konzept der textbildenden Potenzen», in : Steyer, Kathrin (ed.) : *Wortverbindungen – mehr oder weniger fest*, Berlin / New York, de Gruyter, 238-261.

Sabban, Annette, 2007. «Textbildende Potenzen von Phrasemen», in : Burger, Harald (ed.) : *Phraseologie. Ein internationales Handbuch*, Berlin / New York, de Gruyter [*Handbücher der Sprach- und Kommunikationswissenschaft* 28.1], 237-253.