VERS LA DÉFINITION PRAGMATIQUE DE LA COLLOCATION : MÉTHODES STATISTIQUES EXEMPLIFIÉES SUR LES ARTICLES JOURNALISTIQUES TRAITANT LA CRISE MIGRATOIRE

Michal Místecký

Université d'Ostrava
République tchèque
mmistecky@seznam.cz

Résumé. L'étude se focalise sur l'application des mesures d'association dans le domaine des textes journalistiques. Dans les articles du *Monde* qui se concentrent sur la crise migratoire, les métriques de la MI, de dés, et le score T ont calculé les collocatifs des expressions « migration », « migrant », « Méditerranée », « français / Français », et « mort ». La période de la recherche était octobre et novembre 2018. Pour que l'investigation soit effectuée sur un grand nombre d'articles, le logiciel *LancsBox*, développé par l'université de Lancaster, a été utilisé pour les calculs. Le but de l'étude est double : primo, de présenter les mesures d'association, et secundo, de les appliquer sur une recherche concrète d'un problème contemporain.

Mots clés. Collocation. Mesures d'association. Corpus. Journalisme. Migration.

Abstract. Towards a Pragmatic Definition of Collocation: Statistical Methods Exemplified on the Newspaper Articles Dealing with the Migration Crisis. The paper focuses on the application of association measures in the sphere of newspaper texts. In the articles of *Le Monde* which deal with the migration crisis, the metrics of MI, Dice, and t-score have calculated the modifiers

for the nodes “migration”, “migrant”, “Mediterranean [region]”, “French [both language, and adjective]”, and “dead / death”. The period of the research was October and November 2018. In order for the investigation to be carried out on a large number of articles, the *LancsBox* software, developed by the University of Lancaster, has been used for the calculations. The goal of the study is double: first, to present the measures of association, and second, to apply them in a piece of research focusing on a contemporary issue.

Keywords. Collocation. Association Measures. Corpus. Journalism. Migration.

1. Introduction

Parallèlement à l’évolution de la linguistique de corpus, l’on assiste à une quantification approfondissante des études langagières, qui amène avec elle une tendance à l’exactitude des définitions des termes lingustiques ; contrairement aux tentatives « métaphysiques », cette tendance est pourtant équilibrée par un accent porté sur l’usage pratique du langage, qui est réfléchi – de façon plus ou moins réussie – dans les corpus. La définition des unités langagières – telles morphèmes, phonèmes, mots et d’autres – ne doit donc pas se fonder sur les directives qui se différencient d’un linguiste à l’autre, mais peut gagner une valeur intersubjective ; ceci est atteint à travers l’usage des formules mathématiques partagées dans la communauté de chercheurs.

Le but de cette étude est d’exemplifier la méthodologie quantitative sur une recherche concrète, à savoir sur les articles du journal du *Monde* dans une période donnée, qui se portent sur le thème de la migration. Premièrement, la méthode du calcul sera présentée ; deuxièmement, on découvrira les paramètres de la recherche ; troisièmement, les résultats généraux de l’investigation seront traités ; quatrièmement, on montrera des visualisations graphiques des réseaux collocatifs et proposera quelques interprétations des chiffres obtenus ; et finalement, les conclusions du travail seront commentées.

2. Méthodologie

Dans la sphère des études du corpus, l’unité la plus discutée est la collocation. Il y avait plusieurs essais de la délimiter (cf. Firth, 1957 ; Coseriu, 1967 ; Manning et Schulze, 2000 ; Březina, McEnery et Wattam, 2015), mais étant donné ce qui a été mentionné, ce ne sont que les définitions intersubjectives qui peuvent être utiles dans l’étude sur le terrain. Quant à l’approche de Jan Šabršula, le linguiste, n’utilisant pas le terme de collocation, essaie d’unir toutes les locutions figées sous la dénomination *unité onomatologique complexe* (cf. Šabršula, 1983 ; Brňáková, 2013). Quoiqu’elle soit suffisamment générale, cette approche ne répond pas à la question importante – à savoir, si un tel syntagme est une collocation systémique.

Pour harmoniser les points de vue des auteurs individuels, on emploiera les formules qu’on appelle mesures d’association (cf. Křen, 2006). Il y en plusieurs, chacune ayant et des qualités, et des désavantages ; pour cette étude, l’on fait usage de trois mesures – le score de l’information mutuelle (MI), le score de dés, et le score T.

En ce qui concerne la MI, c'est une des métriques les plus utilisées dans le domaine (cf. Cvrček, 2015). Se fondant sur la notion de l'entropie, la formule en est la suivante –

$$MI\ (xy) = \log_2 \frac{N * f(xy)}{f(x) * f(y)}\ ; \tag{1}$$

N veut dire le nombre total des mots dans le corpus, $f(x)$ la fréquence de l'occurrence du mot x, $f(y)$ la fréquence de l'occurrence du mot *y* , et $f(xy)$ la fréquence de la co-occurrence des mots *x* et *y* Les chiffres dépassant 7 sont associées aux collocations qui peuvent être considérées comme systémiques ; néanmoins, le désavantage de cette méthode est qu'elle préfère des co-occurrences des mots très rares.

La deuxième mesure est la métrique de dés. Cette formule est assez intuitive, donnant, comme résultats, la probabilité de la co-occurrence des deux mots dans la collocation. Si, par exemple, il y a 10 occurrences du mot x et 10 occurrences du mot *y*, et qu'ensemble, ils apparaissent cinq fois, la métrique compte la probabilité de 0.5. Mathématiquement –

$$Dés\ (xy) = \frac{2f(xy)}{f(x) + f(y)}\ ; \tag{2}$$

la significance des symboles est la même que dans la formule précédente.

Finalement, l'on utilisera la métrique du score T, qui est traditionnel dans le domaine de la statistique. La formule en est la plus compliquée ; voici –

$$T\ (xy) = \frac{f(xy) - \frac{f(x) * f(y)}{N}}{\sqrt{f(xy)}}, \tag{3}$$

la significance des symboles étant la même que dans les formules précédentes. Le désavantage du score est le contraire de celui de la MI : les chiffres élevés apparaissent dans les collocations à haute fréquence dans le corpus ; celles-ci sont donc plutôt des colligations (cf. Firth, 1957) – des locutions figées à valeur grammaticale (par exemple « ces quatre pommes »).

3. Le corpus étudié et les principes de la recherche

Les métriques ci-haut mentionnées ont été appliquées dans la recherche des articles traitant la crise migratoire. Du *Monde*, quotidien français prestigieux, l'on a choisi tous les articles se trouvant sous la section « Immigration en Europe » (voir la Bibliographie) et publiés du 10 octobre jusqu'au 10 novembre 2018 (19 au total). En dehors de ces paramètres, l'on a exclu les reportages en images, les articles réservés pour les abonnés, et les discussions pour les internautes. En ce qui concerne les textes, les titres, les sous-titres, et les questions dans les entretiens font partie du corpus, parce que les collocations sont des expressions de tous ceux qui participent à la création d'un article. Pour que la collocabilité des mots soit

recherchée, leur occurrence minimale doit être *trois* dans tout le corpus ; les mesures vont couvrir un espace de cinq mots à gauche du mot-clé jusqu'à cinq mots à sa droite. Cela veut dire qu'on ne limite pas le terme de collocation à des expressions se trouvant l'une à côté de l'autre.

Les mots de tête[1] recherchés sont « migration », « migrant », « Méditerranée », « français / Français », et « mort ». La sélection est déterminée par l'orientation des articles de même que sur la base du sondage préliminaire dans le corpus. Donc, il s'agit d'une combinaison de mots qui sont généralement utilisés dans le contexte étudié et des expressions à valeur actuelle. La différence entre ces deux groupes se manifeste aussi dans les interprétations présentées.

La recherche elle-même s'effectuera à l'aide du logiciel *LancsBox*, qui a été désigné par l'Université de Lancaster pour les analyses du corpus (cf. Březina, McEnery et Wattam, 2015).

4. Résultats généraux

Pour délimiter un niveau de signification, on a décidé de prendre en compte les collocations gagnant MI > 7, dés > 0.1, score T > 3. Il faut souligner que dans le cas de la MI, la limite est le standard dans le domaine (voir ci-dessus), tandis que les niveaux des dés et du score T ont été choisis pour qu'on puisse obtenir d'autres résultats et présenter des orientations différentes de ces méthodes par rapport à la MI. Ces considérations sont à retenir pendant la recherche.

Les résultats sont récapitulés dans le Tableau 1. Pour commencer, il faut mentionner la prominence du collocatif « SOS », nom d'une organisation humanitaire qui s'occupe du sauvetage en mer, opérant avec le navire Aquarius[2] ; ceci est assez fréquent, d'après tous les calculs, dans l'environnement du lexème « Méditerranée ». Ce qui est particulier est le fait que c'est le seul collocatif qui a réussi à pénétrer dans le tableau avec ce lexème, cela indiquant une certaine « occupation » de la région géographique par l'affaire connectée avec cet ensemble.

L'occurrence de l'organisation SOS s'harmonise avec l'importance d'un autre groupe, Organisation internationale pour les migrations (OIM), selon les métriques de la MI et de dés. Depuis 2016, elle fait partie des Nations unies en tant qu'une des agences[3] ; sa mission est de traiter la question de la migration et de défendre les droits des migrants. De même, d'autres collocatifs – « organisation » et « internationale » – sont liés à l'OIM. L'idée est que dans les articles, il peut y avoir une tendance de souligner les activités des groupes pro-migratoires.

En revanche, les collocatifs apparaissant à la vicinité de l'expression « français / Français » montrent un autre phénomène. Les formules de la MI et des dés ont découvert la fréquente co-occurrence de l'expression avec le nom de Mamoudou Gassama, un migrant

[1] Le « mot de tête » est une traduction littérale du terme headword, qui est utilisé par le logiciel LancsBox (voir ci-dessous). Il veut dire le noyau (mot-clé) d'une collocation ; autour de lui, il y a des collocatifs.

[2] Pour en savoir plus, il est possible de consulter https://sosmediterranee.org/.

[3] Pour en savoir plus, il est possible de consulter https://www.iom.int/fr.

malien qui a sauvé un enfant après avoir escaladé quatre étages d'un immeuble parisien.[4] Cela lui a valu la nationalité française et le sobriquet de presse « Spider-Man français », ceci étant à l'origine de sa prominence dans la présente recherche. La focalisation sur l'histoire exceptionnelle d'un individu est un autre trait des journaux modernes[5] ; ici, la fonction en est de souligner la contribution que les immigrés peuvent apporter à la société française. Donc, elle vise le même but que celui recherché dans le paragraphe précédent.

La parole à comportement particulier est « mort » ; cela apparaît et comme le mot de tête, et comme un collocatif avec l'expression « migrant » (voir la métrique de dés). Comme les articles sont focalisés sur un évènement qui s'est déroulé fin octobre, le lexème colloque avec les enfants, les nombres, et la localité (« la Méditerranée »). Les journalistes s'occupent d'une affaire à valeur négative qui promet de gagner l'attention des lecteurs avec des opinions différenciées.[6] De cette façon, les articles gardent leur position de compassion avec des immigrés, utilisant, en même temps, une stratégie attaquant les émotions. De plus, le nombre des collocatifs du « mort » (d'après la métrique de dés) montre que le traitement de l'affaire est répétitif.

Pour conclure, il faut se concentrer sur la quasi-absence des collocatifs trouvés à travers le calcul du score T. Ce fait peut se comprendre, étant donné que cette formule ne découvre que des collocations très fréquentes, qui semblent absentes du corpus recherché. La présence de l'organisation « SOS », mentionnée ci-dessus, relève d'une exception qui a été déjà interprétée.

Mot de tête	MI	Dés	Score T
migration	OIM ; internationale ; organisation	internationale ; organisation ; OIM	–
migrant	–	mort ; plus	–
Méditerranée	SOS	SOS (70 %)	SOS
français / *Français*	Mamoudou ; Gassama ; intérieur	Mamoudou ; Gassama ; intérieur ; ans ; frontière	–
mort	treize ; enfants ; bord	treize ; enfants ; bord ; Méditerranée ; deux	–

Tableau 1 : Les résultats généraux des trois métriques utilisées dans la recherche.

[4] Pour en savoir plus, consultez http://www.leparisien.fr/paris-75/paris-il-escalade-un-immeuble-pour-sauver-in-extremis-un-enfant-suspendu-dans-le-vide-27-05-2018-7738266.php.

[5] Le point de vue critique est présenté ici : http://www.mediaculture.fr/le-story-telling-contre-linformation/.

[6] Ceci est commenté, par exemple, ici : https://www.theguardian.com/commentisfree/2018/feb/17/steven-pinker-media-negative-news.

Un autre point de vue, plus minutieux, est présenté dans le Tableau 2. Ici, il y a la liste de collocatifs rangés d'après les valeurs de la MI ; le montant du chiffre indique le niveau du caractère systémique de la collocation. Comme déjà mentionné, l'on trouve que les organisations internationales qui s'occupent de l'aide proposée aux migrants se lient avec les mots de tête très étroitement ; il semble que la tendance vers les collocations est assez saillante dans le cas de « migration », qui prend les trois premiers rangs sur quatre. Ceci peut être expliqué par la focalisation des articles sur la situation actuelle, mais aussi par la structure de l'abréviation (« OIM » veut dire « Organisation internationale pour les migrations »). Au contraire, il n'y a aucune collocation systémique avec « migrant » – en tant qu'adjectif ou même substantif –, ce qui indique un usage varié des expressions qui se trouvent dans sa vicinité. La faiblesse de sa force collocative peut signifier une polysémie du terme et son adaptabilité contextuelle.

Mot de tête	**Collocatif**	**MI**
migration	OIM	9.68
Méditerranée	SOS	9.27
migration	internationale	8.9
migration	organisation	8.81
mort	treize	8.68
mort	enfants	8.46
français / *Français*	Mamoudou	7.76
français / *Français*	Gassama	7.76
français / *Français*	intérieur	7.6
mort	bord	7.4
migrant	–	0

Tableau 2 : Les chiffres de la MI des collocations les plus systémiques.

5. Visualisations graphiques et commentaires

Pour que l'on puisse bien s'orienter dans les résultats, on les complétera avec des visualisations graphiques. Celles-ci comprennent le mot de tête, autour duquel il y des rayons indiquant les collocatifs, la saturation de la teinte signifiant leurs fréquences dans le corpus. Les collocatifs sont rangés de gauche (L) à droite (R) ; si rien n'est marqué, la position est variable. Ces visualisations permettent d'analyser les résultats plus en profondeur.

À titre d'exemple, la Figure 1 représente le réseau collocatif de l'expression ci-haut mentionnée, « migrant », selon la métrique de dés ; contrairement aux principes de la recherche, la fréquence minimale de l'occurrence d'un collocatif a été déterminée à 5 (f > 5), pour que la visualisation soit plus observable. Ici, on peut voir que les collocatifs les plus importants – « personnes », « mort » (L tous les deux) – sont très proches du mot de tête, le traitant soit comme adjectif, soit comme substantif. De plus, la plupart des collocatifs sont des expressions grammaticales – auxiliaires, prépositions, ou articles ; il y a peu à dire à propos de leurs occurrences, excepté l'usage du verbe « avoir » (« a », « ont »), qui exprime le passé composé des actions des migrants. Néanmoins, il est probable que l'usage de ce temps verbal est caractéristique du langage journalistique en général.

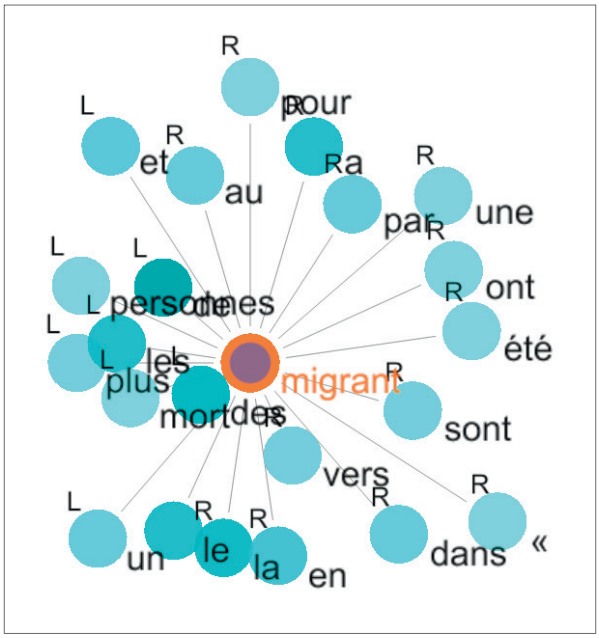

Figure 1 : Le réseau collocatif du mot de tête « migrant » (métrique de dés, f > 5).

Ensuite, on traitera le réseau collocatif du mot « français / Français », toujours selon la métrique de dés. À l'exception des collocatifs déjà analysés, il y aussi des expressions comme « frontière » et « l'intérieur », qui font penser à la protection du pays et à une politique plus stricte envers la migration. Ici, on peut donc voir deux tendances idéologiques opposées : primo, il y a l'acte héroïque de Mamoudou Gassama, et secundo, il y des débats sur la situation en Méditerrannée et la coopération entre les ministres de l'Intérieur français et espagnol quant à la défense des frontières. Le réseau semble donc plus dichotomique que dans le cas des collocatifs du mot « migration » .

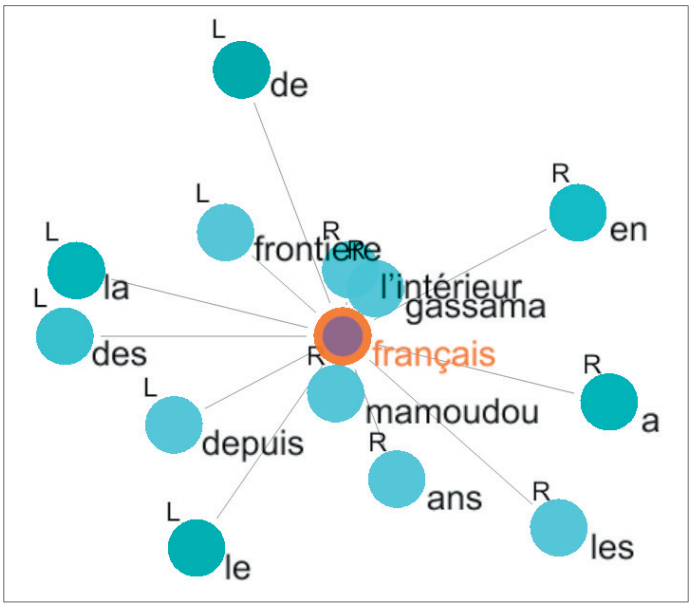

Figure 2 : Le réseau collocatif du mot de tête « français / Français » (métrique de dés).

6. Conclusions

En somme, les résultats de l'étude sont présentés dans les points suivants.

1° En ce qui concerne la comparaison des métriques, il semble que la MI et les dés trouvent les même collocatifs, les réseaux des dés étant plus riches ; toutefois, ceci est causé par de différents niveaux de significance choisis (voir ci-dessus). L'absence des collocatifs dans le score T induit qu'il n'y a pas assez de colligations dans les textes. Il est possible que le discours journalistique ait tendance à varier par défaut, et pas seulement dans le contexte étudié.

2° Les résultats mettent en lumière la focalisation étroite des articles ; la prédominance est gagnée par les organisations qui s'occupent des immigrants, et par des faits divers à valeur négative (le naufrage). Ce souci de détail est dû à la période assez courte qui est couverte par la présente recherche, à la tendance générale du journalisme moderne, et au fait que les articles plus interprétatifs peuvent avoir été exclus de la recherche, étant réservés aux abonnés. La négativité est le principe globalement répandu dans les journaux d'aujourd'hui.

3° Dans le cas du mot de tête « français / Français », la métrique de dés a montré des collocatifs à orientations idéologiques opposées (le personnage de Mamoudou Gassama contre « frontière » et « intérieur »), ce qui évoque une hétéroglossie présente dans les articles. Ceci ne paraît pas être le cas du mot « migration ».

4° Par rapport aux collocations, le mot de tête « migrant » prend le nombre minimal de collocatifs ; cela s'explique par le caractère vague de l'expression, son fonctionnement double (comme adjectif et substantif), et la multitude de contextes dans lesquels elle se trouve.

5° Quant aux interprétations, il faut toujours distinguer le trait stylistique du trait thématique. Bien qu'il soit peu probable que le style d'un auteur importe dans les textes, il se peut qu'il y ait des phénomènes qui sont typiques du journalisme en tant que sphère stylistique, de même que ceux qui partent d'un thème particulier. Pour que l'on puisse distinguer ces deux motivations, il faut procéder à des recherches plus étendues et suivant une direction différente.

Bibliographie

BREZINA, Vaclav ; Tony McENERY ; WATTAM, Stephen (2015). "Collocations in Context: A New Perspective on Collocation Networks". *International Journal of Corpus Linguistics*, 20.2, pp. 139-73.

BRŇÁKOVÁ, Jana (2013). "« L'Unité onomatologique complexe » de Jan Šabršula". *Studia Romanistica*, 13.1, pp. 19-25.

COSERIU, Eugenio (1967). "Lexikalische Solidaritäten". *Poetica*, 1, pp. 293-203.

CVRČEK, Václav (2015). "Asociační (kolokační) míry". [online] Disponible sur : https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry [cit. 2019-01-14].

FIRTH, John Rupert (1957). "Modes of Meaning". In: FIRTH, John Rupert. *Papers in Linguistics.* London : Oxford University Press, pp. 190-215.

FRANK, Cyrille. "Le "story-telling" contre l'information". [online] Disponible sur : http://www.mediaculture.fr/le-story-telling-contre-linformation/ [cit. 2019-01-14].

Immigration en Europe. [online] Disponible sur : https://www.lemonde.fr/immigration-en-europe/ [cit. 2019-01-14].

IOM – OMI. [online] Disponible sur : https://www.iom.int/fr [cit. 2019-01-14].

KŘEN, Michal (2006). "Kolokační míry a čeština: srovnání na datech Českého národního korpusu". In : ČERMÁK, František ; ŠULC, Michal (éds.). *Kolokace*. Praha : NLN, pp. 223-248.

MANNING, Christopher D. ; SCHÜTZE, Hinrich (2000). *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts : MIT Press.

"Paris : il escalade un immeuble pour sauver un enfant suspendu dans le vide". [online] Disponible sur : http://www.leparisien.fr/paris-75/paris-il-escalade-un-immeuble-pour-sauver-in-extremis-un-enfant-suspendu-dans-le-vide-27-05-2018-7738266.php [cit. 2019-01-14].

PINKER, Steven. "The media exaggerates negative news. This distortion has consequences". [online] Disponible sur : https://www.theguardian.com/commentisfree/2018/feb/17/steven-pinker-media-negative-news [cit. 2019-01-14].

SOS Méditérannée. [online] Disponible sur : https://sosmediterranee.org/ [cit. 2019-01-14].
ŠABRŠULA, Jan (1983). *Základy francouzské lexikologie.* Praha : SPN.

Michal Místecký
Katedra českého jazyka
Filozofická fakulta
Ostravská univerzita
Reální 5
701 03 OSTRAVA
République tchèque