Usefulness of corpus linguistic methods in the analysis of phrasemes in journalistic texts

[Utilité des méthodes de linguistiques de corpus dans l'analyse des phrasèmes dans les textes journalistiques]

Iveta Dinzikova

DOI: 10.18355/XL.2018.11.01XL.29

Abstract
The aim of this study is to demonstrate the usefulness of corpus linguistic methods in the analysis of phrasemes in journalistic texts. We analyze the phrasemes containing an ethnonym in the French and/ or Slovak languages in four monolingual corpora of journalistic texts (each of which comprises approximately 55 million to 600 million words) compared to four corpora of contrast (each of which comprises approximately nearly 130 million to 10 billion words). We search and observe these phrasemes using query language and corpus linguistics methods (CQL/ CQP language, relative frequency, and MI-Score test) to demonstrate statistically significant recurrences of these phrasemes and their specificity rate.
Key words: phrasemes, methods of corpus linguistics, ethnonyms, journalistic texts

Résumé
Cette étude vise à démontrer l'utilité des méthodes de linguistiques de corpus dans l'analyse de phrasèmes dans les textes journalistiques. Nous analysons les phrasèmes contenant un ethnonyme en langues française et/ ou slovaque dans quatre corpus monolingues de textes journalistiques (dont chacun comporte à peu près de 55 millions à 600 millions de mots) par rapport à quatre corpus de contraste (dont chacun comporte à peu près de 130 millions à 10 milliards de mots). Nous recherchons et observons ces phrasèmes par le langage de requête et des méthodes de linguistique de corpus (langage CQL/ CQP, fréquence relative et test MI-Score) pour démontrer les récurrences statistiquement significatives de ces phrasèmes et leur taux de spécificité.
Mots-clés : phrasèmes, méthodes de linguistique de corpus, ethnonymes, textes journalistiques

1. Introduction

Les méthodes de linguistique de corpus qui se sont développées pendant les dernières décennies, supportent une observation empirique ainsi qu'une utilité pour d'autres disciplines linguistiques comme la phraséologie. En appliquant quelques-unes de ses méthodes (fréquence relative et test MI-Score), nous analysons les phrasèmes contenant un ethnonyme en langues française et/ ou slovaque dans les textes journalistiques. Nous démontrons ainsi les récurrences significatives de ces phrasèmes dans les textes journalistiques par rapport à un corpus de contraste (corpus de divers types de textes[1]). De plus, nous pouvons aussi classer les phrasèmes composés de deux mots par rang selon le taux de spécificité dans les textes journalistiques.

[1] Selon http://korpus.sk/bibstyle.html : au niveau stylistique, il y a quatre types de textes principaux dans cette recherche : journalistiques, littéraires, spécialisés et la *live communication*. Les corpus *frTenTen12*, *skTenTen11* et *prim-7.0-public-all* sont

Les phrasèmes sont tout d'abord sélectionnés dans plusieurs dictionnaires (monolingues, bilingues, phraséologiques ou étymologies) selon un mot-clé (ethnonyme). Ensuite, ces phrasèmes sont analysés par les méthodes présentées dans quatre corpus monolingues de textes journalistiques (deux français et deux autres slovaques dont chacun comporte à peu près de 55 millions à 600 millions de mots) par rapport à quatre corpus de contraste (dont chacun comporte à peu près de 130 millions à 10 milliards de mots).
Le but de cet article est donc de démontrer l'utilité des méthodes de linguistiques de corpus dans l'analyse des phrasèmes dans les textes journalistiques. ***L'objectif de cette recherche*** est de présenter les phrasèmes avec un taux de représentation élevé dans les textes journalistiques par rapport à un corpus de contraste.
Ce travail relève du domaine des Humanités numériques en Sciences humaines et sociales par son objet de recherche (les phrasèmes des textes journalistiques) et sa méthodologie (celle de la linguistique de corpus).

2. Méthodologie appliquée
À cause de l'hétérogénéité des données, du déséquilibre des sources et d'une représentativité insuffisante, nous utilisons des méthodes, des approches et des sources hybrides (dictionnaires, études, quatre corpus observés – *frTenTen12, skTenTen11, prim-7.0-public-all, Emolex* avec leurs quatre sous-corpus de textes journalistiques). Tout d'abord nous analysons ainsi la description scientifique, linguistique et lexicographique actuelle des phénomènes évoqués, par exemple dans les études et dans les dictionnaires. Ensuite, nous observons le fonctionnement réel des phénomènes langagiers dans les corpus. Il s'agit de **quatre corpus monolingues** dont chacun comporte à peu près de 130 millions à 10 milliards de mots (les corpus *frTenTen12, skTenTen11, Emolex* et le Corpus National Slovaque *prim-7.0-public-all*) et de **leurs quatre sous-corpus monolingues de textes journalistiques** dont chacun comporte à peu près de 55 millions à 600 millions de mots (le corpus *frTenTen12Presse,* le corpus *skTenTen11Presse,* le corpus *EmolexPresse* et le corpus *prim-7.0-public-inf*).
Nous portons surtout notre attention sur la caractéristique et la démonstration de l'utilité des méthodes de linguistique de corpus, dans une démarche déductive *corpus-based* (Hunston, 2002, 2006 ; Biber, 2010), dans l'analyse des phrasèmes dans les textes journalistiques. Plus précisément, nous utilisons le langage de requête CQL/ CQP et deux méthodes quantitatives de linguistique de corpus, c'est-à-dire la fréquence relative et le test MI-Score, pour tester les trois hypothèses que nous avons formulées.
Avant de faire cette recherche (au total 70 phrasèmes), nous avons analysé 23 phrasèmes dans les corpus frTenTen12, skTenTen11, Emolex, prim-7.0-public-all par deux méthodes de linguistique de corpus (fréquence relative et le test logDice). S'appuyant sur les premières observations (Dinžíková, 2017), nous formulons les hypothèses suivantes :

- Hypothèse 1 : nous supposons que le langage journalistique se caractérise par la surreprésentation statistiquement significative de certains phrasèmes analysés qui jouent aussi un certain rôle dans les textes journalistiques.
- Hypothèse 2 : nous supposons que nous sommes capables de justifier empiriquement les phrasèmes statistiquement significatifs dans les textes journalistiques liés à la culture française ainsi que ceux liés à la culture slovaque en même que ceux étant identiques dans les deux cultures.

---

composés de ces quatre types de textes principaux. Cependant dans le corpus *Emolex,* il y a exclusivement des textes journalistiques et littéraires.

– Hypothèse 3 : nous supposons que la plupart des phrasèmes observés représente la preuve que deux mots sont des collocatifs dans les textes journalistiques car nous utilisons le test MI-Score qui est sensible aux phrasèmes.

2.1. Corpus observés

2.1.1. Corpus frTenTen12 et son sous-corpus frTenTen12Presse

Corpus frTenTen12[2] est un corpus monolingue textuel en langue française qui a été créé à partir de textes téléchargés d'Internet en 2012. Sa taille est à peu de 11,5 milliards de tokens (c'est-à-dire à peu près de 10 milliards de mots). Il est composé surtout de quatre types de textes : la *live communication*, les textes journalistiques, littéraires et spécialisés, mais nous ne connaissons pas la proportionnalité de ces textes car ils n'ont pas d'annotation stylistique.

Sous-corpus frTenTen12Presse est un corpus monolingue textuel en langue française qui vient du corpus *frTenTen12*. Il est composé à peu près de 115 millions de tokens (ou bien à peu près de 99 millions de mots). Il s'agit exclusivement de textes journalistiques issus de la presse française : *Le Parisien*, *L'Équipe*, *Le Monde*, *Le Figaro*, *Libération* et *Les Échos* (voir Tableau 1).

Journal	Requête en langage CQL/ CQP
Le Parisien	<doc (urldomain=".*parisien.fr") />
L'Équipe	<doc (urldomain=".*equipe.fr") />
Le Monde	<doc (urldomain=".*lemonde.fr") />
Le Figaro	<doc (urldomain=".*lefigaro.fr") />
Libération	<doc (urldomain=".*liberation.fr") />
Les Échos	<doc (urldomain=".*echos.fr") />

Tableau 1 : Liste de journaux créant le sous-corpus frTenTen12Presse

Nous avons créé ce sous-corpus par l'intermédiaire de la fonction *Faire des sous-corpus* où nous avons mis les sites des journaux français en langage CQL/ CQP.

2.1.2. Corpus skTenTen11 et son sous-corpus skTenTenPresse

Corpus skTenTen11[2] est un corpus monolingue textuel en langue slovaque d'à peu près 656 millions de tokens (c'est-à-dire à peu près 540 millions de mots), à partir de textes téléchargés d'Internet en 2011. Concernant les types de textes et l'annotation stylistique, ce sont les mêmes caractéristiques que dans le corpus frTenTen12.

Sous-corpus skTenTen11Presse est un corpus monolingue textuel en langue slovaque qui a été créé à partir du corpus *skTenTen11*. Sa taille est à peu près de 67 millions de tokens (ou bien à peu près 55 millions de mots).

Journal	Requête en langage CQL/ CQP
Sme	<doc (urldomain=".*sme.sk") />
Hospodárske noviny	<doc (urldomain=".*hnonline.sk") />
Nový čas	<doc (urldomain=".*cas.sk") />
Pravda	<doc (urldomain=".*pravda.sk") />
Bratislavské noviny	<doc (urldomain=".*bratislavskenoviny.sk") />
Plus jeden deň	<doc (urldomain=".*pluska.sk") />

[2] Pour trouver ce corpus, voir https://the.sketchengine.co.uk/

Tableau 2 : Liste de journaux créant le sous-corpus skTenTen11Presse

Ce sous-corpus est créé aussi exclusivement de textes journalistiques, issus de la presse slovaque : *Sme, Hospodárske noviny, Nový čas, Pravda, Bratislavské noviny* et *Plus jeden deň* (voir Tableau 2). Pour constituer ce sous-corpus, nous avons également utilisé la même fonction que pour le sous-corpus frTenTen12Presse.

2.1.3. Corpus National Slovaque prim-7.0-public-all et son sous-corpus prim-7.0-public-inf

Corpus prim-7.0-public-all[3] est un corpus monolingue textuel en langue slovaque utilisant des textes téléchargés d'Internet ainsi que des textes numérisés issus de la presse périodique, d'œuvres littéraires et de littérature spécialisée. Il fait à peu près 1,2 milliards de tokens (c'est-à-dire à peu près 1 milliard de mots). Il contient quatre types de textes : journalistiques (65,1 %), littéraires (15,1 %), spécialisés (9,5 %) et d'autres textes (10,3 %), ils ont donc tous une annotation stylistique.

Sous-corpus *prim-7.0-public-inf*[3] est un corpus monolingue textuel en langue slovaque qui vient du corpus *prim-7.0-public-all*. Sa taille est à peu près de 771 millions de tokens (ou bien à peu près de 597 millions de mots). Il est composé de textes journalistiques, mais nous ne connaissons pas la liste précise des journaux et des magazines.

2.1.4. Corpus Emolex et son sous-corpus EmolexPresse

Corpus Emolex[4] est un corpus monolingue textuel en langue française d'à peu près137 millions de tokens (c'est-à-dire à peu près 128 millions de mots), qui a été créé à partir de textes numérisés issus de la presse périodique et d'œuvres littéraires. Il s'agit de deux types de textes : les textes journalistiques (88,3 %) et littéraires (11,7 %) et ils ont aussi tous une annotation stylistique.

Sous-corpus EmolexPresse[4] est un corpus monolingue textuel en langue française venant du corpus *Emolex*. Il contient est à peu près 121 millions de tokens (ou bien à peu près 112 millions de mots). Il est composé de textes journalistiques, mais nous ne connaissons pas la liste précise des journaux et des magazines.

2.1.5. Comparaison de quatre corpus et de quatre sous-corpus

Chacun des huit corpus et sous-corpus ci-dessus propose d'autres **avantages quantitatifs (**taille importante des corpus : *les corpus frTenTen12, skTenTen11, prim-7.0-public-all,* diversité stylistique : *les corpus frTenTen12, skTenTen11, prim-7.0-public-all***) et/ ou qualitatifs (**homogénéité stylistique des textes : *tous les sous-corpus*, proportionnalité précise des textes : *les corpus Emolex et prim-7.0-public-all avec leur sous-corpus*). À cause de cette hétérogénéité des corpus et des sous-corpus, nous utilisons différentes méthodes quantitatives de linguistiques de corpus.

2.2. Langage de requête CQL/ CQP

Le langage *CQL* (Corpus Query Language) ou bien *CQP* (Corpus Query Processor) est un langage formel de requête utilisé dans les corpus textuel. Il s'agit d'une chaîne de caractères exprimant un motif linguistique (un mot ou une suite de mots) qui est caractérisé par un étiquetage (comme par exemple le lemme, le tag, la catégorie grammaticale) avec la combinaison des expressions régulières (une chaîne de caractères alphanumériques et de symboles spéciaux, créée sur la base des règles syntaxiques concrètes, qui décrit un certain ensemble de formes). Ce langage « articule trois niveaux dotés chacun d'opérateurs (joker, répétition, OU, etc.) : le niveau des occurrences (combinaison des mots), le niveau des propriétés (mobilisation possible de diverses étiquettes attachées aux occurrences), et le niveau des caractères (variations d'expression d'un mot ou plus généralement d'une valeur d'étiquette) »

[3] Pour trouver ce corpus, voir http://korpus.juls.savba.sk/

[4] Pour trouver ce corpus, voir http://emolex.u-grenoble3.fr/emoBase/

(Pincemin et col., 2008 : 93). C'est donc le niveau des occurrences ou bien des cooccurrences des mots qui est important pour l'analyse de nos phrasèmes.
Nous pouvons trouver dans la littérature spécialisée les deux termes pour ce langage formel de requête. La notion CQL est utilisée par exemple par Jakubíček, Kilgarriff, McCarthy et Rychlý (2010), Pecman (2012) et aussi sur les sites : https://www.sketchengine.co.uk/documentation/corpus-querying/ et http://korpus.juls.savba.sk/. D'autres linguistes (Pincemin et col., 2008 ; Schaeffer-Lacroix, 2015 ; Everest et Hardie, 2011) utilisent la notion CQP. De plus, Everest et Hardie (2011 : 8) démontrent les différences entre ces deux termes et ils donnent des arguments pour utiliser celui de CQP.

Grâce au langage de requête *CQL/ CQP***,** nous recherchons tous les phrasèmes dans quatre corpus et quatre sous-corpus observés (voir Tableau 3).

Phrasème	Requête en langage CQL/ CQP
réponse de Normand	[lemma="(?i)reponse" | lemma="(?i)réponse"] [lemma="(?i)de"] [lemma="(?i)Normand"]
été indien	[lemma="(?i)ete" & tag="N.*"] | [lemma="(?i)été" & tag="N.*"] [lemma="(?i)indien"]

Tableau 3 : Exemple de la requête des phrasèmes dans les corpus par le langage CQP/ CQL

2.3. Fréquences absolue et relative

La fréquence est considérée comme un des indicateurs de base pour démontrer les récurrences significatives du phénomène analysé. D'un côté, nous distinguons la *fréquence absolue* et la *fréquence relative*, d'un autre, la *fréquence d'occurrences* et la *fréquence de cooccurrences*. La fréquence absolue représente un nombre d'occurrences ou de cooccurrences (ces termes ont été distingués par Gries, 2010) du phénomène dans tout le corpus observé. Dans cette étude, nous utilisons seulement la fréquence des cooccurrences qui selon Gries (2010 : 269) désigne « la fréquence des éléments linguistiques tels que les morphèmes, les mots, les motifs/ constructions qui se cooccurrent avec un autre élément linguistique de cet ensemble ou une position dans le texte » car nous n'analysons que des phrasèmes. Par exemple, elle désigne combien de fois le phrasème *roulette russe* se trouve dans les corpus frTenTen12 et Emolex (voir Tableau 4, colonne *Fréquence absolue* où le nombre de cooccurrences de *roulette russe* est 4 499 fois pour le corpus frTenTen12 et 68 fois pour le corpus Emolex). Cependant, tous les corpus observés sont de taille différente, donc ils ne sont pas comparables par la fréquence absolue et pour cette raison, nous utilisons la fréquence relative (voir Tableau 4, colonne *Fréquence relative)*.

Phrasème	Fréquence absolue		Fréquence relative	
	Corpus frTenTen12	Corpus Emolex	Corpus frTenTen12	Corpus Emolex
roulette russe	4 499	68	0,4	0,59

Tableau 4 : Exemple des fréquences absolue et relative

La *fréquence relative* représente la *fréquence absolue* divisée par la taille totale d'un corpus (en tokens), donc il s'agit d'une fréquence du phénomène linguistique sur un million d'occurrences. Nous sommes capables de la compter selon la formule suivante :

$$REL = \frac{ABS}{N} \times 1000000$$

Grâce à la fréquence relative, il est donc possible de comparer les résultats de la fréquence relative dans le cadre d'un ou de plusieurs corpus car il s'agit toujours de la fréquence du phrasème sur un million. Plus la valeur de fréquence relative est élevée, plus le rang l'est aussi. Ainsi nous pouvons empiriquement justifier les récurrences statistiquement significatives des phrasèmes analysés dans quatre sous-corpus observés.

2.4. Test MI-Score

Il existe actuellement plusieurs calculs statistiques dans la linguistique de corpus, par exemple les tests *MI-Score*, *logDice*, *MI3*, *log-likelihood* ou *T-score*. Ces tests présentent les valeurs de spécificité statistique de chaque collocation (dans notre cas de chaque phrasème) et les classent par rang selon le taux de spécificité ou bien le taux d'association seulement entre deux mots, donc les collocations de plus de deux mots sont exclues. Plus leur valeur est élevée, plus leur rang l'est aussi. De plus, il est possible de comparer leurs résultats seulement dans le cadre d'un corpus car les variables de chaque corpus sont spécifiques.

Chaque test est sensible pour un autre type de collocation. Nous choisissons le test MI-Score car les phrasèmes sont notre objet de recherche et ce test est sensible aux termes techniques ou à d'autres expressions qui présentent très peu ou pas de variation (Gries, 2010).

Le test *MI-Score* (Mutual information score) dépend de la fréquence de composantes *x* et *y* et de la fréquence de bigramme *xy*) ainsi que de la taille de corpus en tokens (voir la formule ci-dessous).

$$MI(xy) = \log_2 \frac{\frac{f(xy)}{N}}{\frac{f(x)}{N}\frac{f(y)}{N}} = \log_2 \frac{N f(xy)}{f(x) f(y)}$$

Plusieurs linguistes ont interprété les valeurs du test MI-Score, par exemple McEnery et col. (2005 : 56) proposent que « si la valeur de *MI-Score* est proche de zéro, très probablement, cela signifie que deux mots sont cooccurrents par chance. La valeur de *MI-Score* peut être aussi négative si deux mots ont tendance à se repousser ». Hunston (2002 : 71) souligne que « si la valeur du phrasème de *MI-Score* est de 3 ou plus, cela peut être considéré comme une preuve que deux mots sont des collocatifs ». Grâce à cela, nous pouvons donc interpréter les valeurs du test MI-Score de nos phrasèmes.

3. Analyse des phrasèmes dans les textes journalistiques

3.1. Définition de la notion de phrasèmes

Comme il n'existe pas d'homogénéité ni dans la dénomination, ni dans la caractéristique de l'unité fondamentale de phraséologie dans la linguistique française, nous adopterons la dénomination et la caractéristique de cette unité de Rosenbaum Franková (2010 : 27) car elle analyse les définitions des phraséologues français ou bien francophones en prenant en considération des facteurs de la phraséologie slovaque actuelle. Rosenbaum Franková s'est appuyée sur les définitions de plusieurs linguistes francophones, comme Gross (1996), Gréciano (1997b), González-Rey (2002), Mejri (2000), Náray-Szabó (2002), Gautier (2004), Mešková (2004) où elle distingue les *phrasèmes généraux* et *terminologues*.

Elle caractérise le phrasème général par des propriétés distinctives (polylexicalité, figement et opacité sémantique) et par des propriétés complémentaires (figuration, iconicité, reproductibilité, anomalies, expressivité et intraduisibilité). Un exemple de phrasème général est *soûl comme un Polonais*.

Le phrasème terminologique (Rosebaum Franková, 2010 : 59) est une « unité phraséologique se retrouvant régulièrement dans les textes spécialisés qui a des propriétés inhérentes aux phrasèmes, comme la figuration et l'expressivité, et qui a également une fonction dénominative du terme ». Un exemple de phrasème terminologique est *clé anglaise.* Quant à certains phrasèmes terminologiques, il semble qu'il s'agisse seulement de termes même si la frontière n'est pas claire. Cependant, les phrasèmes terminologiques ont au minimum une composante figurée ou métaphorique.

3.2. Sélection des phrasèmes dans les dictionnaires et dans les études

Nous sélectionnons les phrasèmes comprenant un ethnonyme dans les dictionnaires monolingues, bilingues, phraséologiques ou étymologies (Rey et col., 2013 ; Rey et col., 2003 ; Ashraf et Miannay, 1995 ; Dojerová-Danthine, 2006 ; Gründlerová et col., 1991 ; Gründlerová, Škultéty et Taraba, 1992a ; 1992b ; Klein, 2008 ; Tillier et col., 2014 ; Planelles, 2016) et dans les études (Puchovská, 2013 ; Gréciano, 1997a ; Pamies, 2011). Nous trouvons ainsi 70 phrasèmes (40 phrasèmes en langue française et 30 phrasèmes en langue slovaque, cependant il y en a 15 identiques pour ces deux langues, mais nous les comptons séparément pour chaque langue).

3.3. Requête des phrasèmes dans huit corpus et sous-corpus

Nous recherchons tous les phrasèmes sélectionnés dans huit corpus et sous-corpus par le langage de requête CQL/ CQP. Nous présentons tous les phrasèmes et quelques-uns sont accompagnés d'un exemple tiré d'un des sous-corpus ainsi que d'une annotation bibliographique, par exemple le lieu et la date de la publication. Ces phrasèmes sont donc issus de textes journalistiques français et slovaques puisque la presse représente le domaine ciblé pour cette recherche. Nous indiquons la traduction littérale des phrasèmes entre crochets, pour bien mettre en évidence les différences culturelles entre les langues française et slovaque. De plus, tous les phrasèmes sont aussi suivis de leur explication.

3.3.1. Phrasèmes identiques avec le même ethnonyme dans les langues française et slovaque

Nous trouvons 15 phrasèmes identiques dans ces deux langues, nous pouvons les désigner en tant que *phrasèmes entièrement équivalents* (selon la typologie d'équivalence des phrasèmes, Mešková et Kubeková, 2015 ; Ţaran Andreici, 2016), à savoir :

- ***roulette russe – ruská ruleta*** [roulette russe] désigne soit une décision importante, mais prise à la légère, qui est risquée et dangereuse, soit le jeu suicidaire avec un revolver et une cartouche.
- ***comme une horloge suisse – ako švajčiarske hodinky*** [comme une horloge suisse/ une montre suisse] désigne quelqu'un ou quelque chose de vraiment précis, ponctuel, régulier.
  Ex. : *„Ľudský organizmus nefunguje **ako švajčiarske hodinky**, ani náš krvný tlak nie je rovnaký v každej situácii a čase," upozorňuje Lipták.*

  skTenTen11Presse : Web site : sme.sk, doc.url : http://zena.sme.sk/c/4281372/vysoky-tlak-neboli-ale-zabija.html, doc.timestamp : 20101030140333
- ***douche écossaise – škótska sprcha*** [douche écossaise] désigne soit une suite d'événements inattendus entraînant une déconvenue (mais en slovaque la traduction de ce phrasème est sans ethnonyme), soit le changement brutal de température de l'eau sous la douche.
- ***froid sibérien – sibírska zima*** [froid sibérien] désigne un froid vraiment rigoureux, glacial.

Ex. : *En faisant tomber le gouvernement, les jeunes manifestants de Bucarest ont pris conscience de leur force. Ils poursuivent la lutte, dans un* ***froid sibérien****. Elle marche dans la neige comme sur un tapis rouge.*

frTenTen12Presse : Web site : lemonde.fr, Crawl date : 2012-03-04, doc.url : http://www.lemonde.fr/sujet/5c45/mihai-razvan.html

- ***grippe espagnole – španielska chrípka*** [grippe espagnole] désigne la grippe répandue entre 1919 et 1923 dans le monde entier.
- ***guitare espagnole – španielska gitara*** [guitare espagnole] désigne la guitare traditionnelle.
- ***bain turc – turecký kúpeľ*** [bain turc] désigne un bain de vapeur.
  Ex. : *Vývojník pary môže mať zásobník či priehlbinu na pridanie esencií pre aromaterapiu. Teplota v* ***tureckom kúpeli*** *sa pohybuje len okolo 40 – 50 oC s relatívnou vlhkosťou vzduchu až 98 %.*

  prim-7.0-public-inf : doc.bibl : Hospodárske noviny. Bratislava: Ecopress a.s. 2005, roč. 14, 05.01.2005., doc.id : 2008-06-17-n-67041
- ***massage thaïlandais – thajská masáž*** [massage thaïlandais] désigne la thérapie ancienne et traditionnelle issue de Thaïlande.
- ***gazon anglais – anglický trávnik*** [gazon anglais] désigne le gazon bien entretenu, dense et court, d'une couleur très verte.
- ***petit-déjeuner anglais – anglické raňajky*** [petit-déjeuner anglais] désigne un petit-déjeuner se différenciant du petit-déjeuner continental, où il y a des œufs et de la charcuterie.
- ***chiffre arabe – arabská číslica*** [chiffre arabe] désigne les chiffres utilisés aujourd'hui dans la plupart des cultures, inventés par les Arabes, qui comportent des chiffres de zéro à neuf.
  Ex. : *Skomplikovalo nám to situáciu, pretože časť dokumentácie sme už mali vypísanú* ***arabskými číslami****.*
  prim-7.0-public-inf : doc.bibl MY Hornonitrianske noviny. Bratislava: Petit Press 2009, roč. 16, č. 38., doc.id : 2012-11-29-a-5863
- ***chiffre romain – rímska číslica*** [chiffre romain] désigne des chiffres inventés par les Romains, qui comportent les caractères *I*, *V*, *X*, *L*, *C*, *D* et *M*.
- ***couteau suisse – švajčiarsky nožík*** [couteau suisse] désignant le couteau de l'armée suisse utilisé dans le monde entier, pour sa versatilité.
  Ex. : *Les marchands de journaux en vendent, de même que des enveloppes renforcées pour réexpédier chez soi* ***un couteau suisse*** *ou une lime à ongles, interdits en cabine depuis 2001.*

  EmolexPresse : publisher => LibÃ©ration, pubDate => 04/01/2007
- ***médecine chinoise – čínska medicína*** [médecine chinoise] désigne la médecine non conventionnelle liée à l'acupuncture, à la phytothérapie et à des massages.
- ***signe chinois – čínske znamenie*** [signe chinois] désigne 12 signes astrologiques inventés par les Chinois, comme par exemple le Cochon, le Rat ou le Dragon.

3.3.2. Phrasèmes dans la langue française

Nous trouvons les 25 phrasèmes suivants dans quatre corpus et sous-corpus français :

- ***saoul/ soûl comme un Polonais*** désigne quelqu'un étant complètement ivre.

- ***filer à l'anglaise*** désigne l'acte de quitter une soirée sans saluer les convives.
- ***d'un calme olympien*** désigne une personne imperturbable, dotée d'un grand calme.
- ***clé anglaise*** désigne un outil à main avec deux mâchoires mobiles qui servent à adapter des écrous et des boulons de tailles différentes.
- ***canne anglaise*** désigne une variété de béquille utilisée pour faciliter la station debout ou la marche en cas d'immobilisation.
- ***fort comme un Turc*** désigne quelqu'un très étant fort, plein de force.
- ***tête de Turc*** désigne une personne étant l'objet de moqueries et de méchancetés.
- ***quart d'heure américain*** désigne le moment où ce sont les femmes qui invitent les hommes à danser.
- ***promesse de Gascon*** désigne une fausse promesse.
  Ex. : *Depuis, celui- ci vient d'obtenir du gouvernement irlandais la promesse qu'il userait de son droit de veto, en cas d'accord à l'OMC sur la libéralisation des produits agricoles.* ***Promesse de Gascon****? Un Etat membre de l'UE ne dispose pas du droit de veto dans le domaine de l'agriculture.*

  EmolexPresse : publisher => Le Monde, pubDate => 06/06/2008
- ***aller se faire voir chez les Grecs*** désigne se faire renvoyer, éconduire brutalement.
- ***téléphone arabe*** désigne la diffusion orale et rapide de nouvelles par le bouche à oreille.
- ***été indien*** désigne une période de beau temps en automne.
  Ex. : *Ce serait la fin de la rupture entre médecine occidentale et médecine traditionnelle? S'il était encore possible de se promener en chemisette mi-octobre et de profiter du fameux* ***été indien*** *– il faisait entre 20° et 25° mercredi dernier – il faut dorénavant s'habiller plus chaudement.*

  frTenTen12Presse : Web site : lemonde.fr, Crawl date : 2012-03-01, doc.url http://abroadatau.blog.lemonde.fr/2005/10/27/2005_10_midterm_exams/
- ***assiette anglaise*** désigne un plat froid où il y de la charcuterie, des œufs, des légumes et des condiments.
- ***querelle d'Allemand*** désigne une querelle sans sujet.
- ***querelle byzantine*** désigne une discussion subtile et presque inutile.
- ***travail de Romain*** désigne une tâche difficile ou un travail énorme.
  Ex. : *Faire décoller une entreprise en création, c'est un* ***travail de Romain****. Le plus souvent, et c'est logique, l'entrepreneur et sa jeune équipe ne peuvent ni tout faire ni tout voir.*

  frTenTen12Presse : Web site : lefigaro.fr, Crawl date : 2012-03-03, doc.url : http://blog.lefigaro.fr/legales/2010/06/post.html
- ***file indienne*** désigne une suite de personnes ou de véhicules qui avancent les uns derrière les autres.
- ***signe indien*** désigne un mauvais sort, une malédiction.
  Ex. : *Au retour des vestiaires, la TA Rennes semble bien décidée à vaincre* ***le signe indien*** *(aucune victoire à domicile depuis le début de saison), même si elle pêche dans la finition, à l'image de la tentative non cadrée d'Aneftah.*

EmolexPresse : publisher => Ouest-France, pubDate => 8/12/2008

- ***calendes grecques*** désigne une date inexistante.
- ***casse-tête chinois*** désigne un casse-tête particulièrement difficile à comprendre.
- ***poupée russe*** désigne le jeu en forme de figure de poupées, de taille décroissante, qui se placent les unes à l'intérieur des autres.
- ***supplice chinois*** désigne quelqu'un ou quelque chose très cruel.
- ***réponse de Normand*** désigne une réponse évasive qui ne dit ni oui, ni non.
  Ex. : *Je ferai une réponse de Normand : tout dépend de votre job, de votre entreprise et de votre personnalité!*

EmolexPresse : publisher => Le Figaro, pubDate => 04/06/2007

- ***auberge espagnole*** désigne une auberge où nous ne trouvons que ce que nous y avons apporté.
- ***trou normand*** désigne une coutume gastronomique française consistant à boire un petit verre d'alcool français (calvados) au milieu d'un repas.

3.3.3. Phrasèmes dans la langue slovaque

Nous trouvons les 10 phrasèmes suivants dans quatre corpus et sous-corpus slovaques :

- ***španielska dedina*** [c'est un village espagnol] désigne quelque chose d'incompréhensible.
- ***s kľudom Angličana*** [d'un calme d'un Anglais] désigne, comme le phrasème *d'un calme olympien*, une personne imperturbable, dotée d'un grand calme.
- ***francúzsky kľúč*** [clé française] désigne, comme le phrasème *clé anglaise*, un outil à main avec deux mâchoires mobiles qui servent à adapter des écrous et des boulons de tailles différentes.
  Ex. : *O tejto profesii prevláda názor, že je to povolanie spojené so zaolejovanými rukami, s kladivom,* ***francúzskym kľúčom****. Pozrite si však dnešné špičkové autoservisy.*
  prim-7.0-public-inf : doc.bibl : Hospodárske noviny. Bratislava: Ecopress a.s. 2004, roč. 13, 30.03.2004, doc.id : 2008-06-17-n-47116
- ***francúzska barla*** [canne française] désigne, comme le phrasème *canne anglaise*, une variété de béquille utilisée pour faciliter la station debout ou la marche en cas d'immobilisation.
- ***fajčiť ako Turek*** [fumer comme un Turc] désigne quelqu'un qui fume du tabac de façon excessive.
- ***raz za uhorský rok*** [une fois par an hongrois] désigne un événement très rare ou qui n'arrive jamais.
  Ex. : *Užite si letnú túru vo Vysokých Tatrách! Tú, ktorú vám prinášame teraz, zvládnu aj „bábovky", ktoré do vysokohorského prostredia zavítajú* ***raz za uhorský rok***.
  skTenTen11Presse : Web site : cas.sk, doc.url : http://vas.cas.sk/clanok/2542/najkrajsie-tury-vysokych-tatier-studenovodska-je-aj-pre-zaciatocnikov.html, doc.timestamp : 20101029192751
- ***zima ako v ruskom filme*** [froid comme dans un film russe] désigne un froid intense.
- ***ako Tatár*** [comme un Tartare] désigne quelqu'un de très maladroit.
- ***švédska trojka*** [triade suédoise] désigne un couple accueillant une troisième personne dans leur relation. amoureuse.
- ***ruské kolo*** [roue russe] désigne une attraction de très grande taille et en forme de roue.

- ***kanadský žart*** [blague canadienne] désigne une mauvaise blague, par exemple, mettre un seau d'eau en équilibre sur une porte entrouverte.
- ***mexická vlna*** [vague mexicaine] désigne le mouvement de foule, qui se manifeste dans des rassemblements importants.
  Ex. : *Žarnovický Autoslide je najmasovejším podujatím tohto druhu v rámci Slovenska. „Odhadujem, že je tu približne 9-tisíc divákov. Vytvárajú fantastickú atmosféru, dnes už robili aj* ***mexické vlny****," pokračoval Valent, podľa ktorého návštevnosť z roka na rok narastá.*
  skTenTen11Presse : Web site : sme.sk, doc.url : http://ziar.sme.sk/c/5188386/na-zarnovicky-autoslide-prisli-majstri-zo-styroch-statov.html, doc.timestamp : 20101029135009
- ***švédsky stôl*** [table suédoise] désigne un buffet : une table où il y a de la nourriture et des boissons disposées pour une réception.
- ***turecké hospodárstvo*** [économie turque] désigne une mauvaise gestion économique.
- ***turecký sed*** [position assise turque] désigne la position assise avec les jambes croisées.

3.4. Fréquence relative des phrasèmes

Nous analysons d'abord tous les phrasèmes par l'intermédiaire de la fréquence relative (voir Tableaux 5 et 6) pour justifier empiriquement les récurrences statistiquement significatives dans quatre sous-corpus de textes journalistiques. Nous comparons ensuite les valeurs de la fréquence relative des phrasèmes.

Dans un premier temps, cette comparaison se fait dans le cadre de deux sous-corpus dans la langue française (grâce au calcul de la moyenne de leurs valeurs de fréquence relative de tous les phrasèmes observés dans ces deux sous-corpus).

Dans un deuxième temps, elle repose sur le même calcul dans le cadre de deux sous-corpus dans la langue slovaque.

Dans un troisième temps, elle s'effectue dans quatre sous-corpus par rapport à leurs corpus de contraste (en comparant les valeurs de la fréquence relative de chaque phrasème dans chaque sous-corpus par rapport à des valeurs de la fréquence relative du même phrasème dans le cadre de son propre corpus). Par exemple, dans le Tableau 5, la récurrence de *roulette russe* dans le sous-corpus frTenTen12Presse (fréquence relative égale à 0,58), est statistiquement significative par rapport à celle dans le corpus frTenTen12 (fréquence relative égale à 0,4). Nous sommes intéressés par les phrasèmes qui ont une valeur de fréquence relative dans les textes journalistiques d'au minimum 20 % de plus que dans le corpus de contraste.

Phrasème	Fréquence relative			Fréquence relative		Phrasème
	Corpus frTenTen12	Emolex		Corpus skTenTen11	prim-7.0-public-all	
	Corpus frTenTen12 Presse	EmolexPresse		Corpus skTenTen11 Presse	prim-7.0-public-inf	
roulette russe	0,4	0,59	*	0,25	0,26	ruská ruleta
	0,58	0,55		0,36	0,3	
comme une horloge suisse	0,03	0,02	*	0,15	0,09	ako švajčiarske hodinky
	0,02	0,02		0,31	0,11	
douche écossaise	0,05	0,18	*	0,01	0,01	škótska sprcha
	0,06	0,16		0,01	0,01	
froid sibérien	0,1	0,04	*	0,02	0,07	sibírska zima
	0,06	0,04		0,06	0,08	
grippe espagnole	0,21	X	*	0,28	0,22	španielska chrípka
	0,4	X		0,39	0,25	
guitare espagnole	0,03	0,03	*	0,05	0,03	španielska gitara
	X	0,03		0,07	0,04	
bain turc	0,15	0,1	*	0,59	0,13	turecký kúpeľ
	0,01	0,07		0,21	0,1	
massage thaïlandais	0,15	X	*	0,58	0,26	thajská masáž
	0,03	X		0,52	0,2	
gazon anglais	0,03	0,11	*	0,19	0,16	anglický trávnik
	0,03	0,12		0,45	0,17	
petit-déjeuner anglais	0,03	X	*	0,12	0,08	anglické raňajky
	X	X		0,16	0,07	
chiffre arabe	0,18	0,02	*	0,24	0,12	arabská číslica
	0,24	X		0,13	0,05	
chiffre romain	0,27	0,06	*	0,31	0,29	rímska číslica
	0,12	0,02		0,34	0,2	
couteau suisse	0,47	0,4	*	0,06	0,1	švajčiarsky nožík
	0,28	0,33		0,18	0,04	
médecine chinoise	0,63	0,31	*	1,4	0,38	čínska medicína
	0,21	0,33		0,87	0,46	
signe chinois	0,1	X	*	0,04	0,01	čínske znamenie
	0,03	X		0,03	0,02	

Tableau 5 : Fréquences relatives des phrasèmes
X = phrasème sans occurrence dans le corpus ou dans le sous-corpus

** = phrasèmes identiques avec le même ethnonymes dans les langues française et slovaque*

= *phrasèmes statistiquement significatifs ayant la valeur de fréquence relative au minimum 0,17*

= *corpus de contraste pour les phrasèmes statistiquement significatifs dans les textes journalistiques*

Phrasème	Fréquence relative				Fréquence relative		Phrasème
	Corpus frTenTen12	Emolex			Corpus skTenTen11	prim-7.0-public-all	
	Corpus frTenTen12 Presse	EmolexPresse			Corpus skTenTen11 Presse	prim-7.0-public-inf	
soûl/ saoul comme un Polonais	0,01	0,01			0,03	0,05	turecké hospodárstvo/ hospodárenie
	0,03	0,01			0,07	0,07	
(d'un) calme olympien	0,21	0,11			0,35	0,35	mexická vlna
	0,08	0,11			0,63	0,36	
clé anglaise	0,05	0,09			0,18	X	švédsky stôl
	0,01	0,07			0,06	X	
canne anglaise	0,05	0,15			0,02	0,04	kanadský žart
	0,04	0,04			0,04	0,03	
filer à l'anglaise	0,08	0,24			0,02	0,04	fajčiť ako Turek
	0,06	0,23			0,09	0,02	
fort comme un Turc	0,01	X			0,25	0,1	raz za uhorský rok
	0,06	X			0,37	0,08	
tête de Turc	0,3	0,08			0,03	0,02	zima ako v ruskom filme
	0,9	0,09			0,01	0,02	
quart d'heure américain	0,02	X			0,53	0,26	španielska dedina
	0,01	X			1,05	0,25	
promesse de Gascon	0,01	0,01			0,08	0,03	ako Tatár
	0,07	0,01			0,07	0,04	
aller se faire voir chez les Grecs	0,01	X			0,2	0,16	turecký sed
	0,03	X			0,42	0,1	
téléphone arabe	0,1	0,04			0,06	0,04	švédska trojka
	0,3	0,05			0,18	0,02	
été indien	0,28	X			0,11	0,06	ruské kolo
	0,4	X			0,25	0,06	
assiette anglaise	0,01	X			0,14	0,03	(s) kľudom Angličana
	0,04	X			0,43	0,03	
querelle d'Allemand	0,04	0,04			0,04	0,15	francúzsky kľúč
	0,03	0,02			0,1	0,05	
querelle byzantine	0,02	0,04			0,17	0,14	francúzska barla
	0,11	0,05			0,52	0,18	
travail de Romain	0,02	X					
	0,05	X					
file indienne	0,29	0,53					
	0,26	0,32					
signe indien	0,12	0,06					
	0,06	0,07					
calendes grecques	0,11	0,5					
	0,01	0,53					
poupée russe	0,31	0,43					
	0,58	0,46					
supplice chinois	0,02	0,04					
	0,1	0,03					
réponse de Normand	0,04	0,03					
	0,08	0,04					
auberge espagnole	0,29	0,51					
	0,74	0,57					
trou normand	0,07	X					
	0,05	X					
casse-tête chinois	0,1	0,11					
	0,13	0,1					

Tableau 6 : Fréquences relatives des phrasèmes

X = phrasème sans occurrence dans le corpus ou dans le sous-corpus

= *phrasèmes statistiquement significatifs ayant une valeur de fréquence relative d'au minimum 0,17*

= *corpus de contraste pour les phrasèmes statistiquement significatifs dans les textes journalistiques*

3.5. Test MI-Score des phrasèmes

Nous utilisons le test MI-Score pour classer les phrasèmes (composés de deux mots) par rang selon le taux de spécificité. Nous démontrons trois phrasèmes avec ce taux le plus élevé dans le cadre de chaque sous-corpus des textes journalistiques (voir Tableau 7).

Phrasème	Test MI-Score		Phrasème	Test MI-Score	
	Corpus frTenTen 12 Presse	Corpus Emolex Presse		Corpus skTenTen 11 Presse	Corpus prim-7.0-public-inf
roulette russe	10,85	12,134	**ruská ruleta**	11,12	10,207
douche écossaise	10,718	10,795	**škótska sprcha**	7,594	5,901
froid sibérien	11,278	11,58	**sibírska zima**	8,548	8,459
grippe espagnole	8,84	X	**španielska chrípka**	8,666	6,799
guitare espagnole	X	6,137	**španielska gitara**	6,628	5,059
bain turc	7,294	6,583	**turecký kúpeľ**	8,786	6,647
massage thaïlandais	11,693	X	**thajská masáž**	13,463	12,217
gazon anglais	6,564	9,495	**anglický trávnik**	7,78	5,806
petit-déjeuner anglais	X	X	**anglické raňajky**	5,129	6,487
chiffre arabe	3,188	X	**arabská číslica**	4,659	3,349
chiffre romain	5,02	4,75	**rímska číslica**	5,91	5,297
couteau suisse	8,363	9,137	**švajčiarsky nožík**	8,405	5,716
médecine chinoise	5,563	8,57	**čínska medicína**	9,307	8,48
signe chinois	1,429	X	**čínske znamenie**	4,323	2,902
clé anglaise	0,136	5,451	**španielska dedina**	7,856	5,513
canne anglaise	3,785	7,556	**turecký sed**	12,943	3,367
téléphone arabe	12,359	4,723	**francúzsky kľúč**	5,854	3,831
été indien	5,783	X	**francúzska barla**	10,442	8,562
assiette anglaise	4,618	X	**ako Tatár**	2,332	0,273
querelle byzantine	11,204	10,035	**švédska trojka**	4,841	3,926
file indienne	8,09	9,02	**ruské kolo**	5,16	0,133
signe indien	3,575	6,078	**kanadský žart**	8,046	7,005
calendes grecques	8,18	16,532	**mexická vlna**	10,412	8,838
poupée russe	9,882	10,708	**švédsky stôl**	10,685	X

supplice chinois	7,892	8,787
auberge espagnole	11,392	11,702
trou normand	8,528	X
casse-tête chinois	7,79	8,53

Tableau 7 : Fréquences relatives des phrasèmes
X = phrasème sans occurrence dans le corpus ou dans le sous-corpus

= phrasèmes ayant la valeur la plus élevée du test MI-Score dans le cadre de chaque sous-corpus

4. Résultats et discussion

4.1. Requête CQL/ CQP des phrasèmes dans les textes journalistiques

I. Nous pouvons prouver la présence de presque tous les phrasèmes dans les quatre corpus. **II.** Nous ne pouvons pas confirmer la récurrence de quelques phrasèmes (voir Tableaux 5 et 6, le signe X, c'est-à-dire sans occurrence) dans le corpus ou le sous-corpus, par exemple été indien dans le corpus Emolex (ainsi que dans son sous-corpus) ou chiffre arabe dans le sous-corpus EmolexPresse. Par ailleurs, le phrasème petit déjeuner-anglais ne se trouve pas dans les textes journalistiques, nous le trouvons seulement dans le corpus frTenTen12.

4.2. Fréquence relative des phrasèmes dans les textes journalistiques

I. Grâce au calcul de la moyenne des valeurs de fréquence relative de tous les phrasèmes observés, qui est de 0,17 dans les deux sous-corpus de la langue française, nous trouvons 13 phrasèmes statistiquement significatifs dans les textes journalistiques (voir Tableaux 5 et 6). Ce sont : *roulette russe*, *grippe espagnole*, *chiffre arabe*, *couteau suisse*, *médecine chinoise*, *filer à l'anglaise*, *tête de Turc*, *téléphone arabe*, *été indien*, *file indien*, *calendes grecques*, *poupée russe* et *auberge espagnole*. **II.** Grâce au même calcul (de nouveau d'une valeur de 0,17), nous proposons 17 phrasèmes statistiquement significatifs dans les textes journalistiques dans la langue slovaque, à savoir *ruská ruleta*, *ako švajčiarske hodinky*, *španielska chrípka*, *turecký kúpeľ*, *thajská masáž*, *anglický trávnik*, *rímska číslica*, *švajčiarsky nožík*, *čínska medicína*, *mexická vlna*, *raz za uhorský rok*, *španielska dedina*, *turecký sed*, *švédska trojka*, *ruské kolo*, *s kľudom Angličana* et *francúzska barla*. **III.** En synthétisant les sections I. et II., nous trouvons 4 phrasèmes statistiquement significatifs présents dans les deux langues : *roulette russe*, *grippe espagnole*, *couteau suisse*, *médecine chinoise*. **IV.** Grâce au calcul de la fréquence relative des phrasèmes dans les textes journalistiques par rapport à leur corpus de contraste, nous trouvons 21 phrasèmes pour les deux cultures, à savoir 7 pour le français : *roulette russe*, *grippe espagnole*, *tête de Turc*, *téléphone arabe*, *été indien*, *file indienne*, *auberge espagnole* et 14 pour le slovaque : *ruská ruleta*, *ako švajčiarske hodinky*, *španielska chrípka*, *anglický trávnik*, *švajčiarsky nožík*, *čínska medicína*, *mexická vlna*, *raz za uhorský rok*, *španielska dedina*, *turecký sed*, *švédska trojka*, *ruské kolo*, *s kľudom Angličana*, *francúzska barla*.

4.3. Test MI-Score des phrasèmes dans les textes journalistiques

I. Nous pouvons confirmer (voir le Tableau 7) que la plupart des phrasèmes observés, représente la preuve que deux mots sont des collocatifs (selon Hunston, 2002 : 71) dans les textes journalistiques car la valeur de MI-Score est de 3 ou plus, sauf cinq phrasèmes dont la valeur du MI-Score est de moins de 3, à savoir : *signe chinois* et *clé anglaise* dans le sous-corpus frTenTen12Presse, *ako Tatár*, *čínske znamenie*, *ruské kolo* dans le sous-corpus prim-7.0-public-inf ainsi que *ako Tatár* dans le sous-corpus skTenTen11Presse. **II.** Nous trouvons trois phrasèmes avec le taux de spécificité le

plus élevé dans le cadre de chaque sous-corpus. Ce sont les phrasèmes suivants, dans le sous-corpus frTenTen12Presse : *massage thaïlandais*, *téléphone arabe*, *auberge espagnole*, dans le sous-corpus EmolexPresse : *calendes grecques*, *auberge espagnole*, *roulette russe*, dans le sous-corpus skTenTen11Presse : *ruská ruleta*, *thajská masáž*, *turecký sed* et dans le sous-corpus prim-7.0-public-inf : *ruská ruleta*, *thajská masáž*, *mexická vlna*.

5. Conclusion

En guise de conclusion, nous voudrions démontrer l'utilité des méthodes de linguistiques de corpus dans l'analyse des phrasèmes dans les textes journalistiques par rapport à un corpus de contraste. Les méthodes de linguistiques de corpus comme la fréquence relative et le test MI-Score, sont efficaces car, grâce à ces méthodes, nous confirmons les trois hypothèses formulées. Nous pouvons ainsi constater que le langage journalistique se caractérise par la surreprésentation statistiquement significative de certains phrasèmes analysés qui jouent aussi un certain rôle dans les textes journalistiques, il s'agit de 21 phrasèmes (voir 4.2. IV.). Nous justifions aussi empiriquement les phrasèmes statistiquement significatifs dans les textes journalistiques dans la culture française (voir 4.2. I.) et dans la culture slovaque (voir 4.2. II.) ainsi que ceux étant présents dans les deux cultures (voir 4.2. III.). En général, les phrasèmes analysés sont des collocatifs, mais pas par chance. De plus, nous classons ces phrasèmes composés de deux mots par rang selon le taux de spécificité dans les textes journalistiques (voir 4.3. I. et II.).

Nous croyons que ce travail contribue à la recherche interdisciplinaire entre la phraséologie, la linguistique de corpus et la stylistique dans le cadre du domaine des Humanités numériques en Sciences humaines et sociales.

Cette contribution fait partie du projet de subvention KEGA 030EU-4 / 2016 Civilisation de la France et des pays francophones – un nouveau concept de la matière.

Bibliographic references

ASHRAF, M. – MIANNAY, D. 1995, Dictionnaire des expressions idiomatiques françaises. Paris: Librairie Generale Française, 416 pp.

BIBER, D. 2010. Corpus-based and corpus-driven analyses of language variation and use. In: Bernd Heine and Heiko Narrog (eds.), The Oxford Handbook of Linguistic Analysis. Oxford University Press, pp. 159-192.

DINZIKOVA, I. 2017. Specific phrasemes with ethnonyms and their study by corpus analysis. In: Journal of Young Scientists, vol. 47, n. 2, Siauliai University, pp. 18-26.

DOPJEROVA – DANTHINE, M. 2006. Francuzske idiomy pod lupou. Bratislava: Remedium, 335 pp.

EVERT, S. – HARDIE, A. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In: Proceedings of the Corpus Linguistics 2011 conference. University of Birmingham, Available online: https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf.

GAUTIER, L. 2004. Terme, phraseoterme, phraseme: question de delimitation en langue specialisee. In: Le Continuum en linguistique. Jatlaoui, Heidi, et al. (Ed.). Sousse: Faculte des Lettres et Sciences Humaines de Sousse, pp. 153- 172.

GONZALEZ-REY, I. et col. 2002. La phraseologie du français. Toulouse: Presses Universitaires du Mirail, 268 pp.

GRECIANO, G. 1997a. La variance du figement. Dans Les formes du sens: Etudes de linguistique française, medievale et generale offertes a Robert Martin a l'occasion de ses 60 ans. Louvain-la-Neuve, Belgique: De Boeck Superieur, pp. 149-156.

GRECIANO, G. 1997b. Collocations rythmologiques. In: Meta, vol. 17, n. 1, pp. 33-44.
GRIES, S. T. 2010. Useful statistics for corpus linguistics. In: Aquilino Sanchez & Moises Almela (eds.), A mosaic of corpus linguistics: selected approaches. Frankfurt am Main: Peter Lang, pp. 269-291.
GROSS, G. 1996. Les expressions figees en français. Paris: Editions Ophrys, 161 pp.
GRUNDLEROVA, V. et col. 1991. Francuzsko-slovensky slovnik. Bratislava: SPN, 704 pp.
GRUNDLEROVA, V. – SKULTETY, J. – TARABA, J. 1992. Francuzsko-slovensky frazeologicky slovnik A – F. Bratislava: SPN, 492 pp.
GRUNDLEROVA, V. – SKULTETY, J. – TARABA, J. 1992. Francuzsko-slovensky frazeologicky slovnik G – Z. Bratislava: SPN, 609 pp.
HUNSTON, S. 2002. Corpora in Applied Linguistics. Cambridge University Press, 241 pp. ISBN 052180583X, 9780521805834.
HUNSTON, S. 2006. Corpus linguistics. In: K. Brown (Eds.), vol. 3, pp. 234-248.
JAKUBICEK, M. – KILGARRIFF, A. – MCCARTHY, D. – RYCHLY, P. 2010. Fast syntactic searching in very large corpora for many languages. In: Otoguro, R., Ishikawa, K., Umemoto, H., Yoshimoto, K. & Harada, Y. (editors), Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24), pp. 741-747. ISBN: 9784905166009.
KLEIN, B. 2008. Les expressions qui ont fait l'histoire. Paris: Librio, 96 pp.
McENERY, A. et col. 2005. Corpus-Based Languages Studies. Routledge, 408 pp.
MEJRI, S. 2000. Figement et denomination. In: Meta, vol. XLV, n. 4, Montreal: Universite de Montreal, pp. 610-621.
MESKOVA, L. 2004. Terminy a frazemy v textoch z odboru financii. In: Odborna komunikacia v zjednotenej Europe II. Banska Bystrica – Praha: Univerzita Mateja Bela v Banskej Bystrici, pp. 65-70.
MESKOVA, L. – KUBEKOVA, J. 2015. Difficulties in Translating Terminological Phrasemes in Economic Print Media from French, Spanish and English into Slovak – a Contrastive Approach. In: Journal of social sciences, vol. 11, n. 3, pp. 304-316.
NARAY – SZABO, M. 2002. Quelques remarques sur la definition du phraseme. In: Revue d'Etudes françaises, n. 7, pp.71-81.
PAMIES, A. 2011. Phraseologie et competence metaphorique: universaux cognitifs vs. heritage culturel. In: Kaldieva, S. y Zaharieva, R. (eds.): Linguistic Studies in honour of prof. Siyka Spasova-Mihaylovna. pp. 58-75.
PECMAN, M. 2012. Etude lexicographique et discursive des collocations en vue de leur integration dans une base de donnees terminologiques. Terminology, Phraseology and Translation. Special issue of The Journal of specialised translation (JoSTrans). pp. 113-138.
PINCEMIN, B. et col. 2008. Usages linguistiques de la textometrie. Analyse qualitative de la consultation de la Base de Français Medieval via le logiciel Weblex. In: Syntaxe et semantique, vol. 9, n. 1, pp. 87-110, DOI : 10.3917/ss.009.0087. Available online: https://www.cairn.info/revue-syntaxe-et-semantique-2008-1-page-87.html.
PLANELLES, G. 2016. Les 1001 expressions preferees des Français. Opportun, 1 171 pp.
PUCHOVSKA, Z. 2013. La representation des prejuges ethniques en français: quelques notes sur les notions du stereotype linguistique et du stereotype de pensee. In: Philologia, vol.23, n. 1, pp. 49-58.
REY, A. et col. 2013. Le Robert pratique. Paris: Dictionnaires Le Robert, 1660 pp.
REY, A. et col. 2003. Dictionnaire des expressions et locutions. Paris: Dictionnaires Le Robert, 1090 pp.

ROSENBAUM FRANKOVA, L. 2010. Dynamicke procesy vo frazeologii odbornych textov z oblasti ekonomie. Bratislava: Filozoficka fakulta Univerzity Komenskeho, 216 pp.
SCHAEFFER-LACROIX, E. 2015. Analyse de trois systemes de gestion de corpus pour l'enseignement – apprentissage des langues etrangeres. Alsic [online], vol. 18, n. 1. Available online: http://journals.openedition.org/alsic/2852. .
ȚARAN ANDREICI, M. 2016. Translation of phrasemes. Equivalence et nonequivalence. In: Professional Communication and Translation Studies, 9/ 2016. Editura Politehnica, pp. 153-166.
TILLIER, M. et col. 2014. Les tresors de notre langue en 1001 expressions. Points, 567 pp.

Words : 7 166
Characters : 48 164 (26,76 standard pages)

Mgr. Iveta Dinžíková, PhD.
Department of Romance and Slavic languages
Faculty of Applied Linguistics
Economic University in Bratislava
Dolnozemská cesta 1
852 35 Bratislava
Slovakia
iveta.dinzikova@euba.sk