UMONS
Université de Mons

Étude en diachronie courte des empreintes de fréquence et de significativité des termes de l'astrophysique: l'exemple des trous noirs

Charlène Meyers
FTI-EII (UMONS)
charlene.meyers@umons.ac.be

Faculté
de Traduction
et d'Interprétation
Ecole d'Interprètes
Internationaux

Objectifs

- Analyser en diachronie courte
  - Les empreintes de fréquence
  - Les empreintes de significativité
- Corpus de vulgarisation (2004 – 2014)
- Corpus spécialisé (1991 – 2014)
- Mots-clefs (keywords)

Introduction

- Diachronie courte

Phénomènes subtils se manifestant sur une période de temps brève comprise entre 10 et 30 ans (Kytö *et al.* 2000; Mair 1997)

- Pertinence en langue de spécialité

> Bien que développée en langue générale, cette notion de diachronie courte est particulièrement pertinente en langue de spécialité où les changements observés sont en lien avec l'évolution d'un domaine scientifique, évolution souvent très rapide. (Picton 2009 : 17)

Introduction

- Empreintes de fréquence : observer la croissance ou la décroissance d'un terme ou son apparition ou disparition à partir de sa fréquence à travers le temps (Ahmad *et al.* 2002)

> Le terme 'empreinte de fréquence' […] renvoie à la possibilité de dessiner la courbe de fréquence d'un terme à travers les différents sous-corpus composant un corpus diachronique. L'hypothèse est de considérer que la courbe de fréquence d'un terme donné dans le corpus reflète l'évolution du terme/concept dans le domaine. (Picton 2009 : 106)

Introduction

- Empreintes de significativité: (par analogie) les marqueurs de la croissance ou la décroissance de la significativité d'un terme à travers le temps ou encore l'apparition ou la disparition d'un terme significatif dans les textes les plus récents d'un corpus.

Hypothèses

- Variation de la fréquence des mots-clefs en diachronie
- Variation de la significativité des mots-clefs en diachronie
- Variation de la fréquence ≠ variation de la significativité

Méthodologie

- Corpus de vulgarisation (VULG): 6 textes (S*cientific American*) sur les trous noirs
- Corpus spécialisé (SPEC) : 13 textes (« to explore more »)
- Textes nettoyés, convertis. Substitution des formules, tableaux et figures par [FRML], [TABL] et [FIGR].

Méthodologie

	Sous-corpus											
VULG	/	/	/	/	/	2004	2005	/	2009	2012	2013	2014
Tokens : 22444												
Types : 3683												
SPEC	1991	1999	2000	2001	2002	2004	2005	2006	2009	/	/	2014
Tokens : 55448												
Types : 4709												

Méthodologie

- Liste des mots-clefs pour SPEC et VULG dans AntConc
- Corpus de référence : BROWN corpus
- Classement des mots-clefs selon leur importance (= *keyness*), mesurée par un test statistique

Méthodologie

- Test de significativité dans AntConc: Khi-carré et rapport de vraisemblance (*Loglikelihood test* = LL)
- Dans cette étude: rapport de vraisemblance
- Test adapté:
- Aux petits échantillons (Anthony 2017)
- Aux données réparties de façon non aléatoire (dans un contexte naturel, les mots ne s'assemblent pas de façon aléatoire) (Xiao 2013)

Liste des mots-clefs dans SPEC

Rang	Freq	Keyness (LL)	Keyword
1	273	1330,292	hole
2	252	1321,867	horizon
3	305	1196,025	black
4	194	1152,013	FRML
5	188	1070,509	quantum
6	163	955,84	singularity
7	139	825,411	Hawking
8	188	759,044	energy

Liste des mots-clefs dans VULG

Rang	Freq	Keyness (LL)	Keyword
1	159	980,375	hole
2	187	940,426	black
3	141	834,105	universe
4	84	634,708	singularity
5	66	476,717	gravitational
6	64	456,18	quantum
7	72	411,091	holes
8	64	359,197	star

Résultats

- 13 mots en commun aux 2 listes (43,33%)
- FRML
- Nom propre (« Hawking »)
- Mots thématiques
- SPEC → mathématiques / sciences / modèles computationnels (equations, scale, FRML, computer, bits, computation)
- VULG→ contexte plus général (stars, light, galaxy, cosmic)

Empreintes de fréquence

- A partir des 2 listes : analyse de la fréquence dans les sous-corpus diachroniques
- Empreintes de fréquence : croissance, décroissance, apparition, disparition d'un terme au cours du temps (Ahmad & Musacchio 2004).

Empreintes de fréquence

- Utilité :
- Même empreinte = regroupement thématique?
- Empreintes opposées = phénomène de mort terminologique, substitution, changement thématique?
- Picton a observé des changements thématiques. Ex.: Obsolescence de certaines fonctionnalités d'une balise spatiale

Empreintes de fréquence

Ex. dans SPEC: « Computation »

	1991	1999	2000	2001	2002	2004	2005	2006	2009	2014
14. computation	4	1	43	14	17	4	0	0	0	0

Empreintes de fréquence

- 3 catégories d'empreintes de fréquence:
- Empreintes discontinues, sans rupture : fluctuation de fréquence sans apparition ou disparition d'un mot. Ex.: gravitational, black, hole
- Empreintes discontinues, avec rupture : fluctuation de la fréquence avec apparition ou disparition d'un mot. Ex: laser, computation, brane
- Empreintes continues sur au moins 3 périodes (croissance ou décroissance). Ex.: field, modes, scale

Empreintes de fréquence

- Recherche contextuelle = explication des phénomènes (croissance, disparition, etc.)
- Ex. : disparition de « computation » → limite du modèle computationnel?

Whether or not it is possible to make computation take place in the extreme regimes envisaged in this paper is an open question. The answer to this question lies in future technological development, which is difficult to predict. (Extrait de SPEC 2000)

Still one would like to understand how to modify Hawking's original computation so that it is consistent with unitarity. (Extrait de SPEC 2004)

Empreintes de fréquence

- Fréquences des mots-clefs de SPEC et VULG varient au cours du temps
- Certaines empreintes sont liées : pics de fréquence de « computer », « computation » et « bits » à la même période. Idem pour « field » et « collapse ». Idem pour « naked » et « singularity »
- Vérification de la première hypothèse

Empreintes de significativité

- Pourquoi étudier la significativité?
- Fréquence = statistiques descriptives
- Fréquence = tributaire de la taille du corpus
- Empreintes de significativité: marqueurs de croissance ou de décroissance de la significativité d'un terme au cours du temps ou de l'apparition ou de la disparition d'un terme significatif

Interprétation du rapport de vraisemblance

- La valeur LL dans la colonne « keyness » = significativité (et donc importance) du mot
- + la valeur de LL est élevée, + le mot est significatif
- Seuil de significativité LL ≥ 6,635 (p<0,01)

→ permet de dire avec 99% de certitude que les résultats obtenus au-dessus de ce seuil ne sont pas dus au hasard.

Empreintes de significativité

- 3 catégories d'empreintes de significativité:
- Empreintes discontinues sans rupture: fluctuation de la valeur LL du mot mais sans qu'elle ne descende en-dessous du seuil. Ex.: gravity
- Empreintes discontinues avec rupture: fluctuation de la valeur LL du mot mais celle-ci est parfois inexistante ou passe en-dessous du seuil. Ex.: gravitational, laser, brane
- Empreintes continues sur au moins 3 périodes (croissance ou décroissance). Ex.: field, modes, quantum

Empreintes de significativité

Ex. : Computation

	1991	1999	2000	2001	2002	2004	2005	2006	2009	2014
14. computation	25,309	4,372	386,606	119,504	155,888	30,53				

- Variations de la significativité en diachronie = 2e hypothèse confirmée

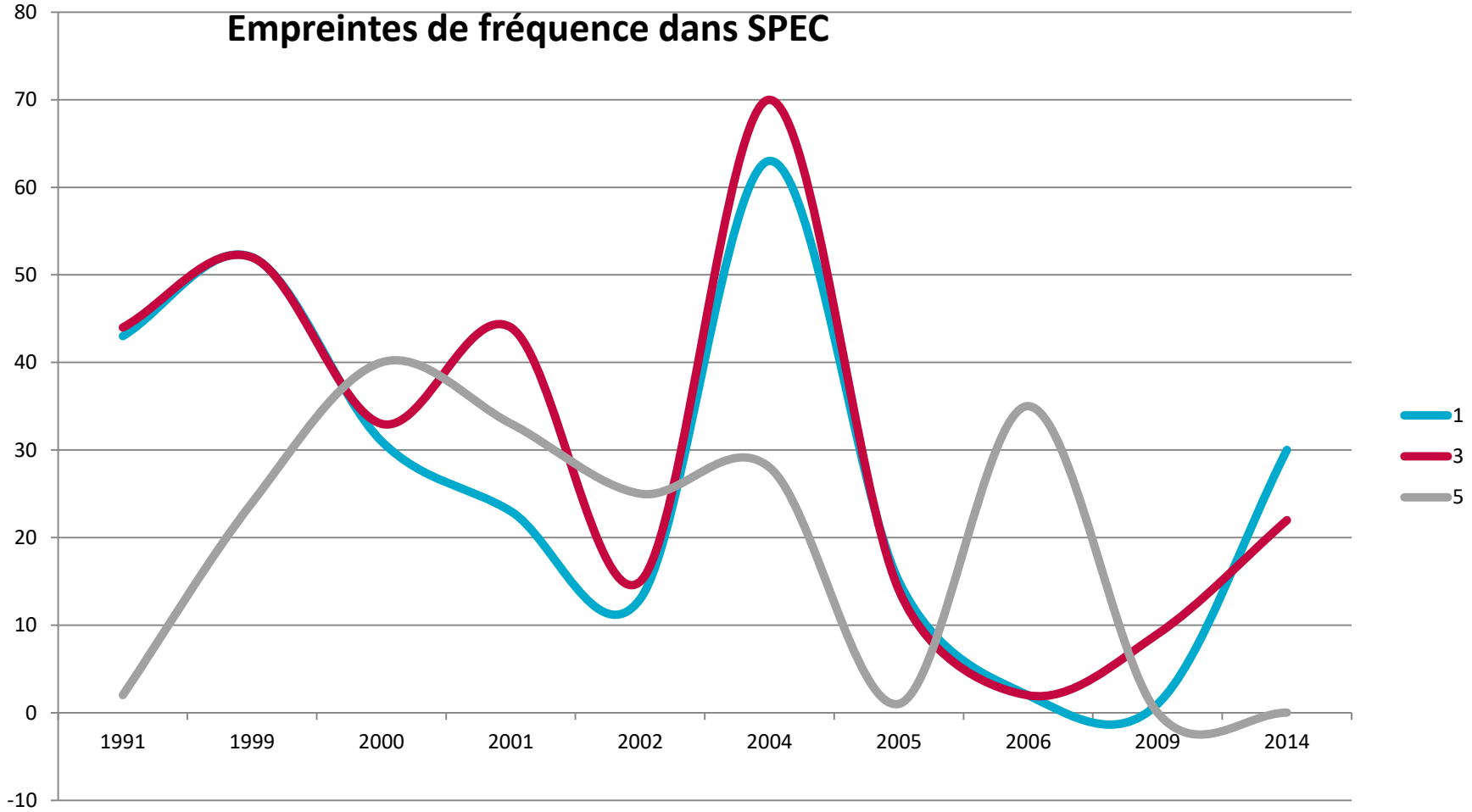
Empreintes de significativité

- Globalement : les empreintes de significativité suivent les empreintes de fréquence
- Observation la plus importante : chute de la significativité en-dessous du seuil établi
- Ex. : « computation » =
- Fréquence fluctuante mais sans rupture entre 1991 et 2004.
- Rupture de significativité en 1999

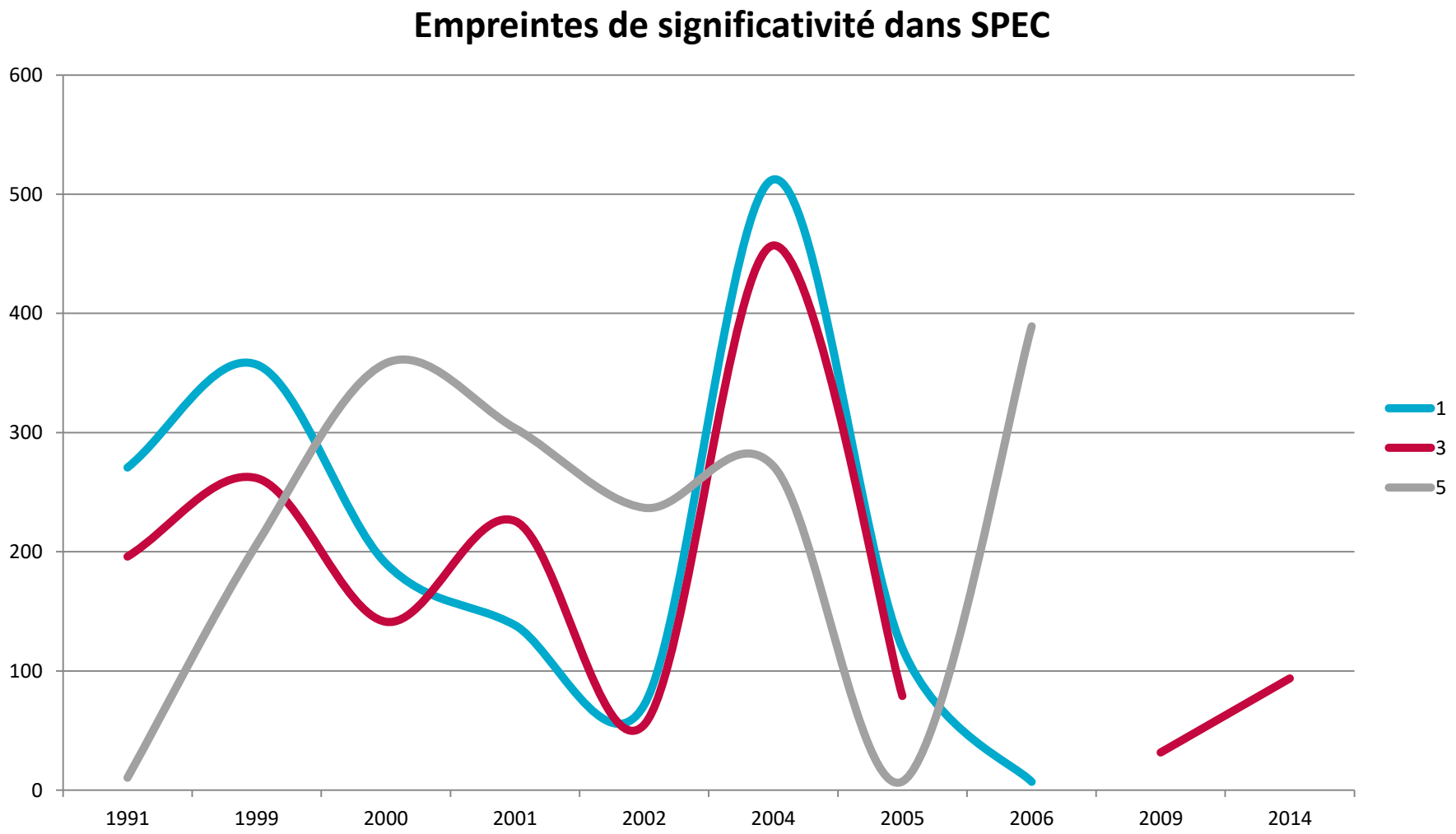
Empreintes de significativité

- Un mot peut avoir une fréquence supérieure à un autre mot pour une année donnée mais être moins significatif que cet autre mot pour la même année (et vice versa)

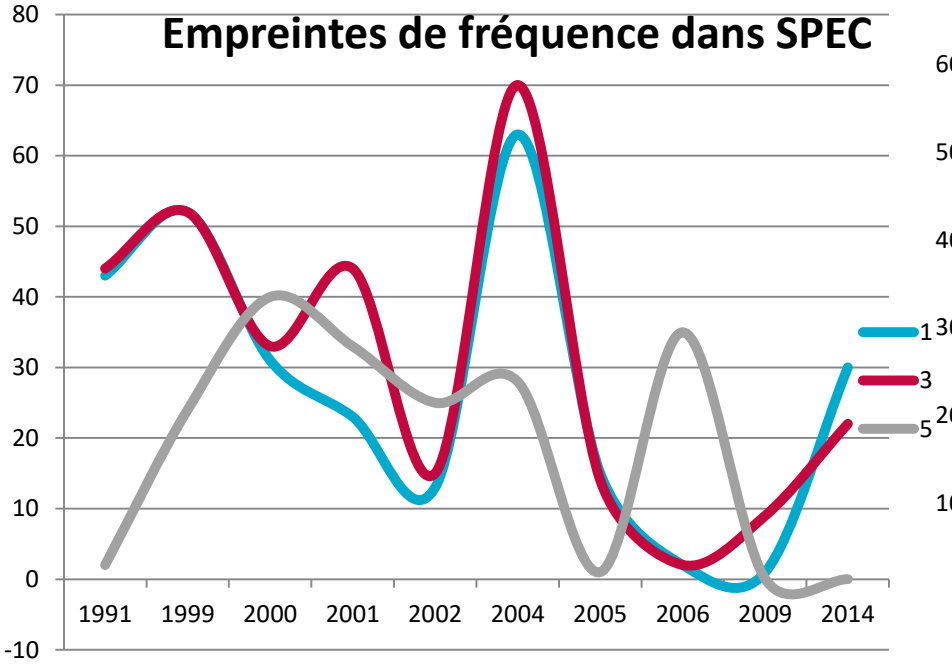
Courbe de fréquence

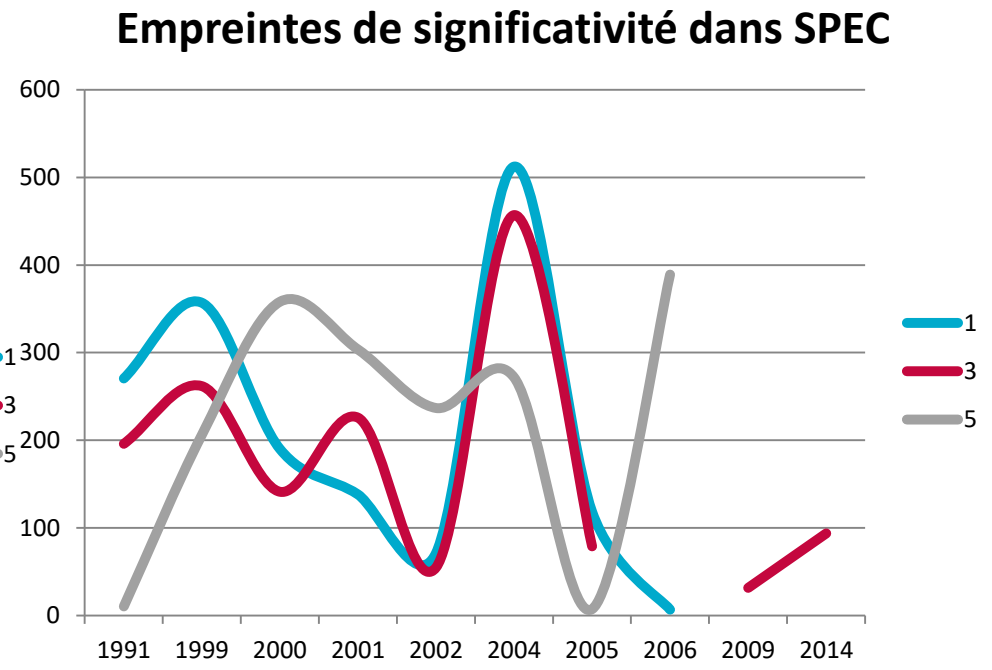


Courbe de significativité



Comparaison des courbes





Empreintes de fréquence ≠ empreintes de significativité
= Hypothèse 3 vérifiée

Conclusion

- Étude des empreintes de fréquence et de significativité confirme les 3 hypothèses
  - Diverses manifestations d'empreintes de fréquence sont observables dans SPEC et VULG
  - Diverses manifestations d'empreintes de significativité sont observables dans SPEC et VULG
  - L'évolution des empreintes de significativité ne suit pas toujours celle des empreintes de fréquence
- Diachronie courte permet d'étudier des changements subtils même sur des petits corpus
- Importance de combler les statistiques descriptives par les statistiques inférentielles

Bibliographie (1)

AHMAD, K. & MUSACCHIO, M. T. (2004) “Discovery of (New) Knowledge and the Analysis of Text Corpora”. *In Actes de la 4ème conference international “Language Resources and Evaluation”,* Lisbonne, Portugal, 24-30 mai 2004, pp.1567-1570.

ANTCONC DISCUSSION GROUP. (2017). « the p value and log likelihood in Keyword ». Google group. <https://groups.google.com/forum/#!topic/antconc/HkFbCkPsmII>

ANTHONY, L. (2014). AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from <http://www.laurenceanthony.net/software>

DUNNING, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19, 1, March 1993, pp. 61-74

GRANGER, S. (1998). « The computer learner corpus: a versatile new source of data for SLA research ». In S. Granger (ed.) Learner English on Computer. Longman, London, pp. 3 – 18.

HUNSTON, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

KYTÖ, M., RUDANKO, J. & SMITTERBERGE, E. (2000) “Building a Bridge Between the Present and the Past: A Corpus of 19th – Century English”. *ICAME Journal, 24, pp.85-97.*

LEVISHNA, N. (2015). *How to do Linguistics with R: Data exploration and statistical analysis.* Amsterdam : John Benjamins.

MAIR, C. (1997) « Parallel Corpora: A Real-Time Approach to the Study of Language Change in Progress ». In M. Ljung (Éd.), Corpus-Based Studies in English, GA-Rodopi, Amsterdam & Atlanta, pp.195-209.

Bibliographie (2)

PICTON, A. (2009) « Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial », Thèse de doctorat. Université Toulouse le Mirail.

RAYSON P., BERRIDGE, D. and FRANCIS, B. (2004). « Extending the Cochran rule for the comparison of word frequencies between corpora ». In *Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgique, Mars 10-12, 2004, Presses universitaires de Louvain, pp. 926-936.*

RAYSON, P. (2003). « Matrix : A statistical method and software tool for linguistic analysis through corpus comparison ». Thèse de doctorat. Lancaster University.

RAYSON, P. & GARSIDE, M. (2000). « Comparing Corpora using Frequency Profiling ». In Proceedings of the workshop on Comparing corpora. Vol. 9 : 1-6.

SCOTT, M. (2000). « Focusing on the text and its key words ». In Burnard, L. and McEnery, T. (eds.) Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora. Peter Lang, Frankfurt, pp. 104 – 121.

XIAO, R. (2013). « Making statistic claims ». Présentation Power Point. Lancaster University.