# Projet transdiciplinaire "Cooccurrences et collocations"

**1 - Éléments de présentation** (nom du porteur, laboratoire(s) associé(s)

Didier Josselin (géographie/informatique) est le porteur de ce projet, en association avec :
- Marc El Bèze (informatique)
- Loïc Grasland (géographie)
- Andrea Venturelli (mathématiques)

L'UMR ESPACE et le LIA de la SFR Agor@ntic sont associés dans cette action interdisciplinaire. Le laboratoire de mathématique y participe également. Les autres laboratoires de la SFR pourront y participer, tant le thème est propice aux points de vue et aux échanges entre les disciplines. De même, le séminaire du projet est ouvert aux membres de la SFR Tersys.

La notion de cooccurrence (*vs* de collocation) est explicitement uin champ de recherche méthodologique et quantitatif en géographie, en statistique-mathématique et en linguistique computationnelle (informatique).

**2 - Descriptif du projet**

Le dictionnaire Larousse définit la cooccurrence comme l' "apparition dans un même énoncé de plusieurs éléments linguistiques distincts, c'est à dire la relation qui existe entre ces éléments. Dans la phrase *Le chat dort*, *chat* est en relation de cooccurrence avec *le* et *dort*." C'est la présence de ces deux termes et de leur association qui constitue le sens de la phrase, le tout étant plus que la somme des parties. On retrouve cette notion, en géographie avec une signification sensiblement différente, où la cooccurrence est le fait que des objets ou des individus qui se ressemblent sont d'autant plus proches dans l'espace. Elle s'oppose à une répartition aléatoire des observations, montrant que l'espace n'est pas isotrope : dans une certaine mesure « qui se ressemble s'assemble ». Du point de vue social, cela renvoie à des comportements grégaires ou des ségrégations que l'on peut observer, mesurer, dans des réseaux sociaux par exemple. La cooccurrence peut même être généralisée avec la dimension temporelle. On parle alors de cooccurrence spatio-temporelle, dont l'analyse est utile par exemple pour la description des faits historiques. En bouclant sur l'analyse des textes, une séquence de mots contient aussi implicitement cette dimension temporelle.

*Un point de vue de géographe.*
La cooccurrence peut cacher des réalités différentes. Si des collègues avignonnais géographes s'éloignent physiquement du campus Ste-Marthe tout en gardant un contact fort avec ce site, comment interpréter la baisse de la mesure de l'« autocorrélation spatiale » dans la localisation des géographes d'Avignon ? L'intérêt scientifique commun de ce projet pour l'ensemble des participants relève de la ressemblance/proximité dans la thématique d'étude. En d'autres temps, il se serait traduit par un rapprochement géographique. Paradoxalement, la mesure d'autocorrélation a sans doute diminué aujourd'hui dans ce cas. Pourtant, l'intérêt commun interdisciplinaire scelle la cooccurrence. Autrement dit, des dispositifs, en particulier les outils

numériques mais pas seulement, conduiraient peut-être à faire éclater cette mesure de « qui se ressemble s'assemble ». Des phénomènes notables de clustering (géographiques) ont aussi émergé depuis quelques années. Il serait intéressant de savoir si les phénomènes d'autocorrélation spatiale diminuent à travers le temps, et dans quels domaines, dans quelles conditions,... notamment dans une période récente (effets du numérique ? en lien avec les thématiques de la SFR). L'intérêt de recherche est donc autant thématique que méthodologique.

*Un point de vue d'informaticien.*
Si on a des méthodes pour calculer les collocations et en lister les propriétés :
- *non – compositionnel : Vin blanc (le vin blanc n'est pas blanc ...)* ,
- *non-substituable : Vin jaune (ne colle pas)* ,
- *non modifiable : Chercher midi (trente?) à 14 h* ,
il est encore difficile de définir ce que sont précisément les collocations . Pour s'en convaincre, trois exemples de définitions données par 3 sources différentes :
*"An expression consisting of two or more words that correspond to some conventional way of saying things."* Ch. 5 de FSNLP
*"Collocations of a given word are statements of the habitual or customary places of that word."* Firth (1957)
*"A phrase that means more than the sum of its parts."* Dustin
Une des idées serait donc de réfléchir ensemble sur ce que l'on peut dire des co occurrences en les opposant de façon contrastive aux collocations ou à d'autres phénomènes étudiés dans chacune des disciplines concernées. Les approches employées pour les détecter et les exploiter sont elles réutilisables, ou transposables ? Et si ce n'est pas le cas à nous d'en proposer de nouvelles ...

## 3 – Objectifs généraux et résultats attendus

Le projet « cooccurrence » vise à réunir un groupe de chercheurs de différents horizons pour discuter de cette notion de façon transversale et évaluer en quoi sa conception multiple peut apporter des avancées dans chaque discipline et pour la science en général. L'approche prônée est méthodologique. On se posera notamment les questions suivantes :
- quelle définition de la cooccurrence est donnée dans la discipline ?
- qu'apporte particulièrement sa prise en compte dans les problématiques ?
- comment l'évaluer qualitativement, la mesurer quantitativement ?
- quelles dimensions sont concernées (spatiale, temporelle, abstraite) ?
- quels problèmes spécifiques se posent à son sujet dans la discipline ?
- que peut-on transposer de la cooccurrence d'une discipline à l'autre et comment ?
- que peut-on trouver de spécifique et de commun à la cooccurrence ?
Le projet vise à l'organisation d'un séminaire d'une journée (ou deux), suivi(s) de la rédaction d'une série d'articles sur le thème de la cooccurrence en sciences exactes et humaines et sociales sociales. L'objectif est d'éditer un numéro spécial dans une des rares revues pouvant l'accepter :
- soit en français (revues plutôt SHS comme, Sciences de la Société, http://w3.scsoc.univ-tlse2.fr/, A contrario, http://www.asso-unil.ch/acontrario/, revue Tracés (sciences humaines) http://www.fabula.org/actualites/traces-revue-interdisciplinaire-de-recherche-en-sciences-humaines_4038.php, voire les Editions Universitaires d'Avignon

ou les Presses Universitaires de Provence, ou Hermès, http://documents.irevues.inist.fr/handle/2042/8538,
- soit en anglais (de type « online international interdisciplinary research journal », http://www.oiirj.org/oiirj/?page_id=850).
Une autre possibilité serait de créer une revue spécifique auprès d'Hermès-Lavoisier, en langue française et anglaise, dans le cadre d'une procédure qu'a commencé à engager Didier Josselin, mais qui pour l'instant est en attente, faute de contenu. Le projet « cooccurrence » pourrait servir de tremplin à cette démarche plus large et permettre de lancer la dynamique de cette revue réellement interdisciplinaire, et non pas basée sur des thématiques ou des régions géographiques, mais sur des mots clés méthodologiques.

## 4 – Caractère innovant de ce projet

Le projet est à la fois modeste (durée assez courte, objectif ciblé, séminaire, faible budget) et ambitieux (ouverture interdisciplinaire, publication d'un numéro spécial). Dans notre recherche de revues interdisciplinaires, nous avons constaté que beaucoup de ces revues, même si elles intègrent dans le titre le terme « interdisciplinaire », sont très ciblées (revue interdisciplinaire *des travaux sur les Amériques, d'études juridiques, d'études hispaniques, sur la Grèce ancienne, du développement cognitif, sur le management et l'urbanisme*, etc.). Ces revues escomptent probablement que cette cooccurrence leur apporte des soumissions d'articles ! Notre projet tente d'aller au-delà et est en quelque sorte un argument pour réfléchir à la création ou au recours adéquat à une vraie revue interdisciplinaire, de bon impact, pour diffuser les résultats de ces regards croisées sur la cooccurrence. Un autre problème que nous avons identifié est le fait que les revues dans lesquelles il serait éventuellement possible de publier ce numéro thématique sont essentiellement du domaine des SHS. En effet, il ne faudrait pas que ce choix occulte la forte dimension méthodologique du projet, qui, pour le coup, se rapproche des préoccupations des sciences exactes. Cet objectif constitue en soi un verrou qui, s'il est surmonté, revêt un réel caractère innovant. La dimension épistémique (et épistémologique) du projet est également, en soi, un point scientifique fort.

## 5 – Sa dimension interdisciplinaire

Dans ce paragraphe, nous citons une recherche par discipline (dont d'ailleurs les contours et l'appartenance à telle ou telle discipline ne sont pas si clairs), pour montrer l'existence de la cooccurrence (ou de notions apparentées) dans les problématiques des disciplines. Ces citations n'ont pas la prétention de couvrir le champ, mais simplement d'illustrer le propos, vues la variété des approches et la portée limitée de notre connaissance, à l'heure actuelle, sur ce sujet transdisciplinaire. C'est l'objet même du séminaire et de sa publication conséquente de nous éclairer sur cette vaste bibliographie transdisciplinaire.

Pour illustrer le caractère holistique de la notion du point de vue scientifique et son adéquation avec les objectifs de la SFR Agor@ntic (culture numérique) citons par exemple [1].

La dimension interdisciplinaire est fondamentale dans ce projet. Il n'existerait pas sans elle. Les disciplines concernées par la cooccurrence sont

nombreuses :
- l'informatique (analyse des régularités et associations dans les textes...) [2]
- la géographie (mesure de l'autocorrélation spatiale positive, ségrégation socio-spatiale...) [3]
- l'histoire (phénomènes de co-présence dans le temps ou lors d'évènements, dans un ordre total ou un ordre partiel...) [4]
- la statistique (méthodes de clustering, probabilités conditionnelles...) [5]
- la sociologie (définition de profils et de comportements types, effet de clubs...) [6]
- les sciences juridiques (rigueur et pertinence des textes de loi par l'association de termes appropriés..) [7]
- les mathématiques (matrices de cooccurrences, graphes de similitude, métriques non euclidiennes...) [8]
- les langues (phrasé, instrumentation du suspense, utilisation de registre lexicaux variés, codage de l'expression selon le degré d'initiation, etc.) [9]
- la philosophie [10]
- etc.

Nous ajoutons à ces quelques références des citations plus spécifiques au traitement informatique linguistique [11 à 15].

**6 - Positionnement dans la (S)FR**

Ce projet s'inscrit résolument dans l'axe transversal méthodologique de la FR
Si des propositions des collègues concernent les technologies du numérique au sens large, un lien peut subvenir avec les autres axes de la SFR.

**7 - Partenariats extérieurs (en cours et à venir)**

Ce projet est soutenu par deux actions complémentaires :
- UMR ESPACE CNRS : séminaires méthodologiques inter-sites (Marseille, Nice, Aix, Arles) ; responsables : Christine Voiron (directrice UMR), Loïc Grasland (responsable de l'équipe d'Avignon)
- GDR MAGIS (sciences de l'information géographique) : axes Extraction et Recherche d'Informations Géographiques (responsable Mauro Gaio) et Dynamiques Spatio-temporelles (responsable Léna Sanders)

**8 - Budget prévisionnel et financements envisagés**

Nous demandons **3000 €** pour couvrir les dépenses d'organisation d'un unique séminaire en automne et d'invitation de quelques participants extérieurs.

Voici le détail des dépenses prévues :
- 2000 euros : frais de déplacement des invités (4 invités)
- 1000 euros : bibliographie, organisation du séminaire (repas, logistique)

**9 - Références**

[1] Leydesdorff Loet, Qiu Vaughan Liwen, (2006), Co-occurrence Matrices and their Applications in Information Science: Extending ACA to the Web Environment, *Journal of the American Society for Information Science and Technology (JASIST)*, 57(12) (2006) pp. 1616-1628.

[2] Globerson Amir, Chechik Gal, Pereira fernando, Naftali Tishby, (2007), Euclidean Embedding of Co-occurrence Data, *Journal of Machine Learning Research 8* (2007), pp. 2265-2295

[3] Foltête Jean-Christophe, (2003), Reconstitution d'une diffusion spatiale à partir d'une succession d'états, *L'espace géographique*, 2003/2 (tome 32), pp. 171-183.

[4] Veyne Paul, (1971), *Comment on écrit l'histoire. Essai d'épistémologie.* Paris, Ed. du Seuil, 352 pages.

[5] Moalla Koubaa Ikram, (2009), Caractérisation des écritures médiévales par des méthodes statistiques basées sur la cooccurrences. *Thèse en informatique,* INSA de Lyon.

[6] Moscovici Serge et Henry Paul, Problèmes de l'analyse de contenu (1968), *Langages, socio-linguistique,* Vol. 3, N° 11, pp. 36-60

[7] Pisetta Vincent, Hacid Hakim, Bellal Fazia, Ritschard Gilbert et A. Zighed Djamel, (2006) Traitement automatique de textes juridiques, in R. Lehn, M. Harzallah, N. Aussenac-Gilles, J. Charlet (eds), Semaines de la connaissance, SdC 2006, Nantes 26-30 juin, Actes électroniques sur CD-Rom

[8] Matthias Tauveron (2012), De la cooccurrence généralisée à la variation du sens lexical, in La cooccurrence, du fait statistique au fait textuel, (Damon Mayaffre et Jean-Marie Viprey, eds), CORPUS, 11, 2012

[9] Salazar-orvig Anne, *Les mouvements du discours. Style, référence et dialogue dans des entretiens cliniques*, Paris, L'Harmattan, 1999, 294 p.

[10] Prévost Marie et Debruille Jacques Bruno, (2013), Cooccurrence des croyances religieuses, superstitieuses et de type délirant, *Santé mentale au Québec*, Volume 38, numéro 1, printemps 2013, pp. 279-296

[11] F Smadja , K McKeown, Automatically extracting and representing collocations for language generation, Proc. of the 28th conf.on ACL, p.252-259, juin 1990, Pittsburgh, Pennsylvania

[12] Y. Choueka, T. Klein, E. Neuwitz. Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus. Journal for Literary and Linguistic computing, 4:34-38, 1983.

[13] K Church , P Hanks, Word association norms, mutual information, and lexicography,Proc. of the 27th conf.on ACL, p.76-83, juin 1989, Vancouver, British Columbia, Canada

[14] C. Manning, H Schütze, Foundations of statistical natural language processing. Cambridge (Mass.) ; London : MIT Press, c1999

[15] Kilgarriff, Rose Measures for corpus similarity and homogeneity. In Proc. 3rd Conf. On EMNLP–3, pp 46–52, Granada, Spain, juin 1998