Les propriétés grammaticales du genre de l'offre d'emploi aux fondements d'une méthode de classement automatique.

Romain Loth, Fanny Rinck

▶ **To cite this version:**

Romain Loth, Fanny Rinck. Les propriétés grammaticales du genre de l'offre d'emploi aux fondements d'une méthode de classement automatique.. Claire Despierres & Mustapha Krazem. Quand les genres de discours provoquent la grammaire.. et réciproquement, Editions Lambert Lucas, pp.203-222, 2012, Linguistique, 978-2-35935-028-9. halshs-00591606

HAL Id: halshs-00591606
https://shs.hal.science/halshs-00591606v1

Submitted on 24 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quand les genres de discours provoquent la *grammaire*?et réciproquement !
Colloque à l'université de Bourgogne, *Dijon* les 13 et 14 avril 2011

Les propriétés grammaticales du genre de l'offre d'emploi aux fondements d'une méthode de classement automatique

Romain Loth et Fanny Rinck
Laboratoire Modyco, UMR7114, CNRS/ Université Paris Ouest Nanterre La Défense

Introduction

Notion « bi-face » (Branca-Rosoff, 1999), le genre textuel met en correspondance des conditions de production et des propriétés linguistiques relativement stables. Les régularités caractéristiques d'un genre sont des régularités formelles, mais concernent également le contenu des textes. Citant comme exemple Dante, qui soulignait que le sonnet convient aux armes et aux amours, F. Rastier (2001 : 249) montre que l'interdépendance entre contenu et expression est au coeur des descriptions spontanées des genres et que le genre conduit à poser le problème de la semiosis textuelle : il définit un rapport normé entre signifiant et signifié au palier textuel. Le contenu des textes varie au sein d'un même genre, comme leur forme, mais le genre privilégie des types sémantiques ou une mise en scène de l'information qui constituent un horizon d'attente dans l'interprétation. Dans la lignée des travaux de D. Biber (1988, 1993), qui a mis en évidence des types de textes sur la base d'agglomérats de traits linguistiques, la caractérisation empirique des genres a trouvé un essor considérable en s'appuyant sur l'analyse automatisée des données textuelles. Pour autant, définir un genre ne saurait s'arrêter aux traits morpho-syntaxiques qui se prêtent à une analyse automatisée.
Dans le domaine de l'extraction d'information, le risque pourrait être, inversement, d'occulter la forme des textes et de s'intéresser directement à leur contenu. Pourtant la forme générique est le fondement de ce qu'est l'information à extraire, le texte structurant le propos. Faciliter la navigation documentaire impose alors de faire émerger une sémantique de spécialité et une sémantique du genre (Hjorland 2002).
L'objectif du projet dont nous rendons compte ici, le projet « Sémantique, Internet et Recherche d'Emploi » développé au laboratoire Modyco porte sur les offres d'emploi et vise une méthode d'étiquetage de documents à partir de la caractérisation des propriétés sémantiques des termes qu'il contient. Le principe est le suivant : le classement de termes en familles permet le classement de documents selon certains de leurs termes jugés pertinents (tags). Dans ce domaine de l'étude des langues de spécialité, la méthode d'analyse distributionnelle dite LSA (Latent semantic analysis) et des méthodes cousines, comme les espaces dits « sublexicaux » ont permis de grandes avancées pour modéliser de façon empirique les prédilections thématiques attachées à tel ou tel terme : ainsi pour les offres d'emploi, il est relativement facile de rattacher des termes comme « centre hospitalier », « soins » ou « médecin » aux professions médicales. Cependant, il ne s'agit pas uniquement de classer les offres d'emploi en fonction de métiers.
L'indexation des offres via leur classement ou étiquetage se joue plus largement au niveau des *types d'information* : c'est l'annonce du métier, la description du secteur et de l'entreprise, les missions liées au poste, les compétences, la formation et l'expérience requises. Ils se manifestent dans le découpage du texte en séquences, dans les mots-clefs utilisés pour résumer le texte, et, comme on va le voir, dans les propriétés syntaxiques des termes eux-mêmes. Formant l'horizon d'attente du lecteur, ils sont tout à fait cruciaux pour la navigation dans les bases d'offres d'emploi de façon plus souple que par les catégories de métiers. Le projet SIRE s'est intéressé à a manière dont ces types d'informations se manifestent dans le lexique des offres, qui est à la fois le lexique propre aux métiers et celui du domaine général du recrutement. On fait l'hypothèse que les types d'informations (comme métier, missions, compétences) se situent au niveau textuel, mais s'appuient

sur des classes d'objets dont la sémantique se définit au sein du niveau plus micro des relations syntaxiques. On propose donc de voir dans quelle mesure une approche basée sur une analyse distributionnelle des dépendances et plus précisément sur une approche de type DSM (Distributional Semantic Models) va permettre de retrouver, de spécifier ou de refonder les types d'informations qu'on peut identifier manuellement.

1 Objectifs/ Démarche proposée

En pratique, l'enjeu attaché à cette partie du projet SIRE était d'attribuer aux différents mots-clefs[1] des types informationnels : par exemple, « agro-alimentaire [*secteur*] », « agent de conditionnement [*métier*] », « conditionnement manuel [*tâche*] ». On se propose de suivre la répartition des lexèmes dans différents contextes syntaxiques pour caractériser les termes, et analyser ce faisant le genre comme une configuration d'usages phraséologiques et textuels. L'approche est empirique et inductive et s'appuie sur plusieurs techniques de modélisation des proximités lexicales sur grands corpus (cf. Figure 1). En observant le comportement des termes, on cherche à mettre en évidence un code sémantique pré-établi par les réalités des métiers et des ressources humaines et par la rédaction des offres (copier-coller, reprise de fiches, validation extérieure, entretiens avec la personne ayant occupé le poste...).

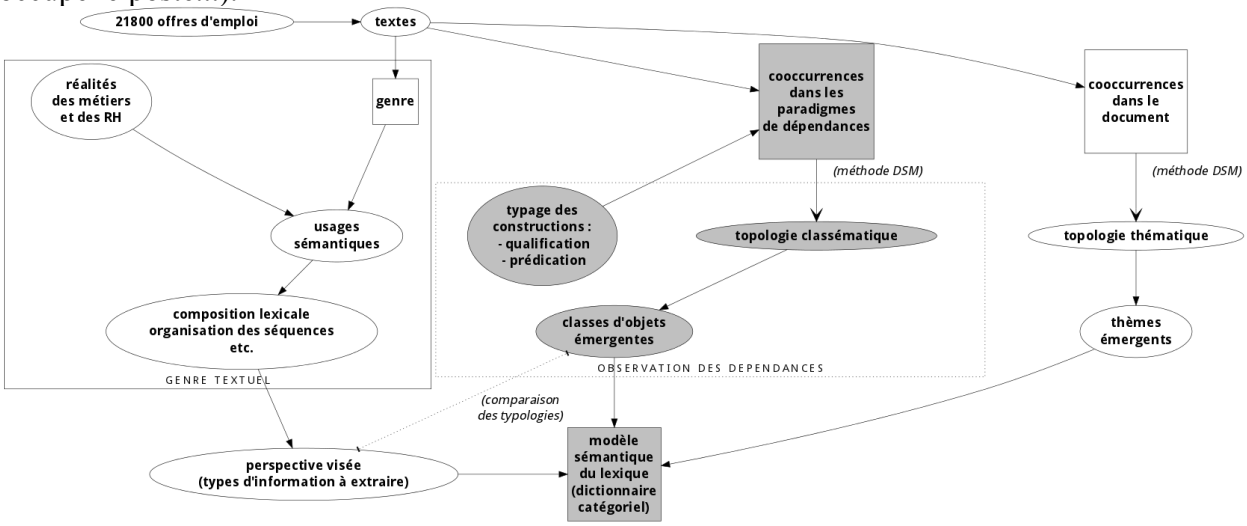

Figure 1: Schéma de la création du lexique des métiers et compétences SIRE. Les noeuds grisés correspondent à la partie développée dans cet article

Une première approche exploratoire avait consisté à annoter manuellement le corpus d'offres d'emploi pour identifier l'apparition de séquences informationnelles propres au genre textuel (Loth, 2010)[2]. L'enjeu suivant est alors d'étudier ces types d'informations sur le plan lexical, en se basant sur une méthode d'apprentissage automatique par analyse distributionnelle. Cette méthode a l'avantage d'opérer sur des mesures de co-occurence généralisées qui permettent des opérations de regroupement de proche en proche pour former un modèle spacial ainsi que la mise au jour de classes émergentes (partition en régions de l'espace ou *clustering*). Cet espace distributionnel qu'on a appelé par la suite « topologie » va ainsi constituer une grille de lecture synthétique pour étudier chaque mot de chaque texte au travers des faisceaux de constructions typiques où il apparait.
Sur la base de ces résultats obtenus de façon non-supervisée, il sera alors relativement facile de faire un travail manuel de « regroupement de regroupements ». Pour résumer, nous allons obtenir pour chaque terme une information automatique sur ses traits sémantiques émergents comme [agent], [lieu], [activité] ou [objet concret] qu'on pourra ensuite retravailler dans une perspective d'extraction de

[1] Les mots-clefs sont des termes soit relevés dans le texte soit voisins sémantiques qui peuvent constituer un *descripteur* pertinent du document (ou « *tag* »).

[2] La typologie qui a été établie à cette occasion peut être consultée dans la Table 1, utilisée en fin de cet article comme outil d'évaluation des classes lexicales obtenues par apprentissage DSM.

métadonnées indexables (*métier, secteur, mission, compétences...*). On retire de cette analyse un dictionnaire catégoriel qui vient s'articuler aux autres travaux du projet.
La méthode traditionnelle en TAL aurait procédé par automates d'extraction, c'est-à-dire par des règles de lecture qui relèvent un terme si elles ont reconnu un syntagme déclencheur. Par exemple, la règle « `(nous )?recrutons (un )?(.....+)` » s'appuie sur un patron pour repérer le nom du métier. Notre méthode, tout en gardant une parenté avec les patrons et les entités nommées, ne pose pas de règles préalables pour extraire l'ensemble des informations pertinentes (métiers, mission, compétences etc.). Elle est donc à la fois moins « savante » mais bien plus adaptable à de nouveaux corpus. Le choix fait dans le cadre du projet SIRE était précisément d'explorer les possibilités offertes par ces approches plus empiriques, dites « induites par les données ».
Outre les applications pratiques visées par le projet (indexation et étiquetage d'offres, panorama de l'emploi), nous proposons ici de présenter la démarche comme une manière de questionner l'interaction entre genres et propriétés grammaticales.

2 A la recherche de la bonne méthode

2.1 Apprentissage automatique du sens ?

Les moteurs de recherche ont généralisé des mesures de similarité automatique entre les documents grace à des relevés de fréquences de termes qui y apparaissent. Mais dès les années 1970, certains auteurs ont signalé que l'on pouvait transposer le principe et que les mêmes relevés de fréquence pouvaient servir à *caractériser les termes* par les documents où ils apparaissent.
Ces modèles de représentation sémantique, dont le premier exemple abouti était la LSA (Latent Semantic Analysis) décrite entre autres dans (Landauer et al. 1998), permettent d'ajouter une étape d'analyse sémantique induite par les données dans les chaînes de traitement en TAL, afin d'aider à constituer des thésaurus puis des ontologies. Ces travaux d'apprentissage sont en pratique souvent effectués dans le cadre d'un genre donné, sans pour autant que ce soit mis en avant par les auteurs.
L'apprentissage automatique nécessite une matière de travail, ce qui implique de chercher dans nos corpus des indices observables du sens d'un terme. Avec la même intuition que celle qui a fait le succès de l'étude des cooccurrences en lexicographie, le modèle DSM pose donc le problème du sens à travers celui de l'étude des usages concrets. Toutefois il définit ses observables d'une façon plus générale comme des « contextes », en se proposant de relever la répartition des fréquences du terme étudié à travers une série aussi grande et variée que possible de ces contextes. Répété pour chaque terme, ce procédé permet ensuite de comparer les termes entre eux (ont-ils la même distribution? sur quoi diffèrent-ils?) et d'étudier la structure globale de leurs oppositions.
Sur ce point les meilleurs indices sont les contextes d'apparition qui constituent l'environnement usuel du terme à définir. Ainsi des contextes comme « **au sein de l'H.** » ; « **directeur d'H.** » ; « **concours de directeur d'H.** », « **H. public** » ; « **H. de province** » peuvent être vus comme co-définissant le sens du terme « **hôpital** ».
Un modèle DSM est avant tout un outil d'observation du corpus modélisé et ses résultats varient considérablement selon le découpage du corpus choisi pour regarder les distributions. On parle de « choix des contextes », et celui-ci, dans notre perspective, détermine la manière dont on analyse la sémantique du genre à travers le corpus. Ce doivent être des contextes possibles à isoler de façon non-supervisée et qui soient pertinents pour le sens des termes. La LSA traditionnelle prenait uniquement en compte le cadre du document. Les variantes présentées dans la littérature choisissent parfois les co-occurrences dans une fenêtre de n mots balayant le texte, ou bien les co-occurrences dans des fragments de texte pré-découpés, etc. Toute la richesse du modèle vient de ce que les contextes où l'on choisit d'effectuer les décomptes peuvent constituer des portions plus ou moins larges du texte. Nous avons par ailleurs expérimenté dans le projet SIRE d'autres manières de créer des cadres pertinents, par exemple avec des contextes formés par des groupes de textes concaténés (pour chercher des voisinages thématiques dans des textes d'une même famille).

Le point qui va nous intéresser ici est que les contextes peuvent aussi suivre des *logiques réticulaires* (en dehors de la linéarité du texte) : par exemple en remplaçant la co-occurence dans une fenêtre choisie par l'appartenance des mots aux mêmes paradigmes de constructions syntaxiques (ex. « manager un projet, une équipe etc. »). C'est cette dernière forme de DSM que nous appliquons ici au sein du genre des offres d'emploi.

2.2 Aperçu des phénomènes de dépendances dans le genre

On peut d'abord remarquer que toute méthode distributionnelle place la collocation au centre du travail de description, ce qui la rend cousine des travaux sur les terminologies spécialisées. C'est sans doute une des raisons qui avait en son temps poussé Harris a articuler son modèle distributionnel classique avec un objectif concret de description des *sublanguages* ou langues de spécialités (Harris 1982).
En effet, la richesse descriptive des langues de spécialité s'appuie fortement sur des formules figées et semi-figées (Mortureux 1995, Krieg-Planque 2009). Les exemples les plus courants s'appuient sur les relations suivantes :

- relation *Nom* ← *Adj*. comme, dans le corpus d'offres d'emploi, « **structure métallique** », « **aisance relationnelle** », « **industrie graphique** », « **outil bureautique** », « **personne âgée** », « **espace vert** » ;
- relation *Nom* ← *Comp. Prép.* (aussi appelées synapsies) comme « **architecte d'intérieur** », « **cabinet de recrutement** », « **chiffre d'affaire** », « **lettre de motivation** »;
- relation *Verbe* ← *Objet* comme « **établir un devis** », « **développer un portefeuille** », « **instruire un dossier** », « **poser un revêtement** », « **adresser sa candidature** ».

Ces considérations terminologiques concernent de manière générale le discours lié à l'emploi et ne sont pas forcément propres au genre des offres. On pourrait par exemple les retrouver dans des genres textuels voisins comme le CV ou la convention collective. Sans évacuer le lexique terminologique des métiers intégré au genre, il faut cependant tenir compte aussi des phénomènes lexico-syntaxiques *non terminologiques* et de leur rôle dans la caractérisation du genre (Rinck 2010).
Les dépendances syntaxiques, entendues au sens large où l'entendait (Tesnière 1959), constituent les relations grammaticales élémentaires au sein de la phrase et en tant que telles on peut s'attendre à ce qu'elles soient moins sensibles au genre que ne le sont les phénomènes macro-textuels (paragraphes et chapitres, titres etc.). La dépendance élémentaire (qualification, détermination) ou encore la coordination et les phénomènes liant prédicat et complément seraient ainsi une couche trop profonde de la constitution des textes, servant moins les besoins rédactionnels propres aux genres que des besoins généraux d'expression sémantique (comme décrire, spécifier, évoquer, etc.).
Notre approche pose que ces « besoins sémantiques » sont mobilisés dans différents genres, mais de manière spécifique selon chacun d'eux. Le genre textuel des offres d'emploi et les langues spécialisées des métiers qu'il porte en lui influent sur les dépendances syntaxiques de multiples manières :

- par la phraséologie et la visée rhétorique : les dépendances unitaires sont soumises à un critère final d'efficacité sémantique (clarté de la phrase) qui variera selon les conditions de production du texte, le public visé et les habitudes stylistiques propres au genre. Citons rapidement quelques exemples :
    1. un premier exemple typique des offres est l'usage de la construction détachée : « **Rattaché au Directeur Commercial et Marketing France, votre mission consiste à développer notre présence et nos ventes de solutions de formation** »[3]
    2. le niveau énonciatif aide également à distinguer des faisceaux de constructions, comme « l'entreprise recrute » ou « nous recrutons » et « vous effectuez », « il effectue » ou « mission : effectuer »).
- par le calibrage du texte : la longueur du message et la densité communicationnelle affectent

[3] Sic

la variété des combinaisons syntaxiques favorables. Cela va du mode paratactique (style télégraphique) aux formes les plus raffinées de complémentation et de qualification ;

1. Le style paratactique des titres exclut par exemple les adjectifs :
   « **Mutuelle recherche , un RESPONSABLE DE CENTRE DE PRESTATIONS SANTE H/F** »
   « **Ingénieur Radio Navigation H/F Toulouse** »
2. A l'inverse la description des compétences favorise l'emploi de la qualification adjectivale :
   « **Vous avez en outre un sens aigu des responsabilités et un bon contact relationnel.** »
3. Le calibrage passe aussi par la construction phrastique, entre autres par une forme très reconnaissable, la séquence de liste des missions :
   « Vous serez en charge de :
   « - L'analyse des comptes.
   « - L'analyse fin de mois de la pertinence des niveaux de dépenses.
   « - Du contrôle des fichiers d'interface comptables.
   « - D'assurer la gestion des immobilisations pour les sociétés du Groupe »

on pourrait ainsi aisément multiplier les exemples de constructions caractéristiques qui vont pouvoir aider nos méthodes d'apprentissage à trouver des régularités d'ordre sémantique propres au genre.

- si de plus le genre textuel est associé à un *domaine de connaissance* au sens des terminologues, les traits en usage dans ce domaine sont transmis au genre (formes syntaxiques utilisées communément pour la dénomination des actants, des objets, des concepts et la description des actions) : les dépendances syntaxiques constituent au sein du genre le lieu pratique où peuvent se forger les combinaisons préférentielles des termes qui font un discours spécialisé, ainsi que la définition même de ces termes.

Pour résumer, l'offre d'emploi fait naturellement appel à la fois au lexique des discours spécialisés et à des formes de construction caractéristiques du genre. L'étiquetage d'un corpus en relations de dépendances permet d'étudier simultanément le figement et les relations autour des termes figés. Pour ces deux objectifs, il faut partir des fréquences des triplets (lexème A, relation de dépendance, lexème B). C'est bien *parce que* ces phénomènes lexico-syntaxiques ont une cohérence au sein du genre que la méthode distributionnelle peut fonctionner sur des contextes en dépendances et apporter des informations d'ordre sémantique.

2.3 La méthode distributionnelle sur les paradigmes de dépendances

L'utilisation de contextes basés sur les relations de dépendances en « input » d'un modèle DSM est recommandée dans la littérature récente (Pado et Lapata 2007) comme une forme plus aboutie de caractérisation sémantique. Bien que le milieu de l'ingénierie linguistique ne s'appuie pas toujours sur les linguistes qui l'on précédé sur les mêmes questions, c'est en quelque sorte par une hypothèse harrissienne que l'on prend l'entourage syntaxique d'un terme comme définitoire de son signifié.

L'approche sémantique de Zellig Harris s'appuie sur une riche mécanique de composition entre les termes, à travers des opérateurs symboliques de transformations sur les combinaisons prédicat/arguments. La méthode proposée par Harris se fixe comme préoccupation cruciale de « coller aux usages », avec dans l'idée le fait que la description exhaustive des compatibilités observées d'un lexème est nécessaire et suffisante pour sa caractérisation syntactico-sémantique. Cette approche à la jonction des structuralismes américain, russe et européen trouve une continuation dans de nombreux travaux de linguistique (Mel'cuk et Aleksandrovič 1988), (Gross 1990). L'évolution récente de l'annotation automatique en graphes de dépendances syntaxiques en fait aujourd'hui l'inspiration centrale des méthodes consistant à entraîner des DSM sur les données grammaticales d'un corpus.

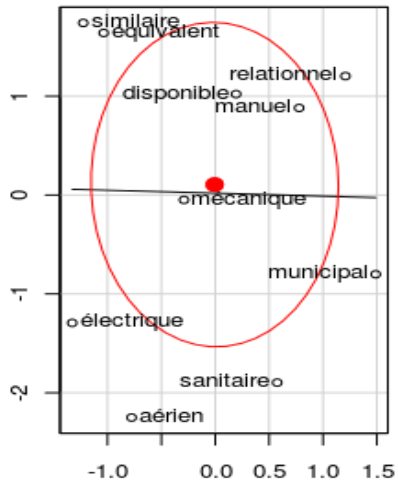

Figure 2: Topologie résultant de la procédure décrite sur les adjectifs

Exemple : une étude miniature de quelques adjectifs

La méthode appliquée ici consiste à noter chaque dépendance observée avec son point d'attache (recteur), et de relever les termes dépendants sous cette entrée. Ainsi la notation $N_7 \leftarrow$ **{disponible, municipal}** indique qu'on constitue un contexte pour la dépendance de qualification N ← Adj ancrée sur « animateur » et qu'on y reporte ses dépendants « disponible » et « municipal ».

Ce relevé par contexte est présenté sous la forme d'une matrice utilisable dans une série de traitements de données dont l'élément principal est la réduction des contextes (colonnes) à un certain nombre k de « dimensions » émergeant selon les faisceaux de recoupement les plus explicatifs pour les données (cf. Figure 2 avec k = 2). L'appelation DSM désigne précisément le modèle d'ensemble soit : un relevé des occurrences de termes noté sur une matrice et cerné par le choix des contextes en entrée, la méthode de décompte, jusqu'aux groupements de colonnes automatiques à la base des faisceaux sous-jacents qui dimensionnent l'espace résultant.

Observations dans le corpus

3x réseau aérien
8x réseau électrique
3x centre sanitaire
3x centre municipal
7x transport aérien
4x transport sanitaire
3x entretien municipal
8x système mécanique
6x système électrique
3x système manuel
9x aptitude manuelle
12x aptitude relationnelle
18x poste équivalent
13x poste disponible
14x poste similaire
4x animateur disponible
6x animateur municipal
4x expérience équivalente
6x expérience similaire

Groupements des adj. par contextes nominaux

$N_1 \leftarrow$ {aérien, électrique}
$N_2 \leftarrow$ {sanitaire, municipal}
$N_3 \leftarrow$ {aérien, sanitaire}
$N_4 \leftarrow$ {électrique, mécanique}
$N_5 \leftarrow$ {manuel, relationnel}
$N_6 \leftarrow$ {similaire, équivalent, disponible}
$N_7 \leftarrow$ {disponible, municipal}
$N_8 \leftarrow$ {équivalent, similaire}

Matrice des adjectifs relevés

	[N1]	[N2]	[N3]	[N4]	[N5]	[N6]	[N7]	[N8]
aérien	3	0	7	0	0	0	0	0
municipal	0	3	0	0	0	0	3	0
sanitaire	0	3	4	0	0	0	0	0
électrique	8	0	0	5	0	0	0	0
mécanique	0	0	0	8	0	0	0	0
manuel	0	0	0	3	9	0	0	0
relationnel	0	0	0	0	12	0	0	0
équivalent	0	0	0	0	0	18	0	4
similaire	0	0	0	0	0	14	0	6
disponible	0	0	0	0	0	13	4	0

Le modèle apporte donc des mesures de distance et de voisinage entre le sens des termes. Ces mesures permettent notamment de travailler sur des agrégats émergents ou clusters. Nous présentons dans la Table 2 en fin d'article ces résultats non-supervisés, qui rejoignent en partie les « types d'information » recherchés par le projet SIRE au niveau textuel.

On observera enfin que notre façon de définir ces contextes se centre sur le paradigmatique. Elle n'est qu'une des variantes possibles pour prendre en compte les dépendances dans un DSM (Purandare et Pedersen 2004).

2.4 Vers une topologie du genre

Il faut souligner qu'en pratique, les études par analyse distributionnelle réalisées dans les dernières années concernent souvent *des genres textuels **ou** des discours particuliers*. La plupart du temps, il s'agit de créer une couche d'interprétation des documents en mettant en place une « ontologie » d'un domaine.
Pour le français, on peut citer (Bourrigault et Lame 2002) sur les textes du domaine légal, (Rousse et De la Clergerie 2005) sur les ressources documentaires en botanique et (Jacquet et Venant 2005) sur le discours journalistique (articles du *Monde*). D'autres travaux mettent plus en avant les questions de construction du texte et du discours comme (Tutin 2007) qui étudie le vocabulaire des articles scientifiques et (Adam et Morlane-Hondère 2009) sur le corpus des articles de Wikipedia. On voit par ces exemples que le genre textuel a une pertinence forte sur les travaux les plus aboutis dans la sphère des études distributionnelles sur dépendances syntaxiques.
Les résultats obtenus sur les offres d'emploi dans notre projet laissent en fait penser qu'un modèle distributionnel entraîné sur les dépendances n'est pas simplement une caractérisation qui affine les distinctions thématiques d'un modèle distributionnel entraîné sur les documents (champs sémantiques liés aux professions, etc.), mais qu'il opère à un niveau du sens différent, lié très nettement au genre textuel, et qu'on peut appeler niveau classématique : il détermine des classes d'objet partageant certains « classèmes », c'est-à-dire des sèmes génériques (Rastier 2001).
L'utilisation de grands corpus annotés automatiquement permet une meilleure compréhension de l'espace d'ensemble formé par toutes les observations dans un corpus. Les travaux qui ont conduit au développement de la méthode distributionnelle se concentraient avant tout sur l'évaluation des relations sémantiques unitaires (de terme à terme) obtenues (comme Pado et Lapata 2008). Dorénavant, on passe progressivement de la caractérisation de chaque mot à la mise en lumière d'une *topologie distributionnelle*. Cela vaut autant pour le niveau local via les faisceaux de combinaisons apparentés pour plusieurs mots (paradigmes, termes voisins) que pour l'analyse de la configuration d'ensemble du vocabulaire au niveau global (groupes émergents, groupes voisins). Les techniques d'analyse de données s'avèrent précieuses pour cette étude d'ensemble (réduction de dimensions, regroupement de cas similaires, propagation d'étiquettes, étude de graphes).

3 Résultats d'un DSM entraîné sur les dépendances dans les offres d'emploi

3.1 Précisions pratiques

Le corpus d'offres d'emploi développé dans le cadre du projet SIRE au laboratoire MoDyCo est constitué de manière à équilibrer les métiers et les sources. Nos scripts de collecte opèrent par parcours de l'arbre de classement de 3 sites (Pôle Emploi, APEC et Monster.fr) afin de prélever un échantillon d'offres pour chaque métier (24700 offres avant dédoublonnage). Le corpus final compte 21817 documents[4] et comprend environ 2,5 millions de mots. L'étiquetage en dépendances est effectué automatiquement par l'analyseur FRMG basé sur un moteur DyALog (De La Clergerie *et al.* 2009) via la chaîne de traitement associée du laboratoire Alpage.
Les termes pris en compte sont les 16000 les plus fréquents, et incluent des termes simples et des séquences composées semi-figées de deux lexèmes. Autour de ces termes, l'algorithme relève les relations de dépendances. et cherche à reconstituer des relations deux à deux entre « termes pleins » (le recteur et les relateurs éventuels (« de », « à »...) sont reportés comme intitulé du contexte et les termes régis sont relevés sous cet intitulé). Notre implémentation ne retient qu'une petite famille de relations entre lexèmes. Ce sont les relations Sujet → Verbe, Verbe ← Objet, Nom ← Complément

[4] On adjoint aussi 1006 fiches métier provenant de l'Onisep (Office National d'Information Sur les Enseignements et les Professions) et du Rome (Répertoire Opérationnel des métiers et des emplois) . La collecte conserve les métadonnées originelles, notamment les codes métier pour les offres Pôle Emploi. Le corpus est ensuite nettoyé par des heuristiques simples qui suppriment les doublons et ce qu'on appelle le « boilerplate » (ou co-texte peu utile de la navigation web).

de Nom, Nom ← Adjectif. Des contextes beaucoup plus riches sont préconisés par Pado et Lapata (relations plus diversifiées, séquences de dépendances à plusieurs termes au lieu des relations simples), mais cette restriction aux 4 relations se prête sans doute mieux au mode de relevé par paradigmes car avoir des contextes simples favorise l'émergence de traits sémantiques plus génériques.

3.2 Un objectif surprise : les classes d'objet

Comme il a été dit plus haut, la méthode permet d'adjoindre à chaque terme une position dans un espace à plusieurs dimensions. Cela a pour effet principal de pouvoir mesurer la proximité ou distance de deux termes à travers leurs contextes de décompte. Dans le projet SIRE, les réseaux de proximités qui apparaissent entre les termes de façon induite étaient explorés avant tout pour retrouver des domaines thématiques et des isotopies liées à tel ou tel métier. Concrètement, nous avons procédé en envisageant des formes de contextes très variés pour tester leur influence sur l'espace résultant.

L'application de la méthode sur des contextes larges comme en LSA (contextes = co-occurences dans document) a déjà permis de modéliser dans notre dictionnaire la parenté thématique entre les termes (champ sémantique professionnel). Les contextes par documents fonctionnent bien avec les offres d'emploi, car chaque offre a une unité thématique imposée par le métier du poste décrit. On obtient donc des familles thématiques (domaine médical, domaine de l'édition, etc.). Dans la première étape de nos travaux, nous nous attendions à ce que les contextes par dépendances prolongent et affinent les liens d'association thématique.

Toutefois, ce n'est pas cette facette de la sémantique des termes qui ressort le plus avec les contextes par dépendance. L'espace est analysé à travers une « clusterisation », ou groupement de termes voisins dans la topologie. Qualitativement, la parenté thématique des termes voisins dans l'espace est faible quoique existante[5]. Dans ce sens, ils échouent aux tests bien plus souvent que les groupes de termes qui émergeaient des espaces par contextes larges. L'étude de notre espace résultant sur les dépendances montre en revanche des régions où les termes appartiennent plus ou moins à une même catégorie abstraite (classes d'objet).

Distribution des termes dans l'espace et meilleurs groupements émergents

La répartition des éléments dans l'espace s'avère intéressante. Dans les espaces par contextes larges, les termes étaient répartis de façon assez homogène et continue alors que notre modélisation par les dépendances simples donne des groupes bien plus marqués, aux frontières discontinues positionnées autour d'une quarantaine de points attracteurs dans la répartition globale, dont une demi-douzaine de zones où le voisinage est totalement resserré. Ces zones correspondent à quelques dizaines de termes « au même endroit dans l'espace », qui partagent tous leurs contextes et sont donc condensés en un point unique[6].

Dans l'ensemble, les axes les plus nets de groupement forment une parenté *classématique.* On veut dire par là que les groupes les plus conséquents sont caractérisés par une unité de l'ordre de la « classe d'objet » : objets concrets, notions, activités, humains, lieux, etc. (Le Pesant et Mathieu-Colas 1998). Voici par exemple des extraits de trois groupes émergents jugés pertinents (avec une partie de leur faisceau de contextes). Les deux premiers clusters correspondent à une classe d'objets « objet inanimé concret » mais à deux familles thématiques (on voit là l'intérêt d'un post-traitement manuel).

Échantillon du cluster 1

système__nc	*matériel__nc*
ligne__nc	*train__nc*

[5] Deux termes voisins dans l'espace ne seront pas spécifiques au même métier, mais ils peuvent apparaître groupés par "environnements" professionnels (métiers manuels, groupes industriels, entreprise de services ou de commerce etc.).

[6] Autrement dit, on observe des points d'attraction dans la répartition des termes. Ces attracteurs sont comme des noyaux prototypiques de régions plus larges correspondant aux classes d'objets émergents dans le genre

équipement__nc
câblage__nc
nacelle__nc
moteur__nc
bateau__nc
ossature__nc

Extraits des contextes observés

hasAdj.électrique__adj	13
isDeN.conduite__nc	11
hasAdj.industriel__adj	9
isObj.réaliser__v	9
isDeN.installation__nc	8
hasAdj.mécanique__adj	7
hasAdj.électronique__adj	7
isDeN.exploitation__nc	7
isDeN.maintenance__nc	7

Échantillon du cluster 2

meuble__nc
composant__nc
papier__nc
chaussure__nc
ameublement__nc
panneau__nc
matériaux--souple__nc--adj
vin__nc
fromage__nc

Extraits des contextes observés

isDeN.fabrication__nc	8
isObj.vendre__v	4
isObj.connaître__v	3
isObj.préparer__v	3
isObj.réaliser__v	3
isObj.fabriquer__v	2
isObj.porter__v	2

Échantillon du cluster 3

personne__nc	*clientèle__nc*
enfant__nc	*particulier__nc,*
public__nc	*patient__nc,*
maître__nc	*famille__nc,*
prospect__nc	*résident__nc,*
étudiant__nc	*débutant__nc,*
usager__nc	*dirigeant__nc,*
locataire__nc	*apprenti__nc*

Extraits des contextes observés

isObj.accompagner__v	26
isObj.accueillir__v	22
isObj.aider__v	13
isObj.conseiller__v	12
isObj.former__v	11
isObj.orienter__v	11
isDeN.accueil__nc	10
isObj.encadrer__v	10
isObj.informer__v	10
isObj.recevoir__v	10
isDeN.accompagnement__nc	9
isObj.assister__v	9
isObj.prendre__v	7
isObj.recruter__v	7
isObj.renseigner__v	7
isSuj.effectuer__v	7
hasAdj.atteint__adj	6
hasAdj.âgé__adj	6
isObj.gérer__v	6
etc.	

Nous avons choisi ces 3 groupes comme exemples parmi les résultats les plus probants, mais certains regroupements sont nettement plus mauvais. On peut dire qu'environ un quart des groupes émergents n'a pas d'unité claire, certains phénomènes de polysémie du recteur nuisant fortement à l'approche mise en place. Malgré ces défauts inhérents à toute méthode non-supervisée, l'utilité des groupes obtenus est manifeste. En observant de près ces résultats, on trouve d'autres généralisations possibles, des proximités paradigmatiques au sens large, simples mais efficaces dans leur effet sémantique de typage. Par exemple les *compétences* relevées s'inscrivent dans une famille de contextes apparentés : on les trouve comme objets de verbes de possession (« **posséder** », « **avoir** ») ou de maîtrise et/ou qualifiées par des adjectifs exprimant la nécessité (« **indispensable** », « **exigée** », « **impératif** »). Ce faisceau distingue les compétences des autres types emblématiques évoqués (métiers, secteurs, activités...). Les termes ayant ces propriétés dans l'espace sont bien groupés à part des autres (opposition avec les *non-objets* de « avoir », etc.).
Malgré la pauvreté de l'implémentation actuelle[7] (ou grâce à elle), le résultat est quelque peu novateur. En réussissant à recouper de façon à peu près cohérente les données lexicales en grands groupes notionnels, elle confirme que la répartition des propriétés lexico-syntactiques dans le genre des offres d'emploi est caractérisée par des schémas de construction propres à ce corpus. Ces schémas sont transversaux (appraissent dans les différentes thématiques professionnelles), mais elles sont les traces syntaxiques caractéristiques d'une façon qu'à l'offre d'emploi de *mettre en scène l'information* à travers des types informationnels (noms de métiers, expressions des compétences,

[7] cf. début du §4,2

lieu ou secteur d'activité). Il rejoint là-dessus les résultats de (Jacquet et Venant 2005) qui se servent d'un modèle similaire pour désambiguïser des termes paronymes. Mais chez nous ces traces syntaxiques servent simplement à faire émerger des classes sémantiques (comme celle des « savoir-faire »). On confirme au passage que certaines de ces classes ont une pertinence strictement liée au corpus spécialisé et dans les usages linguistiques qui lui sont associés, tandis que d'autres sont à portée générale (humain, objet...). On fait ainsi émerger une perspective d'interprétation lexicale propre au genre.

En approfondissant, on peut par exemple utiliser les constructions pour classer sémantiquement les compétences en variantes : les expressions décrivant des savoir-faire et connaissances pratiques (« **expertise comptable**» ou « **maîtrise de la PAO** ») peuvent être qualifiées par une autre famille d'adjectifs exprimant la précision des connaissances (« **technique** », « **approfondi** ») et se retrouver dans des contextes propres aux activités (« **exercer** », « **assurer** »). A l'inverse, les critères de personnalité comme « **sens du relationnel** », « **qualité de l'expression** », apparaissent également comme objets de verbes de possession mais sont les seuls à complémenter des verbes comme« **faire preuve** » ou « **démontrer** ».

On peut dire pour résumer que le genre appelle un découpage du texte en types informationnels, et que ces types se reflètent dans des classes lexicales appropriées, définissables de façon empirique par leurs cadres de sous-catégorisation syntaxiques. Pour évaluer ce phénomène, nous avons confronté les groupes lexicaux émergent à une typologie qui avait été élaborée indépendamment et manuellement (il s'agissait d'une typologie ayant une visée de découpage du texte en séquences types pour une extraction d'information plus traditionnelle (Loth 2010)).

Type d'information	**Principales formes**	**Fréq. moy.** (par offre)	**Long. moy.** (en mots)
Poste	intitulé du poste, mention du domaine professionnel	2,36	3,98
Mobilité	lieu de travail, déplacements prévus, zone commerciale	0,85	2,67
Etablissement	nom, groupe, site web, données quantitatives (CA, effectifs)	2,43	3,15
Secteur	activité, branche, parfois lieux, souvent produit ou service vendu	2,06	5,02
Environnement	responsable, interlocuteurs, équipe/service, conditions de travail	2,09	4,63
Missions	fonction principale, tâches, objectifs liés au poste	**7,99**	8,06
Contrat	type de contrat, durée, horaires, salaire	1,10	3,41
Expérience	[ancienneté + nature de l'expérience] (sous forme phrastique composée)	1,29	9,45
Compétences.: Savoir-faire	connaissance d'un champ, objets spéc., compétences techniques, langues	2,04	3,74
Compétences : Personnalité	*(pas d'articulation secondaire émergente)*	3,25	2,63
Compétences: Formation	diplôme, niveau, filière de qualification	0,86	5,26
Contact	e-mail, personne à contacter, adresse, n° de réf., procédure à suivre	0,79	7,20

Table 1: Typologie préalable des informations « à extraire » : briques logiques dans la construction de l'offre, utilisée pour évaluer les types émergents ci-dessous

Une comparaison manuelle montre des groupes quelque peu différents, mais apparentés aux premiers.[8] La Table 2 en présente une typologie provisoire (sous réserve d'une amélioration de la

[8] Une comparaison à grande échelle est aussi menée via une tâche de prédiction d'étiquettes de type sur un corpus d'expressions connues. Les regroupements émergents de l'espace par dépendances sont pris comme attributs (*feature vector*) dans un régression logistique et permettent de prédire 78% d'étiquettes correctes.

méthode : la prise en compte des relations prépositionnelles autres que « de » serait particulièrement importante). Cette vision lexicale nous rappelle aussi que le typage sémantique s'effectue dans la phrase, en contexte. Avec le seul lexème, il est périlleux de vouloir créer de tels groupes. Un terme comme **« administration scolaire »** ne peut-il pas désigner tout à la fois un *lieu*, une *mission*, un *secteur d'activité*, une *compétence*, une *notion*, voire un *interlocuteur* ? On voit bien les limites de toute catégorisation rigide. On rappellera seulement qu'avec la notion de distance (métrique continue), notre modèle ne force pas à faire cette catégorisation rigide. Elle est testée ici pour apprécier les résultats et évaluer la convergence entre les types d'information attendus et les types émergents.

Les grand groupes présentés dans la Table 2 correspondent à des regroupements des groupes émergents. Nous appelons ces grands regroupements une *perspective* de recodage, et elle est la seule étape « faite à la main » du traitement. Les groupes émergents étaient ici au nombre fixé arbitrairement de 40, de tailles inégales. Par exemple pour la ligne « Secteur, domaine », on a trouvé un groupe émergent d'adjectifs et 4 groupes émergents de noms simples ou d'expressions nominales. Les exemples et le pourcentage associé à la case (14,6%) correspondent à l'union de ces 5 groupes émergents.

Intitulé subjectif du grand groupe	**Exemples de termes**	**Part des termes relevés**
Mission, tâche, objectif	accompagnement de projet, assembler, chiffrage des solutions, dépannage, encadrer, entretien préventif, ~~effectuer~~, placer un produit, promotion de produit, rédaction des spécifications, suivi de chantier, surveiller la voie, tenue de dossier, vendre	15,5%
Objets concrets	article funéraire, communiqué de presse, ~~étanchéité toiture~~, garniture, ~~gymnastique médicale~~, ~~interview~~, instrument du musique, moteur d'avion, panneau de bois, pièce défectueuse	9,2%
Objets matière	~~élément~~, matières premières, pierre, ~~repas~~, revêtement, sol, verre, viande	1,4%
Humains métiers	agent d'entretien, chef d'atelier, chef de chantier, directeur de développement, enseignant, infirmier, médecin, pâtissier, pharmacien, régleur	8,0%
Humains interlocuteurs	adolescent, ~~animal~~, candidat, clientèle, élève, patient, personne âgée, personnel d'agence, prospect, responsable hiérarchique, ~~voyage~~	1,7%
Compétence, formation	brevet, BTS, formation supérieure, permis, spécialité vidéo	4,7%
Compétence, savoir-faire	langue anglaise, outil bureautique, questions juridiques, resistance physique, technique de vente	5%?[9]
Compétence, personnalité	aimer le terrain, esprit commercial, esprit d'initiative, qualités d'expression	3,8%
Neutres, intypables	~~assurer la rentabilité~~, avoir besoin, baigner, changer tout, compléter l'équipe, donner l'opportunité, endosser la responsabilité, faire le succès, gravir les échelons, ~~poser la charpente~~, ~~renseigner un tableau~~, ~~soigner la présentation~~, spécifier, suppléer	9,0%
Lieu, Service, Atelier, etc.	association, atelier, bâtiment communal, cabinet d'expertise, département, enseigne, établissement de formation, filiale, magasin, ~~mutualité~~, service de direction	8,2%
Notions abstraites de domaine	alpin, arboricole, donnée commerciale, évenement sportif, information médicale, lien, saison, stratégie, sujet, système d'information	10,3%
Secteur, domaine	~~adjacent~~, agroalimentaire, artisanal, chimique, forestier, graphique, pharmaceutique, psychologique, sanitaire, socioculturel, ~~téléphonique~~, touristique ; arts graphiques, banque d'investissement, blanchisserie, ~~blanchisseur~~, ~~cloison~~, construction de logements sociaux, énergies nouvelles, fabrication de chaussures, génie climatique, industrie mécanique, ~~outil de découpe~~, ~~préparation de pâte~~, recherche clinique, sécurité des installations, télésurveillance, urbanisme	14,6%
Autres		8.5%

Table 2: Typologie provisoire des groupements émergents de la distribution de 16000 termes parmi 21800 offres

[9] La catégorie des savoir-faire existe et est centrale, mais les décomptes effectués pour l'instant ont rangé ses représentants dans certaines catégories voisines : objets concrets, notions et missions

La classe la plus importante est constituée par ce qu'on a appelé les *missions,* ou taches à effectuer pour le poste. Elle comprend plusieurs groupes émergents qui dans leur forme naturelle étaient coupés entre formes à tête verbale et formes à tête déverbale. S'il est naturel qu'une telle opposition émerge dans la topologie des dépendances, c'est bien le rôle du recodage en perspectives de ne rechercher que les classes d'oppositions qui seront pertinentes à la navigation dans les offres.
Dans la table 2, les termes barrés indiquent des exemples de termes dont on estime qu'ils n'auraient pas dû être placés dans le grand groupe final. La part réelle de ces « erreurs » d'apprentissage est d'environ 20% à 30% sur les 8000 termes les plus fréquents[10]. Elle semble malheureusement augmenter rapidement pour les termes plus rares. Toutefois les développements de la méthode d'apprentissage, un travail plus conséquent sur le recodage en « perspectives recherchées » et l'utilisation d'un corpus plus grand en entrée permettent d'espérer plus de précision sous peu.

3.3 Des classes d'objets lexicales aux types d'information textuelle

Le recodage en perspectives utile pour la typologie de la Table 2 montre la voie pour que la reconnaissance des classes d'objets soit mise au service d'une indexation au niveau plus macroscopique du texte. Les classes sont les unités sous-jacentes aux types d'information que nous cherchions à extraire et elles s'articulent via les dépendances en une sorte de grammaire des types, autrement dit des règles de combinaison, comme dans les exemples suivants :

- « **conditionnement de produits frais** » *(type résultant = mission)*
  [activité] [objet concret]
- « **responsable de conditionnement de produits frais** » *(type résultant = métier)*
  [agent] [activité] [objet concret]
- « **2 ans à un poste de responsable de conditionnement** » *(type résultant = expérience)*
  [ancienneté] [agent] [activité]

L'emploi est une réalité qui pré-existe à l'offre, mais le genre des offres détermine ce qui est dit de l'emploi. On peut envisager à l'avenir une modélisation à plusieurs niveaux du sens des expressions. Le genre fait émerger des types informationnels qui constituent un point d'échange entre les paliers de la sémantique et de la syntaxe : ce sont des séquences propres au genre de l'offre d'emploi, mettant en correspondance forme et contenu selon une combinatoire *restreinte* de relations syntaxiques et de classes lexicales qu'on est capable de retrouver par apprentissage sur des corpus.

4 Conclusion et pistes

4.1 Hypothèse de complémentarité des niveaux en modélisation sémantique

Au vu des résultats, on peut formuler l'hypothèse suivante : *La caractérisation d'un lexème par son environnement local (voisins immédiats du mot ou lexèmes ayant les mêmes dépendances syntaxiques) est complémentaire de la caractérisation par environnement larges (cooccurrences dans tout un document ou dans un groupe thématique de documents, méthodes dites LSA*. Les deux peuvent servir à l'étude lexicale spécialisée, mais ils interviennent sur des facettes du sens différentes, ou si on veut à des niveaux de grain sémantique différent.
Certes, la caractérisation distributionnelle en contextes larges replace le lexème dans une isotopie thématique (un champ d'évocation commun aux documents sources où il apparaît). Mais l'observation locale circonscrit mieux la classe d'objets propre au lexème étudié. En effet, comme le mettent en avant les travaux sur les classes d'objets (Le Pesant et Colas 1999), les restrictions de sélection pour remplir une relation de dépendance sont équivalentes à des traits définitionnels de classes (par exemple : boire + <LIQUIDE>). L'analyse sur des contextes de voisinage syntaxique immédiat ne vient pas améliorer la qualité des groupes thématiques, mais la compléter par une

[10] Décompte effectué sur un échantillon de 1000 termes parmi les 8000 les plus fréquents.

répartition qu'on peut qualifier de classématique.

4.2 Questions posées au corpus de genre

Les faisceaux de contextes apparaissant dans un genre sont à la fois ceux qu'il hérite de la langue générale et ceux qu'il introduit ou renforce lui-même. Notamment, la spécialisation thématique propre aux genres s'accompagne d'une recrudescence de certains emplois de termes. D'autres propriétés émergent des besoins stylistiques (phraséologie adaptée à la rhétorique typique du genre).

Après recensement des dépendances, les classes d'objets les plus importantes (en nombre de termes observés) sont ainsi souvent spécifiques au genre et elles participent de la construction des séquences discursives qui le caractérisent. L'approche distributionnelle par les dépendances syntaxiques offre une représentation plus claire des grandes tendances abstraites dans la sémantique des mots pour étudier ce point de passage entre sémantique lexicale et sémantique du texte. Pour en tirer des perspectives de navigation, ces tendances peuvent être liées à des types d'information *génériques*. Les repérer équivaudrait alors à décrire une topologie des combinaisons syntaxiques autorisées entre les occurrences des termes d'un genre donné, en y cherchant des régions correspondant aux « tags » les plus pertinents.

Ces développements permettent par ailleurs d'envisager le modèle sémantique distributionnel comme un outil d'observation syntaxique et stylistique des corpus spécialisés. La méthode distributionnelle permet de mettre en avant les groupes élémentaires de termes munis d'une approximation de leur « connectique » (axes de leurs relations/oppositions sémantiques). Elle peut être directement utilisée pour typer les séquences d'information propres à un genre structuré. On peut ainsi chercher à établir un ensemble de classes émergeant du lexique employé comme "signature caractéristique" d'un genre textuel.

4.3 Remerciements

Le projet SIRE est un projet d'analyse automatique des offres d'emploi financé par l'Union Européenne et la région Ile-de-France dans le cadre du fonds FEDER (Fonds européen de développement régional). C'est un projet de recherche sur 3 ans organisé par la société Lingway depuis 2009, en partenariat avec la société Proxem et le laboratoire MoDyCo, UMR 7114 du CNRS. Ces travaux donnent lieu à la thèse sur *La sémantique des offres d'emploi* de l'École Doctorale 139 « Connaissance, Langage, Modélisation » à l'université de Nanterre.

4.4 Bibliographie

Adam, C. et **Morlane-Hondère**, F. (2009) « Détection de la cohésion lexicale par voisinage distributionnel: application à la segmentation thématique. », *Actes de RECITAL'09 à Senlis.*

Biber, D. (1988), *Variation across speech and writing*, Cambridge, Cambridge University Press.

Bourigault, D. et **Lame**, G. (2002). Analyse distributionnelle et structuration de terminologie: Application à la construction d'une ontologie documentaire du Droit. *Traitement automatique des langues* 43(1), pp. 129-150.

Branca-Rosoff, S., (ed) (1999), *Types, modes et genres de discours*, *Langage et Société*, 87, Paris, Maison des Sciences de l'Homme.

Condamines, A. (2006), « Modes de construction du sens en corpus spécialisé », *Cahiers de grammaire*, 30, pp. 75-88.

De La Clergerie, E. et al (2009), « FRMG: Evolutions d'un analyseur syntaxique TAG du français », *Journée ATALA: Quels analyseurs syntaxiques pour le français ?* (Paris, 2009)

Gross, M. (1990), « Sur la notion harrissienne de transformation et son application au français. »

Langages, 25(99), pp. 39-56.

Harris, Z. (1982), « Discourse and Sublanguages » Dans R. Kittredge & J. Lehrberger, eds. *Sublanguage: studies of language in restricted semantic domains*. Walter de Gruyter, pp. 231-245.

Jacquet, G. et **Venant**, F. (2005) « Construction automatique de classes de sélection distributionnelle », *Actes de TALN 2005 (Dourdan),* pp. ???

Le Pesant, D. et **Mathieu-Colas**, M. (1998) « Introduction aux classes d'objets » *Langages* 32 (131) pp. 6-33.

Marchal, E. et **Torny**, D. (2003), « Des petites aux grandes annonces: le marché des offres d'emploi depuis 1960 », *Travail et Emploi*, 95, pp. 59-71.

Krieg-Planque, A. (2009). *La notion de « formule » en analyse du discours. Cadre théorique et méthodologique*. Besançon, Presses Universitaires de Franche-Comté.

Loth, R. (2010), « Les offres d'emploi comme texte ? », *Les Cahiers de l'ED 139 2010,* pp. 97-106.

Malrieu, D., et **Rastier**, F., (2001), « Genres et variations morpho-syntaxiques », *Traitement Automatique des Langues*, 42 (2), pp. 548-577.

Mel′cuk, I.A. et **Aleksandrovĭc**, I. (1988), *Dependency syntax: theory and practice*, New York, State University Press.

Mortureux, M.-F. (1995), « Les vocabulaires scientifiques et techniques », Beacco et Moirand (éds), *Les Carnets du Cediscor* (3) : *Les enjeux des discours spécialisés,* pp. 13-25.

Rastier, F. (2001), *Arts et sciences du texte*, Paris, Presses Universitaires de France.

Padó S. et **Lapata**, M. (2007). « Dependency-based construction of semantic space models », *Computational Linguistics,* 33(2), pp. 161-199.

Paltridge, B. (1997), *Genres, frames and writing in research settings*, Amsterdam, Philadelphia, John Benjamins.

Purandare, A. et **Pedersen**, T. (2004). « Word sense discrimination by clustering contexts in vector and similarity spaces ». *Proceedings of the Conference on Computational Natural Language Learning,* pp. 41-48.

Rinck, F. (2010), « L'analyse linguistique des enjeux de connaissance dans le discours scientifique », *Revue d'anthropologie des connaissances*, 4(3), pp. 427-450.

Rousse, G. et **De La Clergerie**, E. (2005), « Analyse automatique de documents botaniques: le projet Biotim », *Actes de TIA 2005,* pp. 95-104.

Tesnière, L. (1959), *Élements de syntaxe structurale*, Paris, Libraire Klincksieck.

Tutin, A. (2007), « Traitement sémantique par analyse distributionnelles des noms transdisciplinaires des écrits scientifiques », *Actes de TALN 2007*, pp. 283-292.